Volatility Targeting with GARCH Forecasts:
A Controlled Evaluation of Risk Stabilization and Forecaster Choice

Eugen Soloviov

July 2026

Abstract

Volatility targeting sizes a position inversely to a forecast of its volatility. It is widely credited with improving Sharpe ratios; it is less often asked what part of that credit is risk management and what part is a claim about forecasting skill. We answer both questions on synthetic, seeded data-generating processes with *known* time-varying volatility — a GARCH(1,1) with volatility regime shifts and a deliberately constant per-observation Sharpe drift — so that ground truth is available by construction. Four reproducible experiments follow. First, against a constant full-exposure position on the same path, targeting holds realized annualized volatility at 0.201 versus a 0.20 target (the untargeted position runs at 0.397), cuts the instability of rolling realized risk by a factor of 5.2, and shrinks maximum drawdown from 0.844 to 0.563 — while lifting the Sharpe ratio only modestly, from 0.168 to 0.180. Second, driving the same rule with five one-step forecasters (GARCH(1,1), GJR-t, EWMA(0.94), rolling realized standard deviation, and HAR-RV on an intraday realized-variance proxy), we find the GARCH accuracy edge over EWMA is a QLIKE gap of -0.0006 that a Diebold–Mariano test cannot distinguish from zero ($p = 0.57$), and it reverses sign across regimes (EWMA better in calm, GARCH better in high volatility); the resulting targeted strategies' Sharpe ratios span just 0.037 across all five forecasters. Third, because the drift is constant by design, targeting leaves the raw mean return essentially unchanged (targeted 0.036 versus a matched-exposure fixed 0.037) while improving Sharpe and drawdown; the Moreira–Muir timing alpha is a modest 0.030 annualized and, at this sample length, statistically insignificant ($t = 0.55$). Fourth, a small sensitivity grid over the exposure cap, refit lag, and transaction-cost level, plus a no-look-ahead check in which a one-period-leaked forecast inflates every metric. The deliverable is a calibrated decomposition: volatility targeting is a risk-management transformation — constant risk, tighter drawdowns, largely forecaster-agnostic among reasonable choices — and not, in this controlled setting, a source of alpha.

1 Introduction

A volatility forecast is worth something only when it changes a decision. The cleanest such decision is volatility targeting: hold an exposure $w_t = \sigma_{\text{tgt}}/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is a forecast of next-period volatility and σ_{tgt} is the volatility the strategy is meant to run at, so that the realized volatility of $w_t r_{t+1}$ is approximately σ_{tgt} whenever the forecast tracks the truth. The construction isolates the forecast's value with minimal contamination from a directional view, which is exactly why it is the right test bed — and exactly why it invites two persistent confusions. The first is attributing to *forecasting skill* a benefit that is really *risk management*: the improvement volatility targeting delivers is overwhelmingly about the shape of the risk distribution, not the level of the mean. The second is treating the choice of volatility model as decisive when, among reasonable forecasters, it usually is not.

Both confusions are hard to dispel on real data, because on real data the true conditional volatility is never observed and the true expected return is contested. We therefore work entirely in a controlled setting. We simulate returns from a data-generating process (DGP) whose conditional volatility we know exactly at every step — a GARCH(1,1) volatility recursion (Bollerslev, 1986) with occasional volatility regime shifts — and we give it a per-observation Sharpe ratio that is *constant by construction*. That last choice is deliberate and central: by fixing the risk–return tradeoff, we remove the one channel (time-varying price of risk) through which volatility targeting could genuinely manufacture alpha, so that whatever risk-adjusted improvement survives must be a pure risk-shaping effect. The deliverable is the calibrated decomposition, not a trading strategy: we download no market data, and nothing here should be read as a live-market result.

