Within-Venue Monitoring of BTC/USDT Liquidity and Resiliency on Binance: A Queueing-Theoretic Framework

Samir Varma *

Posted Date: 3 April 2026

doi: 10.20944/preprints202604.0256.v1

Keywords: Bitcoin; market microstructure; order flow; queueing theory; market quality; liquidity risk; digital assets; market resiliency

 doi:10.20944/preprints202604.0256.v1

Article

Within-Venue Monitoring of BTC/USDT Liquidity and Resiliency on Binance: A Queueing-Theoretic Framework

Samir Varma

VS Asset Management, LLC, Cos Cob, CT 06807, USA; samir@vsasset.com

Abstract

We develop a queueing-organized framework for within-venue monitoring of BTC/USDT liquidity, signed-flow pressure, and resiliency on Binance. The model treats latent buy and sell pressure as occupancy processes and uses that state space to organize three empirical diagnostics: the variance-per-BTC liquidity measure R_r, the effective mean-reversion rate θ_{eff}, and the companion signed-flow proxy $\beta_{\text{eff}}^{\text{proxy}}$. Using Binance trade data from 2020–2025, we find a pooled first-order variance–volume regularity away from the highest-volume tail and substantial time variation in rolling liquidity and resiliency. In overlapping 30-day windows, θ_{eff} is positive by point estimate in roughly two-thirds of windows but clearly positive in only about two-fifths under a simple uncertainty buffer, implying that local recovery is often fragile or ambiguous. The intended users are short-horizon risk managers, execution desks, market makers, and exchange surveillance teams that need auditable venue-level indicators of when liquidity is thinning, recovery is weakening, and signed flow is turning one-sided. Queueing is useful here because it turns those signals into one coherent monitoring dashboard for venue-level market quality and short-horizon risk.

Keywords: Bitcoin; market microstructure; order flow; queueing theory; market quality; liquidity risk; digital assets; market resiliency

1. Introduction

Binance spot BTC/USDT is a useful venue for studying short-horizon liquidity and resiliency because it combines continuous high-turnover trading with auditable trade prints and aggressor-side information. Those features let us build venue-level diagnostics directly from transaction data and track how trading activity, signed flow, and recovery conditions evolve through time inside a single market.

This paper develops a queueing-organized monitoring framework for three linked objects: the variance-per-BTC liquidity diagnostic R_r, the effective mean-reversion rate θ_{eff}, and the companion signed-flow proxy $\beta_{\text{eff}}^{\text{proxy}}$. The queueing layer models latent buy and sell pressure as occupancy processes. We do not directly observe customer-level queues, but this state-space view places liquidity, signed-flow pressure, and recovery on one common stock-flow accounting system instead of treating them as unrelated reduced-form measures. The practical question is simple: when do trade prints suggest that this venue is liquid and resilient, and when do they instead suggest that it is becoming thin, one-sided, and fragile?

The paper sits at the intersection of four literatures: queueing and inventory-based market models (Cont & de Larrard, 2013; Garriott et al., 2025), order-flow and price-impact models (Amihud & Mendelson, 1980; Ho & Stoll, 1981; Kim & Stoll, 2014; Kyle, 1985; Stoll, 1978), cryptocurrency market microstructure (Alexander et al., 2023; Anastasopoulos et al., 2026; Dimpfl, 2017), and short-horizon market-risk monitoring (Foucault et al., 2013). The contribution is a transparent single-venue monitoring framework for a continuously traded and fully auditable market.

 doi:10.20944/preprints202604.0256.v1

Who should care and why

The natural audience includes market makers adjusting inventory, execution desks assessing trading conditions, exchange risk and surveillance teams watching for one-sided markets, and researchers who want auditable market-quality measures. What matters to those users is whether the venue is becoming thin, whether directional pressure is unwinding quickly or only weakly, and whether stressed states are followed by worse near-term risk conditions. The aim is to show that a simple queueing state space can organize those questions using observables that are available in Binance trade data.

The framework models latent buy-pressure and sell-pressure queues as two independent $M/G/\infty$ systems. New directional pressure arrives stochastically, existing pressure unwinds after random holding times, and the resulting net imbalance is mapped into price changes through a linear imbalance-to-price mapping. Throughout, "long" and "short" label latent directional pressure rather than directly observed leveraged financing positions, and the queues should be read as BTC-normalized latent occupancy units rather than as observed limit-order-book queues. These primitives yield a stationary Skellam distribution for net inventory, a variance–volume relation, and a local feedback extension that motivates the observable monitoring diagnostics used below.

Empirically, three results carry most of the paper. First, variance per traded BTC shows a pooled first-order regularity across volume bins, strongest away from the highest-volume tail, and the rolling R_r series shows that liquidity conditions move materially over time. Second, lagged hourly return autocorrelation yields a rolling effective mean-reversion estimate whose point estimate is positive in roughly two-thirds of 30-day windows but clearly positive in only about two-fifths under a simple uncertainty buffer. Third, the rolling diagnostics are not only contemporaneous summaries: high-R_r days are followed by worse next-day variance and tail outcomes than low-R_r days, and non-positive resiliency windows are followed by worse next-day tail outcomes than clearly mean-reverting ones. The appendix recovery event study points in the same direction for isolated large hourly shocks.

What is genuinely learned from queueing here

Queueing earns its place here for three reasons. First, it makes R_r, θ_{eff}, and $\beta_{\text{eff}}^{\text{proxy}}$ measurements on one common stock-flow state space rather than three separate alarms. Second, it gives θ_{eff} a recovery-timescale interpretation in terms of latent pressure unwinding. Third, under the symmetric $M/M/\infty$ special case it supplies one concrete turnover-versus-recovery translation. The empirical core remains observable: R_r and θ_{eff} are directly estimated, $\beta_{\text{eff}}^{\text{proxy}}$ is a companion proxy, and the one-hour symmetry-based translations are kept for interpretation rather than presented as stand-alone structural estimates. For the intended audience, the payoff is that the same framework answers three operational questions at once: how much liquidity the venue is delivering per traded BTC, how quickly it appears to recover from directional pressure, and whether signed flow is becoming one-sided enough to treat current conditions as more fragile.

Table 1 makes that added value concrete. The queueing layer does not replace reduced-form evidence; it organizes that evidence and tells the reader how the main diagnostics fit together.

Table 1. What the queueing layer adds beyond reduced-form monitoring. The table reports the observable objects used in the main paper and the queueing interpretation attached to them.

