Silent Model Failure: Unstable GARCH Value-at-Risk Estimation for Short-History Cryptocurrency Assets

Al Jaber – CUNY The City College of New York – AL.JABER62@login.cuny.edu

Yeasin Arafat Riyad – CUNY The City College of New York - YEASINARAFAT.RIYAD79@login.cuny.edu

Abstract

New cryptocurrencies reach the market almost weekly, and within days investors and trading desks need a number that tells them how much they might lose overnight. The usual way of producing that number, a GARCH model feeding a Value-at-Risk calculation, was designed for assets with years of price history behind them. This paper asks what happens when the same machinery is applied to an asset only a few months old, and the answer is uncomfortable. Working with three young coins over a shared one-year test period and steadily cutting the estimation sample from two years down to two months, we find that once the sample falls below roughly a quarter of a year the fitted model tends to settle into a corner of its parameter space where the volatility it forecasts drifts downward and quietly understates the true risk. In one case this leaves a ninety-nine percent Value-at-Risk that is breached on almost a quarter of all days, and the estimation routine gives no hint of trouble: it converges normally, and the risk figure it returns looks entirely ordinary. What makes the result more than a cautionary anecdote is that the other two coins, fitted in the same way over the same period, settle into different corners and remain well calibrated. Nothing an analyst can see at the moment of estimation separates the dangerous case from the safe ones. We trace the behavior to well-documented small-sample properties of GARCH estimation, show that classical and quantum neural forecasters step around this particular trap without being dependable substitutes in their own right, and close with a simple diagnostic that flags the degenerate fit before it is trusted.

Keywords: Value-at-Risk; Expected Shortfall; GARCH; model risk; cryptocurrency; backtesting.

1. Introduction

A volatility forecast earns its keep in finance only when it becomes a risk number that behaves as promised. The number that matters in practice is Value-at-Risk, and it is judged not by how closely a forecast tracks realized volatility on an average day but by calibration over many days: the loss should exceed the reported threshold about as often as the model claims, and those exceedances should not bunch together in time. Kupiec (1995) and Christoffersen (1998) gave the profession the tests that make this judgement operational, and the distinction they encode is the one that matters here, since a model can describe the general level of volatility well and still misjudge the tail badly enough to be useless for setting capital.

GARCH, introduced by Bollerslev (1986) as a generalization of Engle's (1982) ARCH process, remains the reference model for this purpose. It is compact, it produces conditional variances that usually behave out of sample, and decades of use have made it the natural first choice. All of that, though, depends on having estimated its parameters well, and estimation depends in turn on a return history long enough to identify them. In established markets that requirement is invisible because it is always met. For young cryptocurrencies it is the whole difficulty. A token that began trading a few months ago has no long record to offer, and yet these are often violently volatile instruments whose owners have every reason to want a dependable estimate of what they stand to lose.

That GARCH estimation misbehaves on small samples is not itself news to econometricians. Hwang and Valls Pereira (2006) showed that the maximum-likelihood estimates of GARCH(1,1) are badly biased when the sample is short and that a convergent fit respecting the usual non-negativity restrictions often cannot be obtained at all, and they proposed five hundred observations as a floor for trustworthy estimation. Hillebrand (2005) established a companion result, that a change in the level of volatility within the sample drives the estimated persistence toward one. These papers describe how the estimator behaves. They stop short of the question a risk manager actually faces, which is what that behavior does once it is carried through to a regulatory risk figure and tested on a real asset that happens to be young. Does the number fail, and if it fails, does anything warn you?

Our findings say that it can fail badly and that nothing warns you. Across three young coins, shrinking the estimation window drives the fit, below about three months of data, into precisely the degenerate region the small-sample theory describes. On one of the three that degeneracy translates into a ninety-nine percent Value-at-Risk violated on nearly a quarter of test days, while the estimator reports a clean, converged solution and the risk figure looks unexceptional on the page. The failure is silent in exactly that sense. It is also, and this is the part we regard as most important, unpredictable from one asset to the next: the other two coins drift into different degenerate corners that leave them well calibrated, and no observable feature of the fit marks out the coin that is about to fail.

