Does GARCH(1,1) Actually Forecast Volatility?
A Controlled Study with Known Ground Truth

Eugen Soloviov

July 2026

Abstract

GARCH(1,1) is the workhorse of applied volatility modeling, but "it forecasts well" is usually asserted on real data where the true conditional variance is never observed. We instead run four seeded, fully reproducible experiments on a synthetic GARCH(1,1) data-generating process (DGP) whose conditional-variance path is known by construction, and calibrate exactly when the model earns its keep and whether the evaluation can be trusted without observing volatility. First, maximum-likelihood estimation is consistent: on a crypto-like DGP ($\alpha = 0.09$, $\beta = 0.90$, persistence 0.99) the persistence estimate converges to the truth (0.977 $\rightarrow$ 0.989 as the sample grows from 500 to 5000) and every parameter's root-mean-square error falls roughly like $1/\sqrt{T}$ (persistence RMSE 0.028 $\rightarrow$ 0.004). Second, in a one-step-ahead forecast contest scored against the *true* conditional variance by the proxy-robust QLIKE loss, the correctly specified GARCH wins at every DGP, beating the best naive competitor (RiskMetrics EWMA or a rolling window) by 83–93%; the advantage is largest at moderate persistence and shrinks toward the near-random regime (best-naive QLIKE 0.0222 $\rightarrow$ 0.0079 as persistence falls from 0.90 to 0.05), and near-IGARCH the RiskMetrics EWMA is the closest competitor because it *is* a persistence-one GARCH. Third, across a persistence grid the persistence itself is pinned down tightly even near the unit root (RMSE 0.032 $\rightarrow$ 0.005 as persistence rises to 0.995), but everything derived from it explodes: the implied volatility half-life runs from 6.6 to 138.3 observations while its estimate becomes wildly dispersed (interquartile range 2.6 $\rightarrow$ 121.4) and the long-run variance grows increasingly downward biased — the near-IGARCH fragility, quantified. Fourth, re-scoring the same forecasters against the noisy squared-return proxy instead of the true variance leaves the ranking intact for the Patton-robust losses: the proxy inflates every QLIKE by a nearly forecaster-independent constant (at the crypto DGP, GARCH's QLIKE moves 0.0018 $\rightarrow$ 1.269 and EWMA's 0.0119 $\rightarrow$ 1.280, preserving the gap) and GARCH is selected under the proxy at all six DGPs, whereas the non-robust MAE agrees on only 82% of noisy draws versus 96% for QLIKE. GARCH's win here is guaranteed by construction — the DGP is GARCH — so the contribution is not that GARCH beats naive forecasters on real crypto, but the calibration of *when* and *by how much* it should, and the demonstration that the evaluation is trustworthy even when true volatility is unobservable. This study accompanies a marketmaker.cc blog post.

1 Introduction

Financial returns are close to unpredictable in sign but strongly predictable in magnitude: large moves cluster, small moves cluster, and the squared returns carry persistent autocorrelation that the raw returns do not (Mandelbrot, 1963). The GARCH(1,1) model of Bollerslev (1986), generalizing the ARCH model of Engle (1982), captures this with three parameters, and has become the default conditional-variance model in risk management. In cryptocurrency markets the fit is if anything sharper, and typically lands in the near-IGARCH regime with estimated persistence close to one (Katsiampa, 2017; Chu et al., 2017).

The standard evidence that GARCH "forecasts volatility well" (Andersen and Bollerslev, 1998; Hansen and Lunde, 2005) is collected on real data, where the object being forecast — the conditional variance σ_t^2 — is latent and never observed. Two questions are therefore hard to answer cleanly on real markets. When exactly does the model beat a naive volatility estimate, and by how much? And can we even trust a forecast comparison when the target is unobservable and must be replaced by a noisy proxy such as the squared return? This paper answers both under controlled conditions, by generating the data from a known GARCH(1,1) so that the true conditional-variance path is available as ground truth.

