



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

contact@spboucher.ai

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

August 2026

Version 0.1 (draft)

Abstract

[TO BE WRITTEN AFTER THE EMPIRICAL PIPELINE HAS RUN — 150 words, summarising the multi-asset comparison of HAR-type, GARCH-type and machine-learning forecasts of realized volatility, the evaluation design (rolling out-of-sample, QLIKE/MSE, Diebold–Mariano, Model Confidence Set), the headline empirical findings, and the economic-value results.]

Keywords: Realized volatility, HAR-RV, GARCH, Machine learning, Volatility forecasting, Model confidence set, Volatility timing

JEL Classification: C22, C45, C53, G17

1. Related literature

1.1. 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. [Bollerslev](#) showed that standard models forecast well once evaluated against realized rather than squared daily returns; [Bollerslev and Russell](#) documented the distributional properties of realized variance for exchange rates and equities; and [Bollerslev](#) developed its asymptotic theory. [Bollerslev](#) 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 this paper operates in.

At very high sampling frequencies, market microstructure noise breaks the consistency of plain realized variance ([Bollerslev](#)). The literature responded with noise-robust estimators: subsampled and two-scale estimators ([Bollerslev](#)), and realized kernels ([Bollerslev](#)). [Bollerslev](#) compared some 400 realized measures across 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 ([Bollerslev](#)), formal jump tests ([Bollerslev](#)), and threshold refinements ([Bollerslev](#)). [Bollerslev](#) introduced realized semivariances, whose signed decomposition [Bollerslev](#) showed to matter for persistence: negative semivariance predicts future volatility far more strongly than positive.

1.2. HAR models and extensions

[Bollerslev](#) proposed the heterogeneous autoregressive (HAR) model, whose daily–weekly–monthly cascade approximates long memory with three OLS coefficients; it has become the workhorse benchmark of the field. Extensions incorporate the continuous/jump split ([Bollerslev](#)), signed semivariances ([Bollerslev](#)), and attenuation-bias corrections: [Bollerslev](#) let the daily coefficient shrink on days when realized quarticity signals a noisy RV estimate (HARQ). Logarithmic specifications are common because log RV is close to Gaussian ([Bollerslev](#)) and the log transform tames heteroskedastic measurement error. On the GARCH side, [Bollerslev](#) showed that little beats a GARCH(1,1) for daily returns, while [Bollerslev](#) tied the latent-variance recursion to a realized measure (Realized GARCH), typically closing most of the gap to HAR-type models. Our benchmark suite spans exactly this spectrum — five HAR variants, three daily GARCH variants, and Realized GARCH — so that the machine-learning comparison is anchored against the strongest available econometrics rather than a single straw man.

1.3. Machine learning for volatility forecasting

Applications of statistical learning to realized volatility are more recent. ? used the LASSO to select lag structures, finding it recovers HAR-like patterns. ? documented gains from recurrent neural networks for monthly equity-index volatility. ? trained learners on large limit-order-book and news datasets for intraday volatility. Closest to our design, ? ran a systematic horse race of regularized regressions, trees and neural networks against HAR on 29 Dow Jones stocks, finding double-digit QLIKE improvements concentrated in periods of market stress and in longer horizons. In the broader return-prediction literature, ? established the now-standard result that trees and shallow networks dominate linear methods and that pooling across assets is crucial for regularization — a theme our pooled multi-asset model imports into the volatility context. For cryptocurrencies, GARCH-type comparisons dominate (?), with high-frequency ML applications still scarce.

1.4. Forecast evaluation and economic value

Because volatility is observed only through noisy proxies, loss functions must be chosen carefully: ? characterizes the class of robust losses — QLIKE and MSE — for which rankings under a proxy coincide with rankings under the truth. Statistical comparisons rely on ? and ? tests, the model confidence set of ?, and, in unstable environments, the fluctuation analysis of ?. Forecast combination is a persistent empirical winner (??). The economic-value strand converts statistical accuracy into portfolio outcomes: ?? priced volatility timing with mean-variance utility, and risk-management evaluation backtests VaR coverage (??).

1.5. Contribution relative to the literature

Three gaps motivate this paper. First, most ML-versus-HAR evidence comes from single asset classes — typically US equities (?) — leaving open whether the documented gains generalize 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 (?) but barely for realized volatility. 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.