Against this ground truth we ask four questions, each an experiment (Section 4) run by one seeded harness. **(1) Risk stabilization:** how tightly does targeting hold realized risk at the target, and what does it do to the Sharpe ratio and drawdown, relative to a constant full-exposure position? **(2) Forecaster bake-off:** across GARCH(1,1), GJR-t, EWMA(0.94), a rolling realized standard deviation, and HAR-RV, how large is the forecast-accuracy spread (by QLIKE, with Diebold–Mariano tests), and how much of it survives into downstream targeting quality? **(3) Honest decomposition:** under a constant conditional Sharpe, does targeting move the raw mean return, and how much of the Sharpe change is a Moreira–Muir timing effect? **(4) Sensitivity and causality:** how do the exposure cap, refit lag, and cost level move the results, and how much does a deliberate one-period look-ahead inflate them?

We make no methodological claim of novelty; every estimator here is standard. The contribution is calibration evidence on data where the answer is known: volatility targeting is a risk-management transformation that is largely forecaster-agnostic among reasonable choices and, absent a time-varying price of risk, not a source of mean return. This study accompanies a marketmaker.cc blog post.

2 The data-generating process and ground truth

Every experiment draws from one controlled DGP. Innovations are $\epsilon_t = \sigma_t z_t$ with z_t independent, mean zero and unit variance, and the conditional variance follows a GARCH(1,1) recursion whose baseline level shifts between two hidden regimes:

$$\sigma_t^2 = \omega_t + \alpha\,\epsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2, \qquad r_t = S\,\sigma_t + \epsilon_t, \tag{1}$$

where ω_t equals a baseline ω in the calm regime and $\omega \cdot m^2$ in the high-volatility regime, the hidden state flipping with probability p_{sw} each step. We use $\omega = 10^{-5}$, $\alpha = 0.09$, $\beta = 0.88$ (so the GARCH persistence $\alpha + \beta = 0.97$), regime multiplier $m = 2.0$, and switch probability $p_{\text{sw}} = 0.01$. The drift term $S\,\sigma_t$ is the crux: because $r_t/\sigma_t = S + z_t$, the *per-observation conditional Sharpe ratio is exactly S at every t*, which we fix at $S = 0.05$ (annualized 0.79). The risk–return tradeoff is therefore constant through time by construction — no regime is a better or worse deal per unit of risk — which is precisely what lets Section 5.3 separate risk shaping from alpha.

Two ground-truth objects are retained from the simulator: σ_t itself, and the genuine one-step-ahead conditional volatility $\sqrt{\mathbb{E}[\sigma_{t+1}^2 \mid \mathcal{F}_t]}$, the oracle forecast a perfectly specified model would produce. Over the evaluation window the realized annualized asset volatility is 0.407 with the hidden chain in its high-volatility state 0.423 of the time; the standard deviation of the (annualized) conditional volatility — the vol-of-vol that gives targeting something to do — is 0.116. A less noisy volatility proxy than the squared return is built the way an intraday data set would supply one: within each period we simulate 48 intraday sub-returns whose variances sum to σ_t^2, and take

realized variance RV_t as their squared sum. All streams are seeded; the same command reproduces `results/results.json` byte-for-byte.

3 Forecasters, targeting, and evaluation

Forecasters. Five one-step-ahead volatility forecasters drive the targeting rule, each strictly causal: the forecast usable to size the position held over period t is built only from information available at the close of $t-1$. *EWMA* is the RiskMetrics estimator $\hat{\sigma}_t^2 = \lambda\hat{\sigma}_{t-1}^2 + (1-\lambda)r_{t-1}^2$ with $\lambda = 0.94$ (J.P. Morgan and Reuters, 1996), an IGARCH(1,1) with no free parameters. *Rolling* is the sample variance of the last 40 returns. *GARCH(1,1)* (Bollerslev, 1986) and *GJR-t*, the Glosten–Jagannathan–Runkle leverage model with Student-t innovations (Glosten et al., 1993), are estimated by maximum likelihood with the `arch` package (Sheppard et al., 2024) on a trailing window of 756 observations, refit every 5 periods and advanced by the variance recursion in between. *HAR-RV* is Corsi's heterogeneous autoregression (Corsi, 2009) run in logs on the realized-variance series with a Jensen correction. Every forecaster is warmed up before the common evaluation window, which contains 3244 observations.