We make no claim about real crypto. On the contrary: because our DGP *is* GARCH, the correctly specified model is guaranteed to win the forecast contest, and reporting that it wins would be circular. The contribution is the calibration around that fact — the size of the advantage and how it varies with persistence, the consistency of the estimator, the near-IGARCH fragility of every persistence-derived quantity, and, most usefully, a controlled confirmation of the Patton (2011) result that the QLIKE and MSE losses rank forecasters identically whether they are scored against the true variance or against the noisy squared-return proxy. That last result is what licenses trusting a volatility-forecast evaluation on real data at all.

Concretely, we run four experiments (Section 3) on the estimators and forecasters of Section 2:

1. **Parameter recovery.** Simulate a crypto-like GARCH(1,1) and refit by maximum likelihood over 300 seeds at sample sizes $T \in \{500, 1000, 2000, 5000\}$. Bias and RMSE shrink with T and the persistence estimate converges to the true 0.99: MLE consistency, measured.

2. **Forecast contest.** Score one-step-ahead forecasters — the correctly specified GARCH refit, RiskMetrics EWMA ($\lambda = 0.94$), and rolling-window sample variances — against the true conditional variance with MSE and QLIKE, across DGPs from near-IGARCH down to near-random. Quantify the GARCH advantage, its dependence on persistence, and the analytic multi-step mean-reversion of the forecast toward the long-run variance.

3. **Persistence and half-life.** Across a grid of true persistence (0.90 to 0.995), report the estimated persistence, the implied half-life $\ln 0.5 / \ln(\alpha + \beta)$, and how estimation and long-horizon uncertainty blow up as persistence approaches one.

4. **Proxy robustness.** Re-score the contest against the squared-return proxy and confirm (Patton 2011) that the robust losses preserve the ranking, so the evaluation is trustworthy without observing true volatility.

All results derive from one seeded run of one public harness (`scripts/run_all.py`); a companion script (`scripts/check_paper_numbers.py`) verifies every numeric claim in this manuscript against the run's JSON output.

2 Model, estimator, and forecasters

The data-generating process. We simulate GARCH(1,1) with zero conditional mean and Gaussian innovations,

$$r_t = \sigma_t z_t, \quad z_t \sim \text{iid } \mathcal{N}(0,1), \qquad \sigma_t^2 = \omega + \alpha\, r_{t-1}^2 + \beta\, \sigma_{t-1}^2, \tag{1}$$

with $\omega > 0$, $\alpha, \beta \geq 0$, and $\alpha + \beta < 1$. Because we generate the path, the conditional variance σ_t^2 used to draw each r_t is retained as ground truth. Taking unconditional expectations of Eq. (1)

gives the long-run variance, and iterating gives the persistence and half-life,

$$\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}, \qquad \text{persistence} = \alpha+\beta, \qquad h_{1/2} = \frac{\ln 0.5}{\ln(\alpha+\beta)}, \tag{2}$$

the half-life being the number of observations for a variance shock to decay halfway back to $\bar{\sigma}^2$. Our base "crypto-like" DGP fixes $\omega = 0.04$, $\alpha = 0.09$, $\beta = 0.90$, so the persistence is 0.99 (near-IGARCH), the long-run variance is 4.0 (per-observation volatility 2.0), and the true half-life is 69.0 observations (derived from Eq. (2)). A burn-in is discarded so each retained path starts from the stationary distribution.

Estimation. Parameters are estimated by Gaussian maximum likelihood. With $r_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, \sigma_t^2(\theta))$ the log-likelihood is a sum of one-step densities, $\ell(\theta) = -\frac{1}{2}\sum_t [\ln 2\pi + \ln \sigma_t^2(\theta) + r_t^2/\sigma_t^2(\theta)]$, maximized numerically by the `arch` library of Sheppard (2023). The DGP is scaled so returns are $O(1)$, so the $\times 100$ rescaling that daily returns usually require is unnecessary and is switched off.