Observable object	Reduced-form reading	Queueing reading
R_r	Liquidity / variance ratio per traded BTC	Occupancy-based variance scale linking traded volume to latent imbalance
θ_{eff}	Local short-horizon mean reversion from lag-1 autocorrelation	Effective unwinding / resiliency timescale of latent pressure
$(\tilde{\kappa}_{\text{dir}}, b_F, \beta_{\text{eff}}^{\text{proxy}})$	Signed-flow pressure and feedback proxy	Arrival/exit pressure read in the same state space as occupancy decay

Symmetry-based translations into $\mathrm{E}[H]$, $\lambda_{\pm}$, and $m_{\pm}$ are reported only in the Online Appendix as a joint full-sample illustration under the symmetric $M/M/\infty$ bridge. They are not presented here as row-by-row structural estimates.

 doi:10.20944/preprints202604.0256.v1

The layered assumption stack is summarized once in Table A1 in the Online Appendix. The empirical discipline is simple: R_r and θ_{eff} are directly estimated from hourly variance, volume, and lag-1 autocorrelation; $\beta_{\text{eff}}^{\text{proxy}}$ is a companion splice built from minute directional impact and one-hour signed BTC flow; and symmetry-based quantities such as $\mathrm{E}[H]$, $\lambda_{\pm}$, and $m_{\pm}$ are appendix-only contingent arithmetic. The empirical case for the paper therefore rests on three observable margins: a pooled variance-per-BTC regularity, a rolling resiliency classification built from $\hat{\theta}_{\text{eff}}$, and a companion signed-flow measure that helps organize market phases. The symmetric one-hour increment benchmark and the tail overlay are supporting checks. A full-sample contingent queueing illustration still helps interpretation: under the nearby symmetric one-hour splices reported in Appendix B.4 of the Online Appendix, the observed liquidity and resiliency bundle maps to indicative mean holding times of about 5.20–7.63 h and total latent occupancy of about 8,524–12,510 BTC-normalized units.

Scope and limitations

This is a focused single-venue paper. It is implemented on Binance spot BTC/USDT, uses a linear impact approximation, and does not run an exhaustive comparison with alternative microstructure models. Those choices are deliberate: they keep the framework transparent, auditable, and easy to implement on continuously updated trade data. The paper should therefore be read as a venue-level monitoring contribution. The central question is whether the observable diagnostics—R_r, θ_{eff}, and $\beta_{\text{eff}}^{\text{proxy}}$—remain jointly interpretable and practically useful inside one bookkeeping system.

Table 2 summarizes the main outputs of the framework and their empirical implementation. Section 2 presents the queueing-theoretic framework, the feedback mechanism, the data, and the empirical implementation. Section 3 reports the main results, including variance–volume evidence, directional impact triangulation, distributional fit, rolling stability, and a brief pragmatic tail-risk overlay. Section 4 discusses the implications for short-horizon price formation, liquidity, and market monitoring. Section 5 concludes.

Table 2. Baseline monitoring dashboard for the main paper (BTC/USDT, Binance, 2020–2025). The table emphasizes directly estimated quantities and one companion proxy used in the paper's main claims.

Quantity	Construction
Variance-per-BTC R_r	$\mathrm{Var}[r_t]/\mathbb{E}[\nu_t] = 1.44 \times 10^{-8}$ with bootstrap s.e. 8.09×10^{-10}†
Effective mean-reversion θ_{eff}	Lag-1 return autocorrelation, $\hat{\theta}_{\text{eff}} = 0.0477\,\text{h}^{-1}$, implying descriptive half-life 14.54 h (roughly 10–25 h)
Heuristic clear-positive 30-day windows	One-standard-error rolling classification share, 39.2%
Companion signed-flow proxy $\beta_{\text{eff}}^{\text{proxy}}$ (baseline splice)	$\beta_{\text{eff}}^{\text{proxy}} = \frac{1}{2}\tilde{\kappa}_{\text{dir}} b_F = 7.24 \times 10^{-2}\,\text{h}^{-1}$

† Estimates use a percentile bootstrap with 10,000 resamples of 24-hour blocks. R_r is the main liquidity diagnostic, and θ_{eff} comes from lag-1 autocorrelation under a local OU approximation, so the implied half-life is descriptive rather than literal. Under the baseline and doubled heuristic buffers, the clearly mean-reverting 30-day share ranges from 17.2% to 39.2%, with ambiguity dominant under either rule. $\beta_{\text{eff}}^{\text{proxy}}$ is a companion signed-flow summary whose sign is stable across nearby-frequency alternatives, although its exact level is splice-sensitive. Symmetry translations are reported only in the Online Appendix.

Additional derivations, extensions, and robustness checks are reported in the Online Appendix.

2. Materials and Methods

2.1. Core Assumptions and Notation

(M1) Independent Poisson arrivals of long (λ_+) and short (λ_-) latent occupancy units, BTC-normalized by bookkeeping convention.

(M2) I.i.d. holding times with finite means $\mathbb{E}[H_{\pm}]$ ($M/G/\infty$ queues).

(M3) Linear imbalance-to-price mapping $P_t = P_0 + \kappa X_t$ with $X_t = B_t - S_t$.

All subsequent sections build directly on these three postulates.

 doi:10.20944/preprints202604.0256.v1

Notation and scale conventions

Throughout, $R_r \equiv \mathrm{Var}(r_t)/\mathbb{E}[v_{1h}]$ denotes the observable returns-scale variance-per-BTC moment computed from one-hour return increments and one-hour traded BTC volume, and $\sqrt{R_r}$ its associated variance scale. The structural return-impact slope is $\tilde{\kappa} \equiv \kappa/\overline{P}$, with units BTC^{-1}; once the one-hour holding-time correction c_H is introduced below, the symmetry-based special-case translation is $R_r = c_H\tilde{\kappa}^2$. Price-scale variance-per-BTC is $R = \bar{P}^2 R_r$. For numerical illustrations in USD we set $P^* = \$50{,}000$; all calibrations and tests use returns units, so no reported estimate depends on the choice of P^*. For empirical bookkeeping, queue occupancies are BTC-normalized latent occupancy units throughout: one latent unit corresponds by normalization to 1 BTC. This is a bookkeeping convention for the latent state, not a claim that literal customer orders are observed in exact 1-BTC blocks. A marked or compound-Poisson generalization with random chunk sizes would be the natural extension, but the present paper uses unit-normalized occupancies to keep the monitoring diagnostics tractable. Operationally, the latent state is tied to executed BTC turnover only at the first-moment level: the normalization is chosen so that expected openings and closures of latent occupancy units match expected executed BTC volume in the bookkeeping identity used below. A marked extension would preserve that first-moment bridge while changing higher moments and tail behavior. The stationary occupancy law is the general $M/G/\infty$ result; the one-hour increment bridge used later for contingent structural translation is the narrower symmetric $M/M/\infty$ special case.

Table 3. Core symbols and dimensions.