Targeting. The exposure is

$$w_t = \min\Big(\frac{\sigma_{\text{tgt}}}{\hat{\sigma}_t},\ w_{\max}\Big), \qquad \text{strategy}_t = w_t\, r_t - c\,|w_t - w_{t-1}|, \tag{2}$$

with a target of 0.20 annualized ($\sigma_{\text{tgt}} = 0.20/\sqrt{252}$ per observation, and we treat 252 observations as one year), an exposure cap $w_{\max} = 3.0$, and a linear transaction cost of 5 basis points per unit of turnover ($c = 5\times10^{-4}$). The position formed at the close of $t-1$ earns r_t; the constant full-exposure benchmark is $w_t = 1$.

Evaluation. Forecast accuracy is scored by the QLIKE loss (Patton, 2011), $\text{QLIKE}(\sigma^2, h) = \sigma^2/h - \log(\sigma^2/h) - 1$, which is robust to noise in the variance proxy, penalizes under-prediction more than over-prediction, and is scale-invariant. We compute it against the RV proxy and, as a controlled cross-check, against the known conditional variance. Differences in predictive accuracy are tested with Diebold–Mariano (Diebold and Mariano, 1995), using a Newey–West long-run variance and the Harvey–Leybourne–Newbold small-sample correction (Harvey et al., 1997); a negative statistic favors the first forecaster. We also report the Mincer–Zarnowitz (Mincer and Zarnowitz, 1969) slope and R^2. Strategy performance is summarized by the annualized Sharpe ratio, the maximum drawdown of the compounded equity, and three risk-constancy metrics against the target: the standard deviation of the rolling annualized volatility (risk instability), the mean absolute distance of that rolling volatility from the target (tracking error), and the gap between full-sample realized volatility and the target.

4 Experimental design

All four experiments share the single seeded path of Section 2 and the forecasters of Section 3, computed once. Experiment 1 (Section 5.1) contrasts the GARCH-driven targeted strategy with the constant full-exposure position. Experiment 2 (Section 5.2) ranks the five forecasters two ways: by forecast accuracy (QLIKE, with Diebold–Mariano between GARCH and EWMA, and a regime-conditional split using the known hidden state) and by the downstream quality of the strategy each one drives. Experiment 3 (Section 5.3) isolates what targeting does to the mean by comparing

Table 1: Experiment 1: a GARCH-driven volatility-targeted strategy versus a constant full-exposure position, on the same seeded path of 3244 observations, 0.20 annualized volatility target, exposure cap 3.0, transaction cost 5 basis points per unit turnover. Volatilities, mean, and drawdown are annualized.

	Vol-targeted	Full exposure
Realized volatility (target 0.20)	0.201	0.397
Realized-vol gap to target	0.001	0.197
Rolling-vol instability (std)	0.020	0.101
Tracking error (mean abs. dist.)	0.015	0.186
Sharpe ratio	0.180	0.168
Maximum drawdown	0.563	0.844
Mean return (annualized)	0.036	0.066
Average gross exposure	0.55	1.00

the targeted strategy against a fixed position matched to its *average* gross exposure, so the two deploy the same capital on average, and decomposes the risk-adjusted gain with a Moreira–Muir volatility-managed regression (Moreira and Muir, 2017). Experiment 4 (Section 5.4) sweeps the exposure cap, the transaction-cost level, and the refit lag, and runs a no-look-ahead check: the same rolling estimator, once strictly causal and once leaked by a single period (its window includes the current return, the very return the exposure earns). Because the leaked forecaster is the only one that violates causality, comparing the two attributes any performance difference entirely to the look-ahead.