Forecasters. Each forecaster produces a causal one-step-ahead variance $h_t = \widehat{\mathrm{Var}}(r_t \mid \mathcal{F}_{t-1})$:

- *GARCH* (correctly specified): fit on a training block, then filter $h_t = \hat{\omega} + \hat{\alpha}\, r_{t-1}^2 + \hat{\beta}\, h_{t-1}$ over the test block. With the true parameters this reproduces σ_t^2 exactly after the seed transient; with estimated parameters it carries only estimation error.
- *EWMA / RiskMetrics* (J.P. Morgan/Reuters, 1996), $\lambda = 0.94$: $h_t = (1-\lambda) r_{t-1}^2 + \lambda\, h_{t-1}$. This is the constrained GARCH special case $\omega = 0$, $\alpha = 1-\lambda$, $\beta = \lambda$, with persistence exactly one.
- *Rolling window*, $w \in \{22, 66, 132\}$ observations (about one, three, and six trading months): $h_t = \mathrm{mean}(r_{t-w}^2, \ldots, r_{t-1}^2)$.

Multi-step forecast. Iterating Eq. (1) in conditional expectation, using $\mathbb{E}_T[r_{T+k}^2] = \mathbb{E}_T[\sigma_{T+k}^2]$, gives the analytic term structure

$$\mathbb{E}_T[\sigma_{T+h}^2] = \bar{\sigma}^2 + (\alpha+\beta)^{\,h-1}\big(\sigma_{T+1}^2 - \bar{\sigma}^2\big), \tag{3}$$

the forecast decaying geometrically from today's one-step variance toward the long-run level at rate $(\alpha+\beta)^{\,h-1}$.

Losses. A forecast h is scored against a target X by (Patton, 2011)

$$\mathrm{MSE}(X,h) = \mathbb{E}[(X-h)^2], \quad \mathrm{QLIKE}(X,h) = \mathbb{E}\left[\frac{X}{h} - \ln\frac{X}{h} - 1\right], \quad \mathrm{MAE}(X,h) = \mathbb{E}[|X-h|]. \tag{4}$$

QLIKE is the Bregman divergence generated by $-\ln$: non-negative, uniquely minimized at $h = X$, and dependent only on the ratio X/h (scale free). The target X is the true conditional variance σ_t^2 in Experiments 2–3 and the squared-return proxy r_t^2 in Experiment 4. MSE and QLIKE are the Patton-robust losses (ranking invariant to a conditionally unbiased proxy); MAE, which targets the conditional median, is not and is carried as a contrast.

Table 1: Experiment 1 (parameter recovery): bias and root-mean-square error of the Gaussian-MLE estimates on the crypto-like DGP ($\omega = 0.04$, $\alpha = 0.09$, $\beta = 0.90$, persistence 0.99, true half-life 69.0 observations), over $M = 300$ seeds at each sample size. Errors fall with T; the persistence estimate converges to 0.99.

T	$\widehat{\text{pers.}}$	pers. RMSE	α RMSE	β RMSE	ω RMSE	median $h_{1/2}$
500	0.977	0.028	0.031	0.040	0.078	41.3
1000	0.984	0.013	0.018	0.022	0.031	50.7
2000	0.987	0.0072	0.014	0.014	0.017	56.8
5000	0.989	0.0044	0.0083	0.0092	0.011	63.0

3 Experimental design

All experiments are generated and analyzed by a single seeded Python harness (`scripts/run_all.py`, methods in `scripts/garch.py`) under Python 3.14.6 with NumPy 2.5.1 and `arch` 8.0.0, so every number is bit-reproducible by one command. *Parameter recovery* refits the base DGP over $M = 300$ seeds at each $T \in \{500, 1000, 2000, 5000\}$, reporting the bias, RMSE, and standard deviation of each estimate. *Forecast contest* simulates $M = 200$ paths of length 3000 per DGP, fits GARCH on the first 2000 observations, and scores all forecasters on the final 1000 (so GARCH's parameters never see the test block); it sweeps DGPs of persistence $\{0.99, 0.90, 0.70, 0.40, 0.20, 0.05\}$, holding the long-run variance fixed at 4.0 and α's share of persistence fixed at the base ratio. The multi-step sub-experiment forecasts $h \in \{1, 5, 10, 22\}$ steps ahead from every test origin on the crypto-like DGP over $M = 120$ paths. *Persistence and half-life* refits over $M = 300$ seeds of length 3000 across true persistence $\{0.90, 0.95, 0.97, 0.98, 0.99, 0.995\}$. *Proxy robustness* re-scores the contest against r_t^2 and compares forecaster orderings under the two targets.