Symbol	Meaning	Units
$B_t,\ S_t$	Outstanding long / short latent occupancy units	BTC-normalized latent occupancy units
$X_t = B_t - S_t$	Net inventory imbalance	BTC-normalized latent occupancy units
$P_t,\ P_0$	Price at t / local reference level	USD
$\lambda_\pm$	Arrival rate of long / short latent occupancy units	BTC-normalized latent occupancy units h^{-1}
$H_\pm$	Holding-time random variable (long / short)	h
$\mu_\pm = 1/\mathbb{E}[H_\pm]$	Exit (closure) rate	h^{-1}
$m_\pm = \lambda_\pm \mathbb{E}[H_\pm]$	Steady-state latent occupancy	BTC-normalized latent occupancy units
κ	price-impact (price scale), $\Delta P_t \approx \kappa\,\Delta X_t$	USD $\cdot$ BTC^{-1}
$\tilde{\kappa}$	structural return-impact slope, $\tilde{\kappa} \equiv \kappa/\overline{P}$	BTC^{-1}
R_r	observable variance-per-BTC in returns units, $\mathrm{Var}(r_{1h})/\,\mathrm{E}[v_{1h}]$	BTC^{-1}
$\sqrt{R_r}$	observable returns-scale variance coefficient	$\mathrm{BTC}^{-1/2}$
$\overline{P}$	sample-average price used to relate scales	USD
R	variance-per-BTC (price units; hourly unless otherwise stated)	$\mathrm{USD}^2 \cdot \mathrm{BTC}^{-1}$
β	Latent momentum-feedback slope in the local state splice	BTC h^{-1} per unit return
$\beta_{\mathrm{eff}} = \tilde{\kappa}\beta$	Structural feedback-adjusted Ornstein–Uhlenbeck (OU) drift under the latent-state splice	h^{-1}
$\beta_{\mathrm{eff}}^{\mathrm{proxy}}$	Reduced-form proxy feedback coefficient from lagged return and signed BTC flow	h^{-1}
$z_t = \tilde{\kappa} X_t$	Latent local price-deviation state used in the feedback splice	return
$\theta = \mu_+ + \mu_-$	Aggregate exit intensity	h^{-1}
$\vartheta,\ \vartheta_\pm$	Gamma mixing scale (all / long / short)	h
v_{1h}	Executed BTC volume in a one-hour bar	BTC
$\sigma_{\mathrm{P}}^2 = R\,v_{1h}$	One-hour price-increment variance benchmark	USD^2
$\sigma_{\mathrm{r}} = \sqrt{R_r\,v_{1h}}$	One-hour return standard deviation	return

Queueing mapping

Queueing notation treats latent occupancy-unit openings as arrivals and holding times as service; for an $M/G/\infty$ queue the steady-state occupancy is Poisson with mean $\lambda\mathbb{E}[H]$, yielding a Skellam distribution for the long-short difference. A brief primer and the mapping from queueing language to trading pressure are provided in Appendix A of the Online Appendix.

2.2. Queueing-Theoretic Framework

2.2.1. Model Setup and Intuition

The central modeling idea is to treat the market as a system of queues in which price pressure emerges from the imbalance between outstanding long and short latent occupancy units.

Consider a Bitcoin market populated by two latent directional-pressure types: those contributing buy pressure ("longs") and those contributing sell pressure ("shorts"). At any point in time, there are B_t outstanding long-side occupancy units and S_t outstanding short-side occupancy units. These are

 doi:10.20944/preprints202604.0256.v1

not directly observed financing books; they are state variables summarizing net pressure that remains active in the market. The key state variable is the net inventory imbalance:

$$X_t = B_t - S_t \tag{1}$$

Over short horizons, local price deviations are determined by this inventory imbalance through a linear imbalance-to-price mapping:

$$P_t = P_0 + \kappa X_t \tag{2}$$

where P_0 is a slowly moving local reference level and $\kappa > 0$ measures the local price effect per unit of net inventory.

Impact-kernel interpretation

Let latent occupancy-unit openings arrive at times t_i with sign $s_i \in \{+1, -1\}$ and holding time H_i. Then

$$X_t = \sum_i s_i \, \mathbf{1}\{t_i \le t < t_i + H_i\},$$

and, taking expectations, the mean price follows

$$\mathbb{E}[P_t] = P_0 + \kappa \sum_i s_i \, G(t - t_i), \qquad G(u) = \Pr(H > u).$$

This is a shot-noise / propagator representation of latent directional pressure with kernel G given by the holding-time survival function (Bouchaud et al., 2009). It is not a one-to-one mapping from public trade prints to literal queue openings and exits; executed BTC enters the empirical implementation only through the first-moment bookkeeping bridge in equation (12). In this interpretation, the price effect persists while latent pressure survives and decays as occupancy exits. The mean holding time $\bar{H}$ is therefore an *effective resiliency timescale* of latent order-flow pressure; in the present implementation we infer that timescale from lagged hourly return autocorrelation and then translate it into $\mathbb{E}[H]$ under symmetry (Section 2.5).

The dynamics of long and short occupancy units follow queueing processes. New long-side units arrive at rate λ_+ and each unit is held for a random time H_+ drawn from distribution G_+. Similarly, new short-side units arrive at rate λ_- with holding times H_- drawn from G_-. The heterogeneity in trader horizons is captured by the entire distribution of holding times rather than a single parameter.

We implement the model on Binance BTC/USDT spot transactions over the period 1 January 2020–9 July 2025. The empirical work does not directly estimate customer-level arrivals, exits, or holding times from minute buckets, nor does it observe order-book queue position. Instead, it estimates an observable variance scale, a minute directional-impact proxy, and an effective mean-reversion rate, then translates those objects into contingent structural quantities under symmetry for interpretation. Data conventions and empirical estimates are reported in Sections 2.4 and 3.

These assumptions are most appropriate when trading is continuous, turnover is high, and short-horizon price adjustment is strongly influenced by order flow; BTC/USDT fits those conditions reasonably well over our sample horizon (see Appendix A of the Online Appendix).

2.2.2. Stationary Inventories and Skellam Law

Outstanding long and short occupancy units each follow an $M/G/\infty$ queueing system. New units arrive according to Poisson processes with rates λ_+ and λ_-, and holding times are drawn from G_+ and G_- with means $\mathbb{E}[H_+]$ and $\mathbb{E}[H_-]$.

A standard queueing result is Palm's theorem (Palm, 1943), which characterizes the steady-state distribution of an $M/G/\infty$ system:

Lemma 1 (Palm's Theorem)**.** *In an $M/G/\infty$ queueing system with arrival rate λ and service time distribution with mean $\mathbb{E}[H]$, the steady-state occupancy is Poisson distributed with parameter $m = \lambda \mathbb{E}[H]$.*

 doi:10.20944/preprints202604.0256.v1

Applied to our setting, steady-state occupancies are

$$B_\infty \sim \text{Poisson}(m_+) \tag{3}$$

$$S_\infty \sim \text{Poisson}(m_-) \tag{4}$$

with $m_+ = \lambda_+ \mathbb{E}[H_+]$ and $m_- = \lambda_- \mathbb{E}[H_-]$.

Under the maintained baseline specification, the long and short queues are independent even though they operate in the same market. This independence is imposed through independent Poisson arrivals and independent holding periods across traders.

The steady-state net inventory is therefore:

$$\boxed{X_\infty \sim \text{Skellam}(m_+, m_-) \quad \Longrightarrow \quad P_\infty - P_0 = \kappa X_\infty} \tag{5}$$

This boxed result should be read as a local stationary law for price deviations around the reference level P_0, not as a literal unconditional law for the full Bitcoin price path over 2020–2025. The Skellam distribution—previously applied to tick-level price changes by Koopman et al. (2017)—emerges here as the stationary distribution of net inventory from explicit queueing primitives, linking arrival rates and holding times directly to local price-deviation moments.