5 Results

5.1 Risk stabilization is the headline; the Sharpe gain is modest

Table 1 contrasts the volatility-targeted strategy with a constant full-exposure position on the same path. The targeting mechanic works as advertised: realized annualized volatility lands at 0.201, essentially on the 0.20 target, whereas the untargeted position runs at 0.397 — its realized-volatility-to-target gap is 349 times larger. The instability of risk, measured by the standard deviation of rolling annualized volatility, falls from 0.101 to 0.020, a factor of 5.2; the tracking error (mean absolute distance of rolling volatility from target) falls from 0.186 to 0.015. The drawdown benefit is large and direct: maximum drawdown shrinks from 0.844 to 0.563. All of this is what a risk manager buys.

What it does *not* buy is a large Sharpe improvement. The targeted Sharpe ratio is 0.180 against the full-exposure 0.168 — a real but modest gain of 0.012. This is exactly what the constant-Sharpe DGP predicts. With a constant conditional Sharpe and a perfect forecast, targeting produces a return stream $\sigma_{\text{tgt}}(S + z_{t+1})$ whose conditional Sharpe is again S: the arithmetic Sharpe improves only because holding constant risk removes the vol-of-vol contribution to the denominator, an effect bounded by the dispersion of volatility rather than by any skill. The dramatic numbers in Table 1 are the risk ones; the Sharpe number is deliberately undramatic, and honestly so.

Table 2: Experiment 2: five one-step volatility forecasters, ranked by forecast accuracy (QLIKE against the realized-variance proxy and against the known conditional variance; Mincer–Zarnowitz slope b and R^2) and by the downstream volatility-targeted strategy each drives (annualized Sharpe, realized volatility, annualized turnover). Lower QLIKE is better; an unbiased forecast has $b = 1$.

	Forecast accuracy				Downstream strategy	
Forecaster	QLIKE	QLIKE(true)	MZ b	MZ R^2	Sharpe	Vol
HAR-RV	0.037	0.023	0.98	0.77	0.203	0.199
EWMA(0.94)	0.046	0.034	1.00	0.76	0.211	0.206
GARCH(1,1)	0.046	0.034	0.99	0.77	0.180	0.201
GJR-t	0.052	0.040	0.97	0.75	0.174	0.201
Rolling std (40)	0.084	0.072	0.84	0.54	0.198	0.209

5.2 Forecaster bake-off: the choice barely matters

Table 2 ranks the five forecasters by accuracy and by the strategy each one drives. By QLIKE against the realized-variance proxy, HAR-RV is best (0.037), which is unsurprising — it alone consumes the less noisy intraday proxy — followed by a near-tie among EWMA (0.046), GARCH (0.046), then GJR-t (0.052), with the rolling standard deviation last (0.084). The ranking against the known conditional variance is identical.

The protagonist result is the GARCH-versus-EWMA comparison. Their QLIKE gap is -0.0006 (EWMA fractionally lower), and a Diebold–Mariano test cannot distinguish them: the statistic is 0.56 with a p-value of 0.57. Nor is the tie an average that hides a consistent winner. Splitting the loss by the known hidden regime, GARCH's edge over EWMA is -0.0044 in the calm state (EWMA better) and $+0.0046$ in the high-volatility state (GARCH better) — the sign *reverses*, and the two nearly cancel. GARCH's mean reversion pays off precisely where the blog-level intuition says it should, after a volatility spike that then normalizes, and costs elsewhere; over the full sample it is a wash. The elaboration to GJR-t does not help here: its QLIKE is significantly *worse* than EWMA's (Diebold–Mariano statistic 5.26, $p = 1.5 \times 10^{-7}$), the extra leverage and tail parameters adding estimation noise the symmetric-innovation DGP does not reward. Only HAR-RV is significantly more accurate than GARCH (statistic -6.93, $p = 4.9 \times 10^{-12}$), and that edge is an artifact of its cleaner input, not its functional form.