4 Results

4.1 Parameter recovery: MLE is consistent

Table 1 reports the estimation error on the crypto-like DGP. Every parameter is recovered with a bias and RMSE that shrink steadily with the sample size, at roughly the $1/\sqrt{T}$ rate a consistent estimator should show: the persistence RMSE falls from 0.028 at $T = 500$ to 0.004 at $T = 5000$, and the persistence estimate itself climbs from 0.977 toward the true 0.99 (reaching 0.989), with a small negative small-sample bias (-0.013 at $T = 500$) that is the well-known downward bias of the persistence estimate near the unit root. The reaction and memory parameters land on their targets ($\hat{\alpha} \to 0.089$, $\hat{\beta} \to 0.900$ at $T = 5000$), and the intercept ω, which is the hardest to pin down in the near-IGARCH regime, carries the largest relative error (RMSE $0.078 \to 0.011$, a persistent small upward bias $0.036 \to 0.003$). The implied half-life, a strongly nonlinear function of the estimates, is correspondingly slow to converge: its median climbs from 41.3 toward the true 69.0, reaching only 63.0 at $T = 5000$ — a first sign of the near-IGARCH fragility that Section 4.3 quantifies.

4.2 Forecast contest: GARCH wins, and by how much

Table 2 scores the one-step-ahead forecasters against the true conditional variance. The correctly specified GARCH minimizes both QLIKE and MSE at every DGP — as it must, since it is the DGP — but the interesting content is the margin and its shape. On the crypto-like DGP (persistence

Table 2: Experiment 2 (forecast contest): one-step-ahead forecast losses against the TRUE conditional variance, averaged over $M = 200$ paths, by DGP persistence (long-run variance held at 4.0). "Best naive" is the lowest-QLIKE non-GARCH forecaster and its QLIKE; the gap is the percentage QLIKE reduction of GARCH over it. GARCH also minimizes MSE at every row.

Persistence	GARCH QLIKE	Best naive	Best-naive QLIKE	QLIKE gap	GARCH MSE
0.99	0.0018	EWMA	0.0119	85.0%	0.219
0.90	0.0016	EWMA	0.0222	92.8%	0.070
0.70	0.0016	rolling (132)	0.0148	89.2%	0.060
0.40	0.0012	rolling (132)	0.0093	87.1%	0.043
0.20	0.0011	rolling (132)	0.0085	87.3%	0.037
0.05	0.0013	rolling (132)	0.0079	82.9%	0.042

0.99) GARCH's QLIKE is 0.0018 against 0.0119 for the best naive competitor, an 85.0% reduction; on MSE the gap is far starker (0.219 versus 1.13 for EWMA), because MSE on the variance level is dominated by the rare large-variance spikes that only a conditional model tracks.

The margin is not monotone in persistence, and the reason is instructive. It peaks at moderate persistence — 92.8% at persistence 0.90, where the best naive QLIKE is worst at 0.0222 — and falls off in both directions. Toward the near-random end the absolute advantage collapses because there is little volatility dynamics left to model: the best-naive QLIKE falls monotonically from 0.0222 at persistence 0.90 to 0.0079 at persistence 0.05, and GARCH's edge over it shrinks with it (absolute QLIKE gap 0.0206 $\rightarrow$ 0.0065). Toward the near-IGARCH end the gap also narrows, to 85.0% at persistence 0.99, but for the opposite reason: there the best naive is the RiskMetrics EWMA, and near a unit root EWMA — a persistence-one GARCH — is a good approximation, so it trails GARCH closely rather than badly. Below persistence 0.70 the best naive is instead the longest rolling window, a near-constant-variance estimate that suits a nearly-static DGP.