Illustrative symmetry-based translations for (m_+, m_-) and the associated Skellam parameters are reported only in Appendix B.3 of the Online Appendix.

2.2.3. Price Distribution and Over-Dispersion

The Skellam distribution has probability mass function:

$$\Pr\{X_\infty = k\} = e^{-(m_+ + m_-)} \left(\frac{m_+}{m_-}\right)^{k/2} I_{|k|}(2\sqrt{m_+ m_-}) \tag{6}$$

where I_n is the modified Bessel function of the first kind of order n.[1]

From equation (2), the local price-deviation distribution is:

$$\Pr\{P_\infty = P_0 + \kappa k\} = e^{-(m_+ + m_-)} \left(\frac{m_+}{m_-}\right)^{k/2} I_{|k|}(2\sqrt{m_+ m_-}) \tag{7}$$

All moments of the local price-deviation distribution are available in closed form:

$$\mathbb{E}[P_\infty] = P_0 + \kappa(m_+ - m_-) \tag{8}$$

$$\text{Var}[P_\infty] = \kappa^2 (m_+ + m_-) \tag{9}$$

$$\text{Skew}[P_\infty] = \frac{m_+ - m_-}{(m_+ + m_-)^{3/2}} \tag{10}$$

$$\text{ExKurt}[P_\infty] = \frac{1}{m_+ + m_-} \tag{11}$$

These formulas reveal several important insights.

(i) First, the *expected* local price deviation is proportional to the difference between long and short pressure $(m_+ - m_-)$, formalizing the sign of the local price effect from net order imbalance.

(ii) Second, local price-deviation volatility scales with the sum of these pressures $(m_+ + m_-)$, so intense activity on *either* side widens the stationary distribution of price deviations.

[1] The modified Bessel function of the first kind arises naturally in circular statistics and here characterizes the difference of Poisson variates.

 doi:10.20944/preprints202604.0256.v1

(iii) Third, under the baseline Poisson specification the excess kurtosis is

$$\mathrm{Kurt}(P_t) - 3 = \frac{1}{m_+ + m_-},$$

which vanishes as order-flow intensity grows, implying asymptotically normal tails.

Empirically, hourly returns display extreme excess kurtosis (≈ 47.29), while the hourly buy-count and sell-count series are strongly over-dispersed relative to a Poisson benchmark. Fitting Gamma-mixed Poisson models to the hourly count series gives full-sample dispersion parameters $\alpha_+ = 0.4922$ and $\alpha_- = 0.6423$. These are trade-count burstiness proxies rather than BTC-commensurate structural calibrations of the latent queues, so the count-space extension below is retained only as a supplementary burstiness benchmark.

2.2.4. Negative–Binomial Over-Dispersion

Observed hourly buy and sell trade counts are substantially more bursty than a pure Poisson benchmark. To document that without overloading the main paper, we use a Gamma-mixed Poisson extension only as a count-space burstiness proxy. Under a common-rate approximation for the buy and sell count mixtures, the hourly increment's excess kurtosis can be decomposed into a tiny Poisson benchmark term plus a much larger over-dispersion term. For Binance, the Poisson term is only 1.8×10^{-5} while the over-dispersion contribution is 5.2889, which is why the paper treats burstiness as a real feature of the count data. The exact common-rate derivation and parameterization are deferred to Appendix H.1 of the Online Appendix; in the main paper this point is only supplementary and does not change the interpretation of the core liquidity and resiliency diagnostics.

2.2.5. Variance–Volume Relation

At the first-moment level, the queueing framework implies a disciplined link between trading volume and the ex-ante variance of price *changes*. Under the BTC-normalized latent-unit convention, observed executed BTC volume is bookkept by expected openings and expected closures of latent occupancy units. No event-level one-to-one mapping from public trades to literal latent queue openings or closures is being claimed. Because each side of the book is an $M/G/\infty$ queue with mean inventory $m_\pm = \lambda_\pm \mathbb{E}[H_\pm]$ and exit rate $\mu_\pm = 1/\mathbb{E}[H_\pm]$, the BTC-normalized first-moment bookkeeping rate for executed volume is

$$\nu = (\lambda_+ + \lambda_-) + \mu_+ m_+ + \mu_- m_- = 2(\lambda_+ + \lambda_-), \tag{12}$$

where the first term captures *arrivals* and the second *departures*. The last equality follows by substituting $m_\pm$. Under this normalization, each latent occupancy unit contributes one BTC when opened and one BTC when closed in the first-moment bookkeeping. Equation (12) is a bookkeeping identity under the BTC-normalized latent-unit convention, not direct observation of literal queue openings and closures. The model therefore does not independently predict venue volume. Instead, it uses a BTC-normalized first-moment convention to place observed turnover and latent occupancy on a common scale. The paper's falsifiable content begins at the observable level: the variance-per-BTC moment R_r, the effective mean-reversion estimate θ_{eff}, the proxy feedback diagnostic $\beta_{\text{eff}}^{\text{proxy}}$, and the success or failure of the symmetric increment benchmarks. It does not begin at literal queue openings or at the contingent symmetry translations reported later only for interpretation.

Sample mean executed volume

Given the hour-by-hour executed BTC volume ν_t, the long-run mean hourly executed BTC volume is estimated by

$$\hat{\nu} = \frac{1}{T}\sum_{t=1}^{T} \nu_t, \tag{13}$$

where T is the number of hourly bars in the sample.

 doi:10.20944/preprints202604.0256.v1

Empirical calibration

Section 3.2 reports the empirical estimates of R_r, its square-root scale $\sqrt{R_r}$, and the contingent structural translation $\tilde{\kappa}$ for Binance BTC/USDT. Table 2 summarizes the main outputs.

Returns form (used in all calibrations and Q–Q plots): with $r_t := P_t/P_{t-1} - 1$ and $\tilde{\kappa} \equiv \kappa/\overline{P}$,

$$\mathrm{Var}[r_t] = R_r \, \mathrm{E}[\nu_t] \qquad (h = 1).$$

2.3. Feedback, Stability, and Resiliency Mechanism

At intraday horizons, signed-flow intensities can respond to recent price movements. This extension adds a local state-dependent splice to motivate a proxy-based resiliency classifier within the queueing framework. In the empirical sections below, the object carried forward is the local pair $(\theta_{\text{eff}}, \beta_{\text{eff}}^{\text{proxy}})$, not a claim of direct structural estimation of the latent queue-stability boundary.

2.3.1. State-Dependent Arrivals

We extend the model to allow arrival rates that depend on recent price changes, capturing local return-following or contrarian signed-flow behavior in a latent queueing splice.

Standing assumptions

A1 **(Queueing)** Arrivals of long- and short-side latent occupancy units are independent Poisson with baseline rates λ_+^0, λ_-^0; holding times are independent and identically distributed (i.i.d.) with finite means $\mathbb{E}[H_\pm]$.

A2 **(Linear price impact)** Price obeys $P_t = P_0 + \kappa X_t$ with constant $\kappa > 0$.