Downstream, the differences compress further. The five targeted strategies' Sharpe ratios span only 0.037, from EWMA's 0.211 to GJR-t's 0.174, and the accuracy ranking does not even survive into the strategy ranking: EWMA leads on Sharpe, GARCH is fourth, and the QLIKE-best HAR-RV is second. Every forecaster hits the target volatility within a few points (realized volatilities 0.199 to 0.209). The lesson is blunt: among reasonable forecasters, the choice of volatility model is a second-order decision for a volatility target, and the sophisticated models earn their keep, if at all, only through the cleaner data they can exploit, not their dynamics.

5.3 Honest decomposition: risk shaping, not alpha

Experiment 3 asks whether the Sharpe improvement is a repackaged mean return. It is not. Comparing the targeted strategy against a fixed position matched to its average gross exposure — so both deploy the same capital on average — the annualized mean returns are 0.036 (targeted) and 0.037 (fixed), a difference of -0.001 and a ratio of 0.98: essentially identical, in fact fractionally lower after costs. The gross-of-cost targeted mean is 0.039, so the transaction bill is 0.003 annual-

ized. Over the same comparison the Sharpe ratio rises by 0.012 (0.180 versus 0.168) and maximum drawdown improves by 0.065 (0.563 versus 0.628). The entire risk-adjusted gain comes from reshaping the return distribution — constant volatility, thinner drawdowns — with the mean held fixed.

The Moreira–Muir decomposition (Moreira and Muir, 2017) confirms the interpretation and adds an honest nuance. Regressing the volatility-managed portfolio (scaled by $1/\hat{\sigma}_t^2$ to the base's volatility) on the buy-and-hold return gives an annualized timing alpha of 0.030 with a slope of 0.87 and an appraisal ratio of 0.15. Under a constant conditional Sharpe there is no time-varying price of risk to time, so one might expect exactly zero; the small positive value is the mechanical mean–variance benefit of underweighting forecastable high-variance periods, which exists even without a time-varying tradeoff. Crucially, at this sample length it is not statistically significant — the t-statistic is 0.55 — so the honest reading is that volatility targeting reshapes risk and, in this controlled world with the alpha channel switched off by design, does not deliver a detectable mean-return edge.

5.4 Sensitivity and the cost of look-ahead

Table 3 collects the sensitivity grids. The exposure cap is slack at this target: because the asset's volatility (0.407) exceeds the target, typical exposure is around 0.55 and a cap of 3.0 never binds, so Sharpe is flat at 0.180 for any cap of 1.0 or above. The cap becomes a live constraint only when tightened below the typical exposure — at 0.75 it binds 0.10 of the time and pulls realized volatility down to 0.199, and at 0.50 it binds 0.59 of the time, pulling realized volatility to 0.177 and Sharpe to 0.161. The practical reading is that the cap is a tail guard against dead-calm leverage spikes, of which this persistent-volatility DGP produces few; it costs nothing until it starts overriding the target. Transaction costs bite linearly and predictably: at the strategy's 6.4 annual turnover, moving from 0 to 5 to 10 to 20 basis points walks the Sharpe ratio from 0.196 down through 0.180 and 0.164 to 0.132. The refit lag matters least of the three: refitting the GARCH every 5, 20, or 60 periods moves QLIKE only from 0.046 to 0.050 and Sharpe from 0.180 to 0.172, confirming that the recursion carries most of the daily information and frequent re-estimation is a minor refinement.