Multi-step forecasts mean-revert. Table 3 reports the analytic multi-step forecast of Eq. (3) on the crypto-like DGP. The forecast error against the realized true variance grows with the horizon, from an MSE of 0.58 at $h = 1$ to 12.65 at $h = 22$, as the intervening shocks accumulate. Meanwhile the forecast decays toward the long-run variance 4.0 at the predicted geometric rate $(\alpha + \beta)^{h-1}$: with persistence 0.99 that factor is 0.961 at $h = 5$, 0.914 at $h = 10$, and 0.810 at $h = 22$, and the empirical mean deviation of the forecast from $\bar{\sigma}^2$ tracks it closely (falling to 0.767 of its one-step value by $h = 22$). This is the near-IGARCH signature: even three weeks out, the forecast has reverted less than a fifth of the way to its long-run level, so a short-horizon crypto GARCH forecast is essentially "today's variance, very slowly fading."

4.3 Persistence and half-life: the near-IGARCH fragility

Table 4 sweeps the true persistence and exposes the crypto practitioner's central hazard. The persistence *itself* is estimated tightly, and in fact more tightly as it approaches one: its RMSE falls from 0.032 at persistence 0.90 to 0.005 at 0.995, because a near-unit-root process pins the estimate against the stationarity boundary. But every quantity *derived* from persistence through the $1/(1 - (\alpha + \beta))$ factor becomes treacherous. The volatility half-life $\ln 0.5 / \ln(\alpha + \beta)$ runs from 6.6 observations at persistence 0.90 to 138.3 at 0.995, and its estimate is both downward biased (median 105.5 against the true 138.3 at persistence 0.995) and wildly dispersed: the interquartile range of the estimated half-life explodes from 2.6 observations at persistence 0.90 to 121.4 at 0.995

Table 3: Multi-step forecast on the crypto-like DGP ($M = 120$ paths). MSE is against the realized true variance h steps ahead; the analytic decay is $(\alpha+\beta)^{h-1}$ with persistence 0.99; the deviation ratio is the mean $|\text{forecast} - \bar{\sigma}^2|$ relative to its one-step value, confirming the geometric mean-reversion toward the long-run variance 4.0.

Horizon h	Forecast MSE	Analytic decay $(\alpha+\beta)^{h-1}$	Deviation ratio
1	0.58	1.000	1.000
5	3.22	0.961	0.947
10	6.40	0.914	0.887
22	12.65	0.810	0.767

Table 4: Experiment 3 (persistence and half-life): estimation across the true persistence grid ($M = 300$ seeds, $T = 3000$, long-run variance 4.0). The persistence is estimated ever more tightly toward the unit root, but the implied half-life $\ln 0.5/\ln(\alpha+\beta)$ and long-run variance become severely dispersed and biased — the near-IGARCH fragility.

Persistence	$\widehat{\text{pers.}}$	pers. RMSE	true $h_{1/2}$	median $\hat{h}_{1/2}$	$\hat{h}_{1/2}$ IQR	median $\widehat{\bar{\sigma}^2}$
0.900	0.896	0.032	6.6	6.7	2.6	3.97
0.950	0.947	0.015	13.5	13.4	5.1	4.01
0.970	0.968	0.011	22.8	21.8	10.1	3.97
0.980	0.978	0.0080	34.3	32.4	14.1	3.90
0.990	0.988	0.0061	69.0	62.4	34.9	3.79
0.995	0.993	0.0053	138.3	105.5	121.4	3.56

— nearly as large as the quantity itself. The long-run variance, whose sensitivity to persistence is $\bar{\sigma}^2/(1-(\alpha+\beta))$ and thus grows from 40 to 800 across the grid, drifts increasingly below its true value of 4.0 (median 3.56 at persistence 0.995). The lesson matches the folk warning: in the near-IGARCH regime the point estimate of long-run volatility and half-life should never be trusted without an interval, even though the persistence that generates them looks precise.

4.4 Proxy robustness: the evaluation survives an unobservable target

Real volatility is never observed, so any forecast comparison on live data must score against a proxy. The squared return r_t^2 is a conditionally unbiased but extremely noisy one — $r_t^2 = \sigma_t^2 z_t^2$ with $\mathbb{E}[z_t^2 \mid \mathcal{F}_{t-1}] = 1$ — and Experiment 4 asks whether it still ranks the forecasters correctly. Table 5 re-scores the crypto-like contest against r_t^2. The effect of the proxy is dramatic in level but not in order: it inflates every QLIKE by a nearly forecaster-independent constant of about 1.267 (the mean of the proxy's own noise entropy), moving GARCH from 0.0018 to 1.269 and EWMA from 0.0119 to 1.279, so the gap between them barely changes (0.0101 against true variance, 0.0108 against the proxy). The entire ordering is preserved, and GARCH remains the selected forecaster; across all six DGPs GARCH is chosen under the proxy exactly as under the true variance (6 of 6), for both robust losses.