A3 **(Linear feedback splice)** Latent arrival rates respond linearly to the local latent price-deviation state, equations (14)–(15); the empirical sections below do not observe that state directly and therefore estimate a separate lagged-return proxy coefficient.

Linear feedback specification

Let $\theta = \mu_+ + \mu_-$ denote the aggregate exit intensity [h^{-1}] and define the rescaled latent feedback coefficient

$$\beta_{\text{eff}} := \tilde{\kappa}\,\beta \quad [\beta_{\text{eff}} \text{ is } \mathrm{h}^{-1}].$$

To avoid confusion with observed close-to-close one-hour return increments, let

$$z_t := \tilde{\kappa} X_t$$

denote the model's local latent price-deviation state in returns units. The feedback equations in this subsection are written in terms of z_t, whereas the empirical sections reserve r_t for observed one-hour close-to-close returns and estimate the separate proxy coefficient $\beta_{\text{eff}}^{\text{proxy}}$ from lagged returns and signed BTC flow. Let arrival rates respond linearly to the recent (left-limit) local price-deviation state z_{t-}:

$$\lambda_+(t) = \lambda_+^0 + \beta\, z_{t-} \tag{14}$$

$$\lambda_-(t) = \lambda_-^0 - \beta\, z_{t-} \tag{15}$$

where $\lambda_\pm^0$ are baseline arrival rates and $\beta \in \mathbb{R}$ measures the strength of state dependence in the latent price-deviation state. When $\beta > 0$, positive recent deviations are associated with return-following latent pressure; when $\beta < 0$, positive deviations are associated with contrarian latent pressure. Because z_{t-} is $\mathcal{F}_{t-}$-measurable, this specification is predictable and introduces no simultaneity. We interpret (14)–(15) as a local linearisation and truncate rates at zero if needed; under the full-sample symmetry translation summarized in Appendix B.3 of the Online Appendix, the implied perturbation βz_{t-1} is small relative to the baseline intensities $\lambda_\pm^0$, so truncation does not bind in practice.

 doi:10.20944/preprints202604.0256.v1

Linearized dynamics

For the local OU approximation in this subsection, we additionally impose symmetric exits $\mu_+ = \mu_- =: \mu = \theta/2$. Substituting $z_{t-} = \tilde{\kappa} X_{t-}$ into the net arrival rate gives feedback proportional to $2\beta_{\text{eff}} X_t$, so the drift of the expected net inventory is

$$\frac{d}{dt}\mathbb{E}[X_t] = (\lambda_+^0 - \lambda_-^0) - (\theta/2 - 2\beta_{\text{eff}})\mathbb{E}[X_t]. \tag{16}$$

Whenever the local effective mean-reversion rate

$$\theta_{\text{eff}} := \theta/2 - 2\beta_{\text{eff}}$$

is positive, the corresponding local mean level is

$$\bar{X} = \frac{\lambda_+^0 - \lambda_-^0}{\theta_{\text{eff}}}. \tag{17}$$

Defining the centered state $Y_t := X_t - \bar{X}$, the local linearized dynamics take the Ornstein–Uhlenbeck form

$$dY_t \;=\; -(\theta/2 - 2\beta_{\text{eff}})\, Y_t\, dt \;+\; \sqrt{\nu_0}\, dW_t, \tag{18}$$

where $\nu_0 = 2(\lambda_+^0 + \lambda_-^0)$ is the baseline event intensity and W_t is a standard Wiener process. The continuous-time stability condition is therefore

$$\theta > 4\beta_{\text{eff}}. \tag{19}$$

Stability analysis

The centered local linearization is mean-reverting only when the drift coefficient is positive, i.e. when $\theta_{\text{eff}} = \theta/2 - 2\beta_{\text{eff}} > 0$, equivalently $\theta > 4\beta_{\text{eff}}$ under the symmetric-exit splice. This inequality is best read as a latent heuristic for the sign logic behind the observed pair $(\theta_{\text{eff}}, \beta_{\text{eff}}^{\text{proxy}})$ rather than as a directly identified latent stability boundary. Baseline asymmetry $\lambda_+^0 - \lambda_-^0$ shifts the local mean level $\bar{X}$ but does not by itself change that local sign condition. Empirically, the paper uses this splice only to organize local regime classification through $\hat{\theta}_{\text{eff}}$ and the companion proxy $\beta_{\text{eff}}^{\text{proxy}}$; the corresponding contingent symmetry summary for ρ is reported only in Appendix C.9 of the Online Appendix.

Metrics. The implied resiliency half-life—the time for the centered state Y_t to decay halfway back toward its local mean after a shock—is $t_{1/2} = \ln 2/\theta_{\text{eff}}$, where $\theta_{\text{eff}} = \theta/2 - 2\beta_{\text{eff}}$ is the effective mean-reversion rate from (18). Because the structural θ is not directly observed, the main text focuses on the directly estimated pair $(\theta_{\text{eff}}, t_{1/2})$ and treats any further translation into θ, $\mathbb{E}[H]$, c_H, or ρ as contingent on the symmetric proxy mapping summarized only in the appendix. For the calibrated parameters (Table 4), this yields the descriptive full-sample contingent illustration $t_{1/2} \approx 14.54$ h. The empirical role of this subsection is therefore local and fragile by design: it supplies a queueing-organized interpretation for observed resiliency diagnostics, not a claim that the latent queue boundary is directly estimated from venue data.

2.4. Data

We use Binance BTC/USDT trade data from 1 January 2020 through 9 July 2025, a window that spans the 2021 bull market, the FTX collapse, and the 2024 exchange-traded fund (ETF) launch period. The working sample contains 48,374 hourly bars over 2,017 Coordinated Universal Time (UTC) days. A complete UTC grid over that span would contain 48,408 hours; the realized panel contains 48,374 because 34 hours are missing on 15 UTC days. Those missing hours come from gaps in the source hourly summary files that feed the canonical hourly panel, not from discretionary filtering or from the later aggressor-side merge. We retain those partial UTC days and compute daily variance and volume from the observed hourly bars. Dropping the 15 incomplete UTC days changes the headline variance-

 doi:10.20944/preprints202604.0256.v1

per-BTC estimate by less than 1% and leaves the lag-1 return autocorrelation essentially unchanged, so the main results are not driven by those omissions. For each UTC day we use the raw tick file to construct minute-level signed BTC flow and returns, the workflow-generated hourly summary file for close-to-close returns and traded BTC volume, and the workflow-generated aggressor-side summary file to build an hourly panel of buyer-initiated BTC volume. All timestamps are converted to UTC. Binance spot BTC/USDT is used as a first venue because it combines deep continuous trading with directly auditable trade-print and aggressor-side records, which makes it a practical setting for a queueing-based monitoring exercise even though the broader Bitcoin market is multi-venue.

All data processing, estimation, figure generation, and Monte Carlo validation were performed in Python 3.9.6 using NumPy 1.26.4, SciPy 1.13.1, pandas 2.2.2, and matplotlib 3.9.4.