The no-look-ahead check is the cautionary tale. We take one estimator — the rolling variance — and compute it two ways: strictly causal (window ending at $t-1$) and leaked by a single period (window including r_t, the return the exposure w_t earns). This is the most common backtest bug, an off-by-one in the index, and it is subtle precisely because it looks causal. Yet the leaked version improves *every* metric: QLIKE falls from 0.084 to 0.080 (a ratio of 1.05), tracking error tightens by a factor of 1.15, and the Sharpe ratio rises from 0.198 to 0.205. None of the gain is real — it is entirely the value of knowing r_t when sizing the position that earns r_t. A one-period leak on a 40-period window is about as mild a look-ahead as exists, and it still manufactures free performance across the board; larger leaks scale up accordingly. Strict causality is not a stylistic preference but the difference between a backtest and a fantasy.

6 Discussion

The four experiments compose into a single, deflationary reading of volatility targeting. Its genuine and large benefit is risk stabilization: on our path it converts a 0.397-volatility, 0.844-drawdown position into a 0.201-volatility, 0.563-drawdown one whose rolling risk is 5.2 times steadier, and it does this with the mean return essentially untouched. That combination — constant risk, compressed drawdowns, unchanged mean — is worth a great deal to anyone who must size and hold a position, and it is available regardless of any view on direction. It is also, importantly, *not* an

Table 3: Experiment 4: sensitivity of the GARCH-driven targeted strategy to the exposure cap, transaction-cost level, and GARCH refit lag; and the no-look-ahead check. "Bind" is the fraction of observations at which the cap binds. The look-ahead rows use the same rolling estimator, strictly causal versus leaked by one period (its window includes the current return).

Grid	Setting	Sharpe	Realized vol	Note
Exposure cap	0.50	0.161	0.177	binds 0.59
	0.75	0.169	0.199	binds 0.10
	1.00	0.176	0.200	binds 0.01
	3.00	0.180	0.201	slack
Cost (bps)	0	0.196	0.201	turnover 6.4
	5	0.180	0.201	
	10	0.164	0.201	
	20	0.132	0.201	
Refit lag	5	0.180	0.201	QLIKE 0.046
	20	0.177	0.200	QLIKE 0.047
	60	0.172	0.199	QLIKE 0.050
Look-ahead	Causal rolling	0.198	0.209	QLIKE 0.084
	Leaked rolling	0.205	0.203	QLIKE 0.080

alpha: with the price of risk held constant by construction, the Sharpe improvement is a modest 0.012 and the Moreira–Muir timing alpha is statistically indistinguishable from zero. On real data a time-varying price of risk may add a genuine timing component — that is the Moreira–Muir finding for equity factors — but it is a separate, data-dependent effect, and a controlled study with a constant tradeoff is exactly the instrument that shows how little of targeting's reputation needs it.

The forecaster bake-off is the practical counterpart. The forecasting literature's careful apparatus — QLIKE robustness, Diebold–Mariano tests, Mincer–Zarnowitz diagnostics — earns its place here not by crowning a winner but by refusing to. The GARCH-versus-EWMA difference is statistically zero and regime-contingent in sign; the fancier GJR-t is measurably worse on a symmetric DGP; and the one forecaster that is significantly more accurate, HAR-RV, owes its edge to a cleaner data input rather than a better model. Downstream, a 0.037 Sharpe spread across five forecasters, with the accuracy ranking not even preserved, says that for the specific job of driving a volatility target, model choice among reasonable candidates is second order. The reporting discipline still matters — without the Diebold–Mariano test one would happily crown GARCH on a -0.0006 QLIKE gap — but its main service is to prevent overclaiming.

7 Limitations

- **Synthetic data, by design.** The DGP is a GARCH(1,1) with a two-state volatility regime and Gaussian (or, for one forecaster's fit, Student-t) innovations, chosen so that conditional volatility and the conditional Sharpe are known exactly. Real returns have fatter tails, richer volatility dynamics, and a price of risk that is not constant; our experiments quantify the risk-shaping mechanics precisely but say nothing about the size of any real-market timing alpha. Every number here is a controlled-simulation result, not a market claim.