The contribution here is not a new forecasting model but a diagnosis. We connect a settled econometric result about small-sample estimation to a consequence the applied crypto-risk literature has not examined, because that literature works almost entirely with long records of Bitcoin and Ethereum (Troster et al., 2019; Trucios, 2019; Trucios and Taylor, 2023). We show the failure is real and mechanical, that it is silent and asset-specific, and that switching to a learning model avoids this particular pitfall without removing the underlying fragility. We close with something a practitioner can act on: a short list of parameter signatures that identify the degenerate fit before its risk number is believed. The message that runs through the paper is that a converged model and a reasonable-looking figure are, on a short-history asset, no evidence at all that the risk estimate can be trusted.

2. Related Work

The apparatus for judging a risk model is mature. Kupiec's (1995) proportion-of-failures test and Christoffersen's (1998) conditional-coverage and independence tests remain the standard means of validating Value-at-Risk, and the regulatory shift toward Expected Shortfall has brought dedicated

backtests of its own, most prominently that of Acerbi and Szekely (2014), which asks whether the losses seen on breach days are consistent with the shortfall the model predicted. These are the instruments we use, and they define what trustworthiness means in the pages that follow. The forecasting target itself, a one-day volatility proxy built from the absolute return, follows the practice discussed by Patton (2011), whose treatment of imperfect volatility proxies motivates the debiased, annualized measure adopted in Section 3.

On the modelling side, GARCH and its descendants continue to anchor applied risk work, but their behaviour on short samples is understood to be delicate. Hwang and Valls Pereira (2006) document heavy small-sample bias and outright convergence failures in GARCH(1,1), and Hillebrand (2005) shows that unmodelled level shifts push estimated persistence toward unity in finite samples. These two results are the theoretical spine of what we report. Our part is to show that the degenerate fits they anticipate are not a statistical footnote but a working hazard once a risk number depends on them.

Cryptocurrencies have drawn a large and still-growing risk literature, driven by the recognition that their returns carry heavier tails and shift regime more often than conventional assets. Troster et al. (2019) compare heavy-tailed GARCH and generalized autoregressive score models for Bitcoin and find that the score models give the best tail coverage, while Trucios (2019) reports that only a robust, bootstrap-based procedure yields reliable ninety-nine percent forecasts for Bitcoin. Reviewing a broad set of advanced methods on recent Bitcoin and Ethereum data, Trucios and Taylor (2023) find that long-memory, extreme-event and multi-regime procedures outperform standard ones, though no single method dominates. In a large-scale backtesting study, Alexander and Dakos (2023) go further and show that much simpler exponentially weighted models match GARCH for crypto Value-at-Risk and Expected Shortfall once asymmetry and heavy tails are captured. Closest to our concern, Mappadang et al. (2024) observe directly that GARCH models require a substantial history to calibrate and that newer cryptocurrencies offer too little data for robust parameter estimates, and they turn to a simpler exponentially weighted specification in response. What runs through this body of work is a common setting, long histories of the largest coins, and a common question, which model forecasts best. The present study departs on both counts. It asks not which model wins on abundant data but whether the standard model can be trusted at all on the short samples that define a young asset, and whether its failure, when it comes, is visible in time to matter.

3. Data and Methodology

3.1 Data

The study uses three young cryptocurrencies from different ecosystems and launches cohorts: Solana, Avalanche, and Polkadot. All three began trading in 2020, recent enough to let us build genuinely short estimation windows while still holding back a full year for testing. Daily prices come from a public reference dataset, from which we compute daily logarithmic returns, and a rolling twenty-one-day annualized realized-volatility series used as a model input. Because these markets trade every day of the year, annualization uses 365 days rather than the 252 customaries for equities. We screened each series for missing values, non-positive prices, duplicate dates, and

calendar gaps and found none. The handful of extreme daily moves that remain, such as the drop during the collapse of FTX in November 2022, were checked against the historical record and kept, since removing genuine crashes would quietly bias a study whose subject is tail risk. The test window in every case is the most recent 365 days, held identical across the three assets so that no result depends on one coin having been evaluated in a calmer year than another.