The empirical pipeline therefore uses three aligned objects: an hourly return/volume panel, a day-level minute-regression panel, and an hourly signed-BTC-flow panel. The first supports the variance–volume and mean-reversion estimates, the second provides a directional-impact scale in BTC units, and the third provides the signed-flow input for the feedback and rolling-diagnostic results.

Sign convention. Signed order flow uses the *taker* side: a trade has sign $+1$ if the aggressor is the buyer and -1 if the aggressor is the seller. Concretely, we take `isBuyerMaker=False` as $+1$ and `True` as -1. Buyer-initiated and seller-initiated BTC volume are therefore observed directly at the trade level and then aggregated to minute or hourly frequency as needed.

Cleaning and alignment. The construction is deliberately mechanical. In the canonical hourly panel, we retain only rows with finite hourly close, traded BTC volume, and close-to-close return values, then sort by UTC timestamp. In the minute regressions, raw trade records with non-finite price, quantity, or timestamp fields are dropped, timestamps are normalized to milliseconds, and signed BTC flow is aggregated within exact UTC minutes before minute returns are computed from the first and last trade in each occupied minute. The hourly BTC-flow panel is then matched hour-by-hour to the workflow-generated aggressor-side summary files; the workflow stops if any required day is missing or if the merge produces missing hourly buy quantities. No manual winsorization or discretionary outlier deletion is applied in these steps.

2.5. Empirical Implementation

We implement the framework at the one-hour horizon unless stated otherwise. The empirical procedure distinguishes four objects. First, $\sqrt{R_r}$ is an observable variance scale identified from the variance–volume moment condition. Second, θ_{eff} is an effective mean-reversion rate identified from return autocorrelation. Third, $\tilde{\kappa}_{\text{dir}}$ is a minute directional-impact slope in BTC units. Fourth, $\beta_{\text{eff}}^{\text{proxy}}$ is a proxy-based feedback rate obtained by combining $\tilde{\kappa}_{\text{dir}}$ with one-hour signed BTC flow. Structural quantities such as $\tilde{\kappa}$, θ, $\mathbb{E}[H]$, $\lambda_{\pm}$, and $m_{\pm}$ are reported only as model-implied translations under the symmetry mapping. Throughout this section, all bar-level observables are one-hour totals: price-scale variance satisfies $R = \text{Var}[\Delta P_{1h}]/\nu_{1h}$ and returns-scale variance satisfies $\text{Var}[r_{1h}] = R_r\ \text{E}[\nu_{1h}]$.

Observable variance scale

The variance–volume identity yields the one-hour variance scale

$$\sqrt{R_r} \;:=\; \sqrt{\frac{\text{Var}(r_{1h})}{\text{E}[\nu_{1h}]}}. \tag{20}$$

This is an observable variance coefficient, not a directional-impact slope. Standard errors are computed with a block bootstrap using 10,000 resamples of 24-hour blocks.

Effective mean-reversion

The aggregate mean-reversion rate θ_{eff} is identified from the lag-1 autocorrelation of hourly returns. Under the OU linearisation (18), the lag-1 autocorrelation coefficient $\hat{\rho}_1$ satisfies $\hat{\rho}_1 = (e^{-\theta_{\text{eff}}} - 1)/2$. We estimate

$$\hat{\theta}_{\text{eff}} \;=\; -\ln(1 + 2\,\hat{\rho}_1), \tag{21}$$

 doi:10.20944/preprints202604.0256.v1

which follows because the lag-1 autocorrelation of the OU increment $r_t = X_t - X_{t-1}$ is $(e^{-\theta_{\text{eff}}} - 1)/2$, and is well-defined when $1 + 2\hat{\rho}_1 > 0$, i.e. $\hat{\rho}_1 > -0.5$. Applied to simple hourly returns, this is a local short-horizon approximation rather than an exact identity for the full return process. The resiliency half-life follows as $t_{1/2} = \ln 2/\theta_{\text{eff}}$, and the structural exit rate is implied under the proxy mapping as $\theta = 2\theta_{\text{eff}} + 4\beta_{\text{eff}}$. Because θ_{eff} is identified from a single autocorrelation coefficient (SE$\approx$ 0.005), the half-life is imprecisely estimated; a two-standard-error band spans roughly 10–25 h.

Directional-impact scale

To obtain a unit-consistent flow-feedback diagnostic, we first estimate day-by-day minute regressions of one-minute returns on contemporaneous signed BTC flow:

$$r_{d,m} \;=\; a_d + \kappa_d\, q_{d,m} \;+\; \varepsilon_{d,m}, \tag{22}$$

where $q_{d,m}$ is signed BTC volume (buyer-initiated minus seller-initiated) computed directly from trade records. The directional-impact scale used in the feedback mapping is

$$\hat{\kappa}_{\text{dir}} := \text{median}_d\, |\hat{\kappa}_d|.$$

This scale has units BTC^{-1} and is reported descriptively in Section 3.3. The day intercepts a_d are estimated but not interpreted.

Feedback

We then construct hourly signed BTC flow F_t from buyer-initiated minus seller-initiated BTC volume and estimate

$$F_t \;=\; a_F + b_F\, r_{t-1} + u_t, \qquad \beta_{\text{eff}}^{\text{proxy}} := \frac{1}{2}\,\tilde{\kappa}_{\text{dir}}\, b_F, \tag{23}$$

where F_t is measured in BTC within hour t, so b_F has units BTC per unit return. A short discrete-time bridge makes the sign-and-units convention explicit. Let $\Delta = 1$ hour and define the excess signed flow over hour t by

$$F_t - \bar{F} \;\approx\; \int_{t-1}^{t} \big[(\lambda_+(s) - \lambda_-(s)) - (\lambda_+^0 - \lambda_-^0)\big]\, ds \;\approx\; 2\beta\, z_{t-1}\, \Delta,$$

where the second step is the local left-endpoint approximation under (14)–(15). Using the lagged one-hour return as the empirical proxy for the local deviation and $\tilde{\kappa}_{\text{dir}}$ as the minute-scale BTC-to-return slope proxy gives

$$F_t - \bar{F} \;\approx\; \frac{2\beta\Delta}{\tilde{\kappa}_{\text{dir}}}\, r_{t-1}, \qquad b_F \;\approx\; \frac{2\beta\Delta}{\tilde{\kappa}_{\text{dir}}}, \qquad \beta_{\text{eff}}^{\text{proxy}} \approx \tilde{\kappa}_{\text{dir}}\beta \approx \frac{\tilde{\kappa}_{\text{dir}}\, b_F}{2\Delta}.$$

With $\Delta = 1$ hour, this reduces to (23). This bridge is used only as a latent sign-and-units heuristic. Because r_{t-1} is an observed increment proxy for the unobserved latent state rather than the state itself, the resulting object is reported as $\beta_{\text{eff}}^{\text{proxy}}$, not as a direct estimate of the structural quantity β_{eff} and not as a mapped latent feedback rate. The minute/hour splice is chosen pragmatically because minute regressions stabilize the directional-impact slope while hourly aggregation reduces noise in signed BTC flow; nearby-frequency robustness is reported in Appendix C.8 of the Online Appendix. In the rolling 30-day results below, the signed-flow slope is re-estimated within each rolling window and combined with the same-window rolling median of daily directional-impact slopes, with the observed missing UTC hours carried through the hourly grid rather than filled. When $b_F < 0$ the flow is contrarian; when $b_F > 0$ it is return-following.