2. Methodology

2.1. 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

$$dp_s = \mu_s ds + \sigma_s dW_s + \kappa_s dq_s, \quad (1)$$

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

$$RV_t = \sum_{i=1}^n r_{t,i}^2 \xrightarrow{P} \underbrace{\int_{t-1}^t \sigma_s^2 ds}_{\text{integrated variance}} + \underbrace{\sum_{t-1 < s \leq t} \kappa_s^2}_{\text{jump variation}} \quad (2)$$

consistently estimates total quadratic variation as $n \rightarrow \infty$ (??). In practice market microstructure noise contaminates returns at the very highest frequencies (??), 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 (?). Following ?, 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 flat-top-free Parzen realized kernel of ?,

$$RK_t = \gamma_0 + \sum_{h=1}^H k\left(\frac{h}{H+1}\right) (\gamma_h + \gamma_{-h}), \quad \gamma_h = \sum_i r_{t,i} r_{t,i-h}, \quad (3)$$

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} (?). The non-flat-top Parzen weights guarantee non-negativity.

Bipower variation and jumps. The realized bipower variation of ?,

$$BV_t = \mu_1^{-2} \frac{n}{n-1} \sum_{i=2}^n |r_{t,i}| |r_{t,i-1}|, \quad \mu_1 = \sqrt{2/\pi}, \quad (4)$$

is consistent for integrated variance in the presence of jumps. Jump days are identified with the ratio-form test of ?,

$$Z_t = \frac{1 - BV_t/RV_t}{\sqrt{\theta \max(1, TQ_t/BV_t^2)/n}} \xrightarrow{d} N(0, 1), \quad \theta = \frac{\pi^2}{4} + \pi - 5, \quad (5)$$

where TQ_t is the tripower quarticity. The jump and continuous components are then

$$J_t = \mathbb{1}\{Z_t > \Phi_{1-\alpha}^{-1}\} \cdot \max(RV_t - BV_t, 0), \quad C_t = RV_t - J_t, \quad (6)$$

with $\alpha = 0.001$ (?). To limit noise distortions, BV_t , TQ_t and the test are computed on the sparse 5-minute grid.

Realized semivariances. Following ?, the positive and negative semivariances

$$RS_t^+ = \sum_i r_{t,i}^2 \mathbb{1}\{r_{t,i} > 0\}, \quad RS_t^- = \sum_i r_{t,i}^2 \mathbb{1}\{r_{t,i} < 0\}, \quad (7)$$

decompose RV_t by the sign of the underlying returns; the signed jump variation is $SJ_t = RS_t^+ - RS_t^-$ (?). Finally, the realized quarticity $RQ_t = \frac{n}{3} \sum_i r_{t,i}^4$ measures 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 UTC. Daily close-to-close returns r_t (which include the overnight component for session-limited markets) feed the GARCH-family models and the economic-value exercises.

2.2. Econometric benchmarks

HAR family. The heterogeneous autoregressive model of ? regresses future RV on daily, weekly and monthly averages,

$$\overline{RV}_{t+1:t+h} = \beta_0 + \beta_d RV_t + \beta_w \overline{RV}_{t-4:t} + \beta_m \overline{RV}_{t-21:t} + \varepsilon_{t+h}, \quad (8)$$

where $\overline{RV}_{t+1:t+h} = h^{-1} \sum_{j=1}^h RV_{t+j}$ and all horizons are estimated by direct projection. We consider five extensions: HAR-J adds the jump component J_t ; HAR-CJ replaces the RV averages with continuous-component averages and adds J_t (?); SHAR splits the daily lag into RS_t^+ and RS_t^-

(?); HARQ interacts the daily lag with $\sqrt{RQ_t}$ to discount noisy RV observations (?); and log-HAR estimates (8) in logarithms, with forecasts retransformed as $\exp(\hat{y} + \hat{\sigma}_\varepsilon^2/2)$. Following ?, all forecasts pass through an “insanity filter” that clips them to the range of the target observed in the estimation window.

GARCH family. On daily returns we estimate GARCH(1,1) (?), GJR-GARCH (?) and EGARCH (?) with Gaussian quasi-likelihood. Because these models forecast the conditional variance of close-to-close returns — which includes overnight variation absent from intraday RV — their forecasts are mapped into RV units with the in-sample proportionality factor $\hat{c} = \overline{RV}/\hat{\varepsilon}^2$ re-estimated in each rolling window, in the spirit of ?. Multi-step variance paths are analytic for GARCH and GJR and simulated (500 paths) for EGARCH.