3.2 Forecasting target and models

Because next-day volatility cannot be observed directly, every model is scored against the same proxy: the absolute daily return, scaled by the square root of pi over two so that it is an unbiased estimate of the daily standard deviation under approximately Gaussian returns, then annualized. Holding this target fixed across models is what keeps the comparison fair, since it denies any model the advantage of an easier objective. Three models forecast it. The first is a constant-mean GARCH(1,1) with Gaussian innovations, estimated by maximum likelihood on the training window and rolled forward through the test window to yield one-day-ahead conditional volatilities. This is the benchmark whose behavior the paper is chiefly about. The second is a classical long short-term memory network, included as a purely neural control that tells us whether anything we observe is specific to a given architecture. The third is a quantum long short-term memory network, in which each gate of the recurrent cell is realized by a small variational quantum circuit rather than a linear map. It is included because quantum models have been argued, on the strength of generalization bounds such as those of Caro et al. (2022), to learn effectively from limited data, which is precisely the regime at issue here. Both neural models are trained only where the window provides enough sequences to fit them with any confidence, and where it does not, they abstain rather than return a number that would not mean anything.

3.3 From forecasts to risk measures and backtests

Each annualized volatility forecast becomes a one-day Value-at-Risk at the ninety-five and ninety-nine percent levels under a Gaussian mapping, with a Student-t mapping whose degrees of freedom are estimated on the training data alone kept as a robustness check. A day counts as a breach when the realized loss runs past the reported threshold. We assess coverage with the Kupiec (1995) test, which asks whether breaches occur at the promised rate, and the Christoffersen (1998) test, which asks in addition whether they arrive independently, treating a model as passing at a given level only when both p-values clear 0.05. Expected Shortfall is checked with the test of Acerbi and Szekely (2014). Every scaling constant, and the estimated degrees of freedom for the Student-t mapping, is computed on the training portion alone, and each forecast's input window ends the day before the day being predicted, so no future information reaches any forecast.

3.4 Experimental design

We hold the test window fixed and shrink the estimation window through 730, 365, 182, 91, and 60 days, from about two years of history down to two months. Following calibration across these lengths isolates the single variable of interest, sample size, and it does so cleanly: the models, the target, and the backtests are identical at every window, so whatever changes must be attributed to how much history was available and to nothing else.

4. Results

4.1 A failure that gives no warning

Given enough history, the models do what they should. From 182 days of training upward, GARCH is well calibrated on all three coins, its ninety-nine percent breach rate hovering near the promised one percent with backtest p-values that raise no concern. The trouble begins as the window contracts. On Solana the ninety-nine percent Value-at-Risk is breached on 12.9 percent of test days when the model is fitted on 91 days of data, and on 24.7 percent when it is fitted on 60, against a target of one percent. The Kupiec and Christoffersen tests both reject well below the thousandth percentile, and the Expected Shortfall test registers a heavy under-statement of losses at the shortest window. None of this is flagged anywhere in the estimation output. The routine converges, and the volatility it hands back looks like a perfectly reasonable number. Figure 1 plots the ninety-nine percent breach rate against the length of the estimation window for all three coins, and its shape tells the story briefly: only Solana leaves the calibrated band, and when it leaves it does so steeply.