Workflow and units

The proxy construction is intentionally mechanical. Step 1: estimate the minute directional scale $\tilde{\kappa}_{\text{dir}}$ in units BTC^{-1}. Step 2: estimate the hourly signed-flow slope b_F in units BTC per unit return. Step 3: apply the one-hour discrete approximation above so that $\beta_{\text{eff}}^{\text{proxy}} = \frac{1}{2}\tilde{\kappa}_{\text{dir}} b_F$ is an hourly signed-flow

 doi:10.20944/preprints202604.0256.v1

monitoring proxy in h^{-1}. This makes the splice transparent, but it does not turn $\beta_{\text{eff}}^{\text{proxy}}$ into a directly identified arrival-rate parameter.

Model-implied structural translation

Under the symmetric $M/M/\infty$ special-case translation used for interpretation, the structural holding time is $\mathbb{E}[H] = 2/\theta$. For one-hour return increments, Appendix B.2 of the Online Appendix shows that the holding-time correction factor is

$$c_H = \mathbb{E}[H]\left(1 - e^{-1/\mathbb{E}[H]}\right),$$

so the structural return-impact slope follows as $\tilde{\kappa} = \sqrt{R_r / c_H}$. This bridge is exact for the symmetric exponential-holding-time mapping; it is not a nonparametric $M/G/\infty$ identification result. Additional occupancy translations, including $\lambda_\pm$ and $m_\pm$ under the same symmetry calibration, are reported only in Appendix B.3 of the Online Appendix. We report all such quantities as contingent model translations for interpretation, not as directly estimated facts.

A related distinction matters for the empirical results below. The stationary Skellam law from Section 2.2.2 concerns the level of latent imbalance X_t, whereas the distributional diagnostics in Section 3.4 benchmark one-hour return increments generated from that same core.

3. Results

3.1. Main Parameter Estimates

We estimate the empirical quantities described in Section 2.5. See Appendix F of the Online Appendix for a Monte Carlo confirmation that the variance-scale estimator based on $\text{Var}(r_{1h}) / \text{E}[\nu_{1h}]$ is effectively unbiased in finite samples.

Table 4. Technical pooled-baseline input table for the main empirical implementation (BTC/USDT, Binance, 2020–2025). Table 2 is the baseline monitoring dashboard; this table records the directly estimated and proxy-based ingredients behind it.

Symbol	Description	Value	Status / source
Empirical inputs (Binance, 2020–2025)			
$\text{Var}(r_{1h})$	Variance of hourly close-to-close returns	4.722×10^{-5}	Data moment
$\text{E}[r_{1h}]$	Mean hourly return	8.034×10^{-5}	Data moment
σ_r	Std. dev. of hourly returns, $\sigma_r = \sqrt{\text{Var}(r_{1h})}$	6.871×10^{-3}	Derived
$\text{E}[\nu_{1h}]$	Mean hourly traded volume (BTC)	3,281	Data moment
N_{hourly}	Number of hourly bars	48,374	Sample size
N_{daily}	Number of UTC days	2,017	Sample size
Directly estimated and proxy-based quantities			
$\sqrt{R_r}$	Observable 1-hour variance scale ($\text{BTC}^{-1/2}$), eq. (20)	1.2×10^{-4}	$\sqrt{\text{Var} / \text{E}[\nu]}$
$\tilde{\kappa}_{\text{dir}}$	Descriptive minute signed-flow impact proxy (BTC^{-1}), from the median of daily slopes in eq. (25)	2×10^{-5}	Minute signed-flow regressions
b_F	One-hour signed BTC-flow slope (BTC per unit return) from the lagged-return flow regression	7250.26	Hourly flow on lagged return
$\hat{\rho}_1$	Lag-1 autocorrelation of simple hourly returns	-0.0233	Direct sample autocorrelation
θ_{eff}	Effective mean-reversion rate (h^{-1}), eq. (21)	0.0477	Lag-1 autocorrelation
$t_{1/2}$	Resiliency half-life (h), $\ln 2/\theta_{\text{eff}}$	≈ 14.54	Local OU summary
$\beta_{\text{eff}}^{\text{proxy}}$	Companion signed-flow proxy (h^{-1}), baseline minute/hour splice	7.24×10^{-2}	$\frac{1}{2}\hat{\kappa}_{\text{dir}}\hat{b}_F$
$\Pr(\hat{\theta}_{\text{eff}} > 0 \text{ clearly})$	Heuristic clear-positive share of valid 30-day windows under a one-standard-error autocorrelation buffer	39.2%	Rolling autocorrelation

95% uncertainty ranges for the main objects are $\tilde{\kappa}_{\text{dir}} \in [1.96 \times 10^{-5}, 2.05 \times 10^{-5}]$, $b_F \in [6{,}145, 8{,}355]$, $\theta_{\text{eff}} \in [0.0292, 0.0665]$, and baseline-splice $\beta_{\text{eff}}^{\text{proxy}} \in [0.0613, 0.0837]$. The rolling classification is heuristic: under the baseline and doubled buffers, the clearly mean-reverting share ranges from 17.2% to 39.2%, with ambiguity dominant under either rule. Nearby-frequency alternatives 0.0661 and 0.0417 show that the proxy sign is stable but the exact level is splice-sensitive.

 doi:10.20944/preprints202604.0256.v1

The empirical mapping uses three observables: variance per traded BTC for $\sqrt{R_r}$, lag-1 hourly return autocorrelation for θ_{eff}, and the pair $(\tilde{\kappa}_{\text{dir}}, b_F)$ for $\beta_{\text{eff}}^{\text{proxy}}$. These are the main empirical objects; any further translation into θ, $\mathbb{E}[H]$, c_H, $\lambda_\pm$, or $m_\pm$ is appendix-only and contingent on the symmetric $M/M/\infty$ bridge. Table 4 records the headline inputs. The full-sample values are $\sqrt{R_r} = 1.2 \times 10^{-4}\,\text{BTC}^{-1/2}$, $\theta_{\text{eff}} = 0.0477\,\text{h}^{-1}$, implying a descriptive half-life of 14.54 h, and baseline-splice $\beta_{\text{eff}}^{\text{proxy}} = 7.24 \times 10^{-2}\,\text{h}^{-1}$. Nearby-frequency alternatives, 0.0661 and 0.0417, show that the proxy sign is stable while the exact level is splice-sensitive. Additional tail and supplementary diagnostics are reported in Appendix C of the Online Appendix. Public OHLCV aggregates are not used for calibration because the implementation relies on trade-print volume and aggressor-side information; brief reconciliation notes appear in Appendix D of the Online Appendix.