- **Constant conditional Sharpe is a modeling choice, not a fact.** We fixed it to isolate risk shaping from alpha. It is exactly the assumption under which volatility targeting should *not* raise the mean, so our "not alpha" finding is conditional on it; relaxing it is the obvious next experiment and would reintroduce the Moreira–Muir channel deliberately suppressed here.
- **Single seeded path for the strategy comparisons.** The four experiments share one long simulated path (3244 evaluation observations). The forecast-accuracy tests (QLIKE, Diebold–Mariano) aggregate thousands of observations and are statistically meaningful; the strategy-level Sharpe and drawdown comparisons are single realizations, deliberately so, and their small differences (a 0.012 Sharpe gain, a 0.037 cross-forecaster spread) should be read as calibrated magnitudes on known ground truth, not as significance statements across seeds.
- **Costs and frictions are stylized.** Transaction costs are a linear function of turnover; there is no market impact, no funding cost on leverage, no slippage between the close a signal is computed on and the price it trades at, and no liquidation mechanics at the cap. Each of these makes a live volatility target worse than its backtest, and the no-look-ahead result is a reminder that the gap only ever runs in that direction.

8 Conclusion

On synthetic data with known time-varying volatility and a deliberately constant price of risk, volatility targeting is revealed as what it is: a risk-management transformation. It pins realized volatility to its target (0.201 versus 0.20, against an untargeted 0.397), makes rolling risk 5.2 times steadier, and cuts maximum drawdown from 0.844 to 0.563, all while leaving the mean return essentially unchanged (0.036 versus a matched 0.037) and lifting the Sharpe ratio only modestly (0.180 versus 0.168), with a Moreira–Muir timing alpha that is statistically insignificant at this sample length. The choice of volatility forecaster among GARCH(1,1), GJR-t, EWMA, a rolling standard deviation, and HAR-RV is close to irrelevant for this purpose: the GARCH-versus-EWMA accuracy gap is an insignificant, sign-reversing -0.0006, and the downstream Sharpe ratios span just 0.037. What is not optional is causality: a one-period look-ahead, the mildest of backtest bugs, silently improved every metric we measured. The honest headline is the quiet one — targeting buys constant risk and tame drawdowns, not alpha, and buys them almost regardless of which reasonable forecaster denominates the position.

Reproducibility. All code, tests, and outputs accompany this paper: `scripts/run_all.py` regenerates `results/results.json` from fixed seeds (Python 3.14.6, NumPy 2.5.1); `scripts/check_paper_numbers.py` verifies every numeric claim in this manuscript against that file and fails on any mismatch; `tests/` contains deterministic invariant tests for the estimators and a no-look-ahead invariant test for every forecaster.

References

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. doi: 10.1016/0304-4076(86)90063-1.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009. doi: 10.1093/jjfinec/nbp001.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Lawrence R. Glosten, 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*, 48 (5):1779–1801, 1993. doi: 10.1111/j.1540-6261.1993.tb05128.x.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997. doi: 10.1016/S0169-2070(96)00719-4.

J.P. Morgan and Reuters. Riskmetrics technical document. Technical report, J.P. Morgan/Reuters, New York, 1996. Introduces the exponentially weighted moving-average (EWMA) volatility estimator with decay lambda = 0.94 for daily data, an IGARCH(1,1) with omega = 0 and alpha + beta = 1.

Jacob Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In Jacob Mincer, editor, *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. National Bureau of Economic Research (NBER), 1969.

Alan Moreira and Tyler Muir. Volatility-managed portfolios. *Journal of Finance*, 72(4):1611–1644, 2017. doi: 10.1111/jofi.12513.

Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011. doi: 10.1016/j.jeconom.2010.03.034.

Kevin Sheppard et al. arch: Autoregressive conditional heteroskedasticity (arch) and other tools for financial econometrics (python). Software package, 2024. `https://github.com/bashtage/arch`. Used here for GARCH(1,1) and GJR-GARCH-t maximum-likelihood estimation.