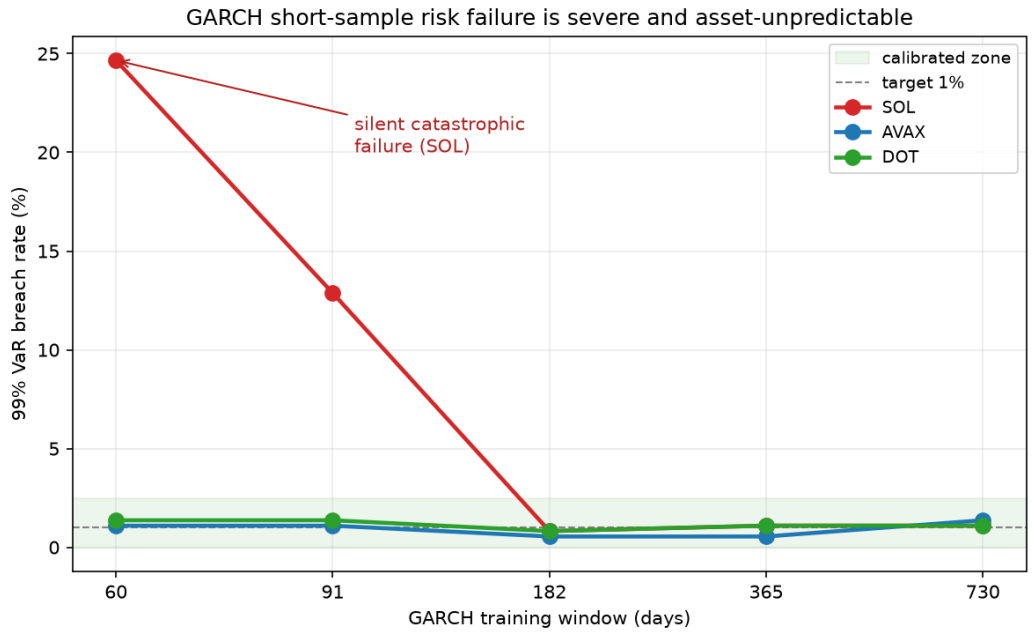

Figure 1. Ninety-nine percent Value-at-Risk breach rate against the length of the GARCH estimation window for the three young coins. The shaded band marks the region consistent with the one percent target. Solana leaves it sharply once the window falls below three months, while Avalanche and Polkadot stay inside it throughout, so the short-sample failure is severe on one asset yet absent on the other two.

4.2 Where the failure comes from

The breakdown is a degenerate maximum-likelihood solution rather than a numerical accident, and it is the very kind the small-sample literature leads one to expect. Table 1 lists the fitted GARCH parameters for Solana at each window. At 60 and 91 days the estimate falls into a corner where the reaction coefficient is zero, the persistence coefficient sits just below one, and the constant is effectively nil. A model of that shape says tomorrow's variance is little more than today's, faded slightly, with nothing to react to a fresh shock and no floor beneath it, so its forecast volatility slides

toward a level about a third of what the test period delivered. The Value-at-Risk threshold shrinks in step, and the losses walk straight through it. This is precisely the near-unit-root persistence Hillebrand (2005) attributes to a level shift inside a short sample, arriving together with the boundary solution and convergence difficulty documented by Hwang and Valls Pereira (2006). Once the window reaches 182 days the fit returns to a sensible interior for Solana, the reaction coefficient becomes meaningful again and the constant recovery, and calibration recovers with it.

Window (days)	**alpha**	**beta**	**omega**	**Persistence (alpha+beta)**	**99% breach rate**
60	0.000	0.987	0.000	0.987	24.7% (degenerate)
91	0.000	0.974	0.067	0.974	12.9% (degenerate)
182	0.095	0.729	4.438	0.825	0.8% (stable)
365	0.090	0.776	2.735	0.866	1.1% (stable)
730	0.084	0.780	2.922	0.864	1.1% (stable)

Table 1. Fitted GARCH(1,1) parameters for Solana across estimation windows. The two shortest windows settle into a degenerate corner, with the reaction coefficient at zero, persistence near one, and the constant near zero, and it is there that the risk figure fails. The longer windows recover a healthy interior fit. The routine reports convergence in every row, so nothing in the output distinguishes the failing fits from the sound ones.

Table 1 shows the mechanism on the asset where it turns catastrophic, but the instability itself is broader, and seeing the full picture is what makes the danger concrete. Table 2 reports the fitted parameters for all three assets across every window, flagging each fit as degenerate when a coefficient reaches a boundary or estimated persistence sits against one. Two features stand out. First, degeneracy is common rather than exceptional: ten of the fifteen fits are degenerate, yet only two of them, both on Solana at the shortest windows, produce a failing risk number. The other eight degenerate fits stay calibrated by luck of which corner they fall into, Avalanche drifting to a boundary that retains a variance floor and Polkadot to one that does not collapse the forecast. Second, and more uncomfortable for anyone relying on a convergence check, degeneracy is not confined to short windows: Avalanche produces degenerate fits even at 365 and 730 days and passes its backtests regardless. Convergence and a clean backtest, then, are not evidence of a healthy fit. They can coincide with a badly degenerate one that simply happened not to bite. This is the heart of why the failure cannot be foreseen from the estimation output alone.