3.2. *Variance–Volume Evidence*

Figure 1 is a visual summary of the empirical relationship between within-day variance and daily trading volume. The grouped scatter averages days within rounded daily volume-percentile bins labeled $0, 10, \ldots, 100$, so the two endpoint bins contain 101 days each and the interior bins contain about 201–202 days each. Within-day variance is computed from hourly close-to-close simple returns within each day; daily volume is total executed BTC volume that day. We do not use log returns anywhere in this test. The heavier evidentiary work comes from the day-level ratio contrast below, the compact grouped-slope companion, and the within-bin dispersion evidence in Appendix C.1 of the Online Appendix. Because that grouped-slope companion is built from daily observations rather than hourly bars, its uncertainty uses a 7-day moving-block bootstrap rather than the 24-hour block bootstrap used for the hourly variance coefficient.

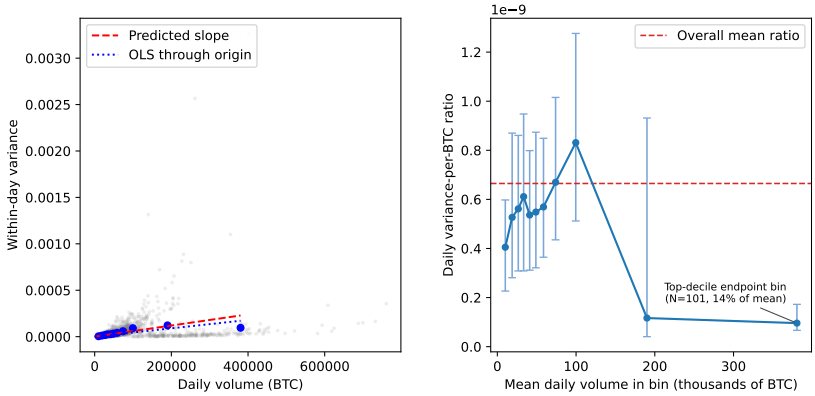

Figure 1. Variance–volume evidence for BTC/USDT, 2020–2025. The left panel shows within-day variance of simple hourly returns against same-day executed BTC volume; gray points are daily observations and blue points are grouped averages formed from rounded daily volume-percentile bins over the 2,017 UTC trading days in the sample. The right panel reports the within-bin daily variance-per-BTC ratio with median and interquartile range by volume bin. The main takeaway is a pooled first-order rise of variance with volume together with a more stable middle-bin ratio pattern and visible attenuation only in the highest-volume endpoint bin.

The visual deviation in Figure 1 is concentrated in the high-volume tail. In the grouped scatter, the highest-volume endpoint bin is visibly attenuated; in a separate day-level contrast, the mean daily ratio in the top 10% of days is only 0.569 times the corresponding ratio in the middle 60% of days. A compact quantitative companion gives the same message: excluding the highest-volume endpoint bin, the grouped n-weighted through-origin slope is 7.027×10^{-10} with 95% block-bootstrap interval $[5.349 \times 10^{-10}, 9.099 \times 10^{-10}]$, versus the theoretical slope 5.996×10^{-10}. The qualitative message is unchanged if all grouped bins are retained (4.581×10^{-10}) or if the top two bins are excluded (7.950×10^{-10}), which is why the endpoint exclusion is best read as a transparent display choice rather than an after-the-fact search for significance. We therefore use the variance–volume evidence as a descriptive first-order compatibility check rather than as a sharp invariance test. Appendix C.1 of the

 doi:10.20944/preprints202604.0256.v1

Online Appendix reports the within-bin dispersion of the daily variance-per-BTC ratio across volume bins, and Table 6 shows that the reduced-form R_r moment also shifts across broad sub-periods. Taken together, those checks support a pooled liquidity-monitoring regularity rather than a time-invariant liquidity law.

Table 5. Compact quantitative companion to the variance–volume figure. The grouped slope excludes only the highest-volume endpoint bin and uses a 7-day moving-block bootstrap.

Companion statistic	Estimate	95% block-bootstrap interval
Grouped slope (all bins)	4.581×10^{-10}	—
Interior grouped slope (bins 0–90)	7.027×10^{-10}	$[5.349 \times 10^{-10}, 9.099 \times 10^{-10}]$
Grouped slope (bins 0–80)	7.950×10^{-10}	—

Table 6. Broad sub-sample reduced-form checks. Including the intermediate 2023 phase shows directly that the reduced-form liquidity and resiliency moments move across major market phases, which is why the paper treats R_r as a pooled first-order diagnostic rather than as a constant law. Hours are reported so the pooled baseline benchmarks are not read as stable venue constants.

Period	Hours	R_r	θ_{eff} (h^{-1})	$\mathbf{E}[v_{1h}]$ (BTC)
2020–2022	26,272	1.63×10^{-8}	0.0471	3,986
2023	8,758	4.94×10^{-9}	0.0850	4,180
2024–2025	13,344	2.29×10^{-8}	0.0336	1,305

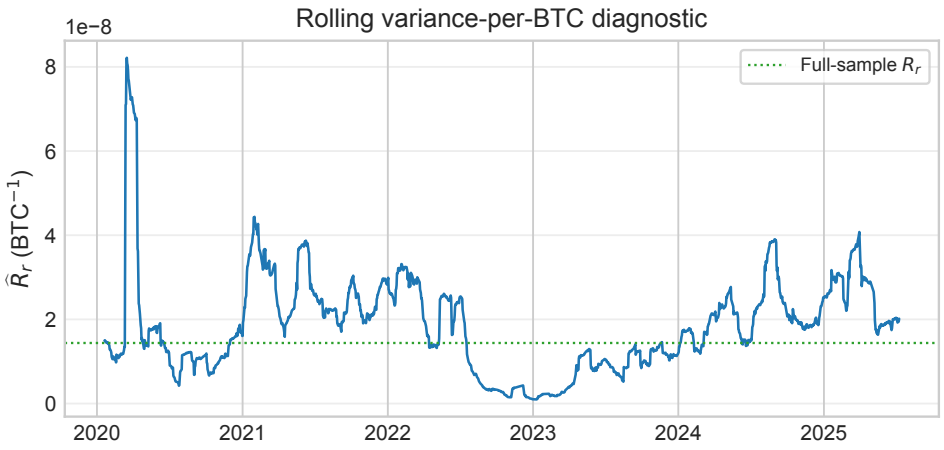

Figure 2. Rolling 30-day estimate of the one-hour variance-per-BTC moment $\widehat{R}_r$ for BTC/USDT, sampled once per UTC day. The line is computed from hourly returns and hourly executed BTC volume inside each trailing 30-day window, and the dashed line marks the full-sample estimate. The main takeaway is that the liquidity diagnostic itself moves over time, so the paper's monitoring interpretation is explicitly dynamic rather than purely pooled.

Figure 2 makes that liquidity movement visible at the same rolling horizon used later for the resiliency diagnostics. The rolling series is noisier than the pooled estimate, but it confirms that the liquidity moment itself is time-varying rather than merely a full-sample summary. We therefore treat R_r as a monitorable liquidity diagnostic whose pooled estimate remains useful for benchmarking, not as a constant law of the