The robustness is a property of the loss, not luck. Repeating the comparison seed by seed on the noisy proxy — where finite-sample noise can flip the ordering of close competitors — the Patton-robust losses agree with the true-variance ordering on 96.2% (QLIKE) and 95.2% (MSE) of the 12000 forecaster-pair comparisons, while the non-robust MAE agrees on only 81.5%. The

Table 5: Experiment 4 (proxy robustness), crypto-like DGP: one-step-ahead QLIKE scored against the true conditional variance versus the squared-return proxy. The proxy adds a nearly forecaster-independent offset (≈ 1.267) but preserves the ordering; GARCH stays first. Across all six DGPs GARCH is the QLIKE-selected forecaster under the proxy as under the truth.

Forecaster	QLIKE vs. true variance	QLIKE vs. r_t^2 proxy
GARCH	0.0018	1.269
EWMA	0.0119	1.279
rolling (22)	0.0290	1.297
rolling (66)	0.0738	1.342
rolling (132)	0.1365	1.407

correctly specified model is dominant enough here that even MAE picks it as the overall winner, but MAE is measurably less reliable on the pairwise orderings, exactly as its lack of proxy-robustness predicts. The practical consequence is the one that makes volatility-forecast evaluation possible at all: scored with QLIKE or MSE, a forecast comparison run on observable squared returns reaches the same verdict it would reach against the latent variance nobody can see.

5 Discussion

What GARCH does and does not buy you. Under a GARCH DGP the correctly specified model wins the one-step contest at every persistence, but the margin over a good naive estimator is not uniform. It is largest where the volatility is genuinely dynamic yet the naive tools are misspecified (moderate persistence), and it shrinks at both extremes: toward a static DGP because there is little to forecast, and toward the near-IGARCH regime because there the RiskMetrics EWMA is itself nearly a GARCH and closes most of the gap. This is the honest calibration behind the workhorse status of GARCH(1,1): the model earns the most where the data have exploitable but non-trivial dynamics, and in the near-unit-root regime typical of crypto a well-tuned EWMA is a respectable one-parameter stand-in for one-step forecasting — consistent with the classic finding that little beats GARCH(1,1) but that the simplest conditional models already capture most of the forecastable variance (Hansen and Lunde, 2005; Andersen and Bollerslev, 1998).

Persistence is easy; everything downstream is not. The near-IGARCH results sharpen a warning that is easy to state and easy to ignore. The persistence is the best-identified feature of a high-persistence GARCH fit, tightening as it approaches one. But the half-life and the long-run variance are $1/(1-(\alpha+\beta))$-type functionals, and a persistence known to ± 0.005 still leaves the half-life uncertain by tens of observations and the long-run variance uncertain by a large factor. Multi-step forecasts and any volatility-targeting rule that leans on the long-run level inherit that fragility; the safe move is to treat the near-IGARCH long-run variance as an interval, refit on rolling windows, and be alert that apparent near-integration can be an artifact of structural breaks rather than genuine persistence.

Why the evaluation can be trusted. The most portable result is the proxy-robustness of Experiment 4. It is what allows the calibration in this paper — all of it obtained against a latent variance we happen to know because we simulated it — to transfer to real data, where the same QLIKE or MSE comparison can be run against observable squared returns and will rank forecasters

the same way. Using a non-robust loss such as MAE forfeits this guarantee; our seed-by-seed comparison shows it degrading measurably even in a setting where the correct model is otherwise dominant.