Asset	**Window**	**alpha**	**beta**	**omega**	**Persistence**	**99% breach / fit**
SOL	60	0.000	0.987	0.000	0.987	24.7% degenerate
SOL	91	0.000	0.974	0.067	0.974	12.9% degenerate
SOL	182	0.095	0.729	4.438	0.825	0.8% stable
SOL	365	0.090	0.776	2.735	0.866	1.1% stable
SOL	730	0.084	0.780	2.922	0.864	1.1% stable

AVAX	60	0.000	0.498	12.577	0.498	1.1% degenerate
AVAX	91	0.000	0.976	0.462	0.976	1.1% degenerate
AVAX	182	0.066	0.075	25.067	0.140	0.5% stable
AVAX	365	0.000	0.991	0.270	0.991	0.5% degenerate
AVAX	730	0.062	0.922	0.433	0.984	1.4% degenerate
DOT	60	0.168	0.000	13.598	0.168	1.4% degenerate
DOT	91	0.195	0.000	14.632	0.195	1.4% degenerate
DOT	182	0.000	0.961	0.697	0.961	0.8% degenerate
DOT	365	0.075	0.896	0.633	0.972	1.1% stable
DOT	730	0.080	0.910	0.247	0.990	1.1% degenerate

Table 2. Fitted GARCH(1,1) parameters for all three assets across every estimation window, with each fit marked degenerate where a coefficient reaches a boundary or estimated persistence sits against one. Ten of the fifteen fits are degenerate, but only the two Solana short-window fits produce a failing risk number. The remaining degenerate fits, including several at long windows, stay calibrated. A converged fit that passes its backtest is therefore not necessarily a healthy one.

4.3 Why the danger is that it cannot be foreseen

The finding we regard as most consequential is that this degeneracy obeys no tidy threshold rule of the form GARCH stops working below some number of days. As Table 2 makes plain, Avalanche and Polkadot also produce degenerate short-window fits, but they land in different corners than Solana, and in both cases the Value-at-Risk that results stays calibrated. Fitted on just 60 days, GARCH breaches on 1.1 percent of days for Avalanche and on 1.4 percent for Polkadot, both passing every test comfortably, while the same procedure on Solana breaches on 24.7 percent. The three fits converge alike and read alike on paper. No convergence flag, no visibly absurd parameter, and no feature of the reported risk number set the catastrophic fit apart from the two harmless ones. The hazard, then, is not that GARCH always breaks on short samples, since two coins out of three shows plainly that it need not, but that whether it breaks is unknowable at the moment one would need to know it, which from a risk manager's chair is the worse of the two problems.

4.4 Do learning models help

A natural response is to reach for a learning model, and there is something to that instinct, though not enough to settle the matter. The classical and quantum recurrent networks do not inherit GARCH's degenerate-fit failure, because they do not estimate a small parametric recursion by maximum likelihood, and their forecast volatility does not slide toward zero on short samples. On Solana at the windows where GARCH is catastrophic, both stayed calibrated, and the quantum

network in particular produced a passing ninety-nine percent Value-at-Risk at 60 days, a window at which the classical network had too few sequences to train. That is a genuine advantage for data-efficient models and worth recording. It does not make them safe, however. Their calibration is uneven across the coins at the shortest windows: the quantum model fails on Polkadot at 60 days, breaching on 2.2 percent, the classical model fails on Polkadot at 91 days, breaching on 3.3 percent, and the classical model fails Solana's independence test at 182 days through clustered breaches. Read honestly, the learning models step around the specific trap that catches GARCH while bringing a short-sample fragility of their own, and none of the three models we study is dependable across all three coins once the window falls to two or three months.

5. A Practical Diagnostic

If failure cannot be predicted from the length of the sample and does not announce itself in the convergence status, the practical question is whether it can be caught by looking at the fit itself. It can, at least in the cases we observe, and the signature is specific enough to be useful. Every catastrophic fit in our study shares the same three marks: a reaction coefficient driven to its zero boundary, an estimated persistence sitting against one, and a constant collapse to near zero. A fit carrying all three has a forecast variance with no floor and no capacity to respond to news, and it is exactly such a fit that produced the ninety-nine percent Value-at-Risk breached on a quarter of day. A workable screen therefore rejects, or at least quarantines for direct backtesting, any short-sample GARCH fit in which the reaction coefficient or the persistence term has reached a boundary, and the implied unconditional variance has grown implausibly large or small relative to the sample. This does not repair the model, and it will occasionally flag a fit that would have been fine, but on short-history assets an over-cautious screen is the correct trade, because the cost of trusting the silent failure is a risk figure wrong by more than an order of magnitude at exactly the moment capital depends on it.