Realized GARCH. The log-linear Realized GARCH(1,1) of ? couples the latent variance h_t with the realized measure x_t :

$$\begin{aligned} r_t &= \sqrt{h_t} z_t, & z_t &\sim N(0, 1), \\ \log h_t &= \omega + \beta \log h_{t-1} + \gamma \log x_{t-1}, \\ \log x_t &= \xi + \phi \log h_t + \tau_1 z_t + \tau_2 (z_t^2 - 1) + u_t, & u_t &\sim N(0, \sigma_u^2). \end{aligned} \tag{9}$$

We estimate (9) by quasi-maximum likelihood and forecast the realized measure directly, using the analytic recursion for $\mathbb{E}[\log h_{t+k}]$ with persistence $\pi = \beta + \gamma\phi$ and a lognormal correction for the accumulated forecast variance.

2.3. 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 ? 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 the jump share J_t/RV_t , log semivariances, normalized signed jumps, log realized quarticity, daily/weekly/monthly returns, log volume growth, the day of the week, the (lagged) log VIX level, and cross-asset averages of log RV within each asset class — fourteen additional predictors, all known at the end of day t .

Learners. We consider penalized linear models — ridge, LASSO (?) and elastic net (?) on standardized predictors — and tree ensembles: random forests (?), XGBoost (?) and LightGBM (?). 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), 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 (?) with 64 hidden units, and a two-block Transformer encoder (?) with model dimension 32 and 4 attention heads. Both are trained with Adam, early stopping on a temporal validation split, and quarterly re-estimation.

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 monthly on the trailing window. The pooled design tests whether volatility dynamics are sufficiently common across assets for cross-sectional information to sharpen asset-level forecasts (??). 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 (?) and permutation importance.

2.4. Forecast evaluation

Rolling protocol. All models are estimated on a rolling window of 1,000 trading days and re-estimated every 22 days; forecasts are produced daily for horizons $h \in \{1, 5, 22\}$, defined as the *average* daily variance over the next h days (cumulative variance equals h times the average). The evaluation sample therefore starts after the first 1,000-plus- h observations of each series and runs through July 2026.

Loss functions. Our primary loss is the QLIKE,

$$L^Q(RV, F) = \frac{RV}{F} - \log \frac{RV}{F} - 1, \quad (10)$$

which together with the MSE is robust to noise in the volatility proxy (?): rankings based on (10) are consistent for the rankings that would obtain under the true conditional variance. QLIKE is also scale-free, which permits averaging across assets; MSE results are reported as a complement.

Statistical tests. Pairwise forecast comparisons use ? tests with Newey–West long-run variances (?) and bandwidth $h - 1$. Joint comparisons use the 90% model confidence set of ? with the $T_{\max}$ statistic and 5,000 stationary bootstrap replications (?). Forecast unbiasedness is assessed with ? levels regressions, $RV_{t+1:t+h} = a + bF_{t,h} + e_{t+h}$, testing $(a, b) = (0, 1)$ with HAC standard errors. Whether machine-learning forecasts contain information absent from HAR is tested with forecast

encompassing regressions (?), $RV = a + b_1 F^{HAR} + b_2 F^{ML} + e$: rejection of $b_2 = 0$ means HAR does not encompass the rival. Finally, we report two forecast combinations (??): the simple mean of all individual models and inverse-MSE weights computed on a trailing 60-day window.

Stability. Because relative performance may drift across regimes, we complement full-sample tests with the fluctuation test of ?, which tracks the standardized DM statistic over a centred rolling window of 30% of the evaluation sample and compares its path to regime-robust critical values.

2.5. Economic value

Volatility timing. For each asset and model we run the volatility-timing strategy of ??: the weight on the risky asset is $w_t = \min(\bar{w}, \sigma^* / \hat{\sigma}_{t+1})$, with annualized target $\sigma^* = 10\%$, leverage cap $\bar{w} = 3$, and $\hat{\sigma}_{t+1}$ the one-day-ahead forecast volatility. Net returns subtract proportional transaction costs of 5 basis points on turnover (0–20 bps in sensitivity checks). We report annualized Sharpe ratios and the ? utility gain — the annualized fee Δ a mean-variance investor with relative risk aversion $\gamma = 5$ would pay to switch from HAR-based to model-based timing.

Value-at-Risk. One-day Gaussian VaR forecasts, $\text{VaR}_{t+1}^q = -z_q \hat{\sigma}_{t+1}^r$, are backtested at the 1% and 5% levels with the unconditional-coverage test of ? and the conditional-coverage test of ?. Because RV omits overnight variance, forecasts are rescaled into close-to-close return units with a trailing 250-day variance ratio, so no future information is used.