6 Limitations

- **Synthetic GARCH DGP, favoring GARCH by construction.** The DGP is a Gaussian GARCH(1,1), so the correctly specified model is guaranteed to win the forecast contest; we never claim GARCH beats naive forecasters on real crypto. The contribution is the calibration of the margin and its shape, the consistency and near-IGARCH fragility results, and the proxy-robustness demonstration — all of which require known ground truth and none of which is a market claim.
- **Gaussian innovations, symmetric variance.** Real crypto returns have heavy tails and an asymmetric (leverage) response that plain Gaussian GARCH(1,1) does not model; those are deliberately out of scope here and are the subject of the companion work on Student-t/skewed innovations and GJR/EGARCH (Nelson, 1991). Our experiments say nothing about model *misspecification*, only about behavior under correct specification.
- **One base parameterization per experiment.** The forecast contest varies persistence but holds the long-run variance and the α-share fixed; the recovery and half-life grids fix the remaining parameters. We did not sweep the full parameter space, the innovation distribution, or the train/test split, so the quantitative margins are specific to these configurations.
- **One-step focus; forecasters are fixed-parameter.** The contest is primarily one-step-ahead, and GARCH is fit once on the training block rather than refit at every step; a fully rolling refit and a broader multi-horizon loss study would sharpen the multi-step numbers.
- **Proxy robustness shown for one proxy.** We use the squared return, the canonical conditionally-unbiased daily proxy; realized-variance proxies from intraday data are more precise and would compress the level offset of Experiment 4, though Patton's ranking-robustness holds for any conditionally unbiased proxy.

7 Conclusion

We asked whether GARCH(1,1) actually forecasts volatility, and answered it where the answer can be checked: on a synthetic GARCH DGP with a known conditional-variance path. Maximum likelihood recovers the parameters consistently, the persistence estimate converging to the true 0.99 as the RMSE falls like $1/\sqrt{T}$. Scored against the true variance by the proxy-robust QLIKE, the correctly specified model beats the best naive forecaster by 83–93%, with the margin largest at moderate persistence and narrowing toward both the static and the near-IGARCH extremes — and near the unit root a RiskMetrics EWMA is its closest competitor because it is a persistence-one GARCH. The persistence is pinned down tightly even near integration, but the half-life and long-run variance derived from it become severely dispersed and biased, quantifying the near-IGARCH fragility that should make any crypto volatility-targeting rule cautious. Finally, re-scoring against the noisy squared-return proxy leaves the QLIKE and MSE rankings intact — the proxy shifts every loss by a near-constant and preserves the order — so the evaluation is trustworthy even though true volatility is never observed. GARCH's victory here is built in; what is not built in, and what

transfers, is the calibration of when it matters and the confirmation that we can measure it without seeing the target.

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, `arch` 8.0.0); `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 DGP, estimator, forecasters, and losses.

References

Torben G. Andersen and Tim Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998. doi: 10.2307/2527343.

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

Jeffrey Chu, Stephen Chan, Saralees Nadarajah, and Joerg Osterrieder. GARCH modelling of cryptocurrencies. *Journal of Risk and Financial Management*, 10(4):17, 2017. doi: 10.3390/jrfm10040017.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. doi: 10.2307/1912773.

Peter R. Hansen and Asger Lunde. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889, 2005. doi: 10.1002/jae.800.

J.P. Morgan/Reuters. RiskMetrics technical document. Technical Report Fourth Edition, J.P. Morgan/Reuters, New York, 1996. Defines the exponentially weighted moving average (EWMA) variance estimator with decay $\lambda = 0.94$ for daily data.

Paraskevi Katsiampa. Volatility estimation for bitcoin: A comparison of GARCH models. *Economics Letters*, 158:3–6, 2017. doi: 10.1016/j.econlet.2017.06.023.

Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963. doi: 10.1086/294632.

Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991. doi: 10.2307/2938260.

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. Characterizes the loss functions (including MSE and QLIKE) whose forecast ranking is robust to replacing the latent variance by a conditionally unbiased proxy such as the squared return.

Kevin Sheppard. arch: Autoregressive conditional heteroskedasticity (ARCH) and other tools for financial econometrics in Python. `https://github.com/bashtage/arch`, 2023. Python package; version 8.0.0 used here. DOI via Zenodo: `https://doi.org/10.5281/zenodo.593254`.