6. Discussion

Three lessons follow, and each is meant for use rather than for the file. The first is that a converged model is not a validated one. In every degenerate case we saw, the estimator reported a normal solution, and an analyst who read only the convergence status and the resulting figure would have had no reason for concern. On a short-history asset a converged GARCH fit should be held as provisional until its parameters have been inspected and its risk number backtested against whatever out-of-sample data can be found. The second is that real exposure is unpredictability. Because one estimation procedure yields disaster on one coin and safety on two others with nothing to tell them apart, no analyst should lean on GARCH Value-at-Risk for an individual young asset without checking that asset on its own terms. A rule that held on the last coin offers false comfort on the next. The third is that changing models helps at the margin but does not dissolve the problem, since the learning models trade one kind of short-sample fragility for another. The safest stance on very short histories is not to swap in a favored model but to withhold trust from any single one until the evidence of a backtest is in hand.

7. Conclusion

Risk numbers for young assets are produced every day with tools designed for old ones, and this paper has shown what that can quietly cost. On short samples the GARCH fit tends to degenerate in the manner the econometric literature has long described, and we have found at least one real case where the degeneracy leaves a ninety-nine percent Value-at-Risk failing on nearly a quarter of days while the model reports a clean fit and says nothing is wrong. That the same procedure stays reliable on two comparable coins, with no signal to separate the safe from the dangerous, is what turns an isolated failure into a general warning. Learning models avoid this specific trap but carry fragilities of their own and so offer no blanket remedy. The lesson is modest and practical. On a young asset, a model that has converged and a figure that looks reasonable prove nothing about whether the risk has been measured, and the only sound response is to check the fit for the degenerate signature described here and to verify the number against whatever history exists before anyone relies on it.

8. Limitations and Future Work

A few boundaries frame these conclusions. The study rests on three assets, one fixed test window, daily data, and quantum circuits run in simulation rather than on hardware. The catastrophic case is shown directly on a single coin, and the wider claim we make is exactly that such cases are unpredictable rather than common, which a broader cross-section of young assets and several rolling test periods would let us quantify more precisely. It would also be worth asking whether the degeneracy can be regularized away, through a variance floor, a prior on persistence, or an alternative specification such as the exponential GARCH of Nelson (1991), whose formulation removes the non-negativity constraints that produce the boundary fits documented here, and whether the diagnostic proposed here can be tuned to trade its false alarms against its misses in a principled way. Establishing that failure is real, tracing it to known small-sample estimation behavior, showing that it is silent and asset-specific, and offering a screen against it is what we take this paper to contribute.

References

[1] Acerbi, C., & Szekely, B. (2014). Back-testing Expected Shortfall. Risk, 27(11), 76-81.

[2] Alexander, C., & Dakos, M. (2023). Assessing the accuracy of exponentially weighted moving average models for value-at-risk and expected shortfall of crypto portfolios. Quantitative Finance, 23(3), 393-427.

[3] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

[4] Caro, M. C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L., & Coles, P. J. (2022). Generalization in quantum machine learning from few training data. Nature Communications, 13, 4919.

[5] Christoffersen, P. F. (1998). Evaluating interval forecasts. International Economic Review, 39(4), 841-862.

[6] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1007.

[7] Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. Journal of Econometrics, 129(1-2), 121-138.

[8] Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of GARCH estimates and persistence. The European Journal of Finance, 12(6-7), 473-494.

[9] Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives, 3(2), 73-84.

[10] Mappadang, A., Nugroho, B. A., Lestari, S. D., Elizabeth, & Lestari, T. K. (2024). Measuring value-at-risk and expected shortfall of newer cryptocurrencies: New insights. Cogent Business & Management, 11(1), 2416096.

[11] Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica, 59(2), 347-370.

[12] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1), 246-256.

[13] Troster, V., Tiwari, A. K., Shahbaz, M., & Macedo, D. N. (2019). Bitcoin returns and risk: A general GARCH and GAS analysis. Finance Research Letters, 30, 187-193.

[14] Trucios, C. (2019). Forecasting Bitcoin risk measures: A robust approach. International Journal of Forecasting, 35(3), 836-847.

[15] Trucios, C., & Taylor, J. W. (2023). A comparison of methods for forecasting value at risk and expected shortfall of cryptocurrencies. Journal of Forecasting, 42(4), 989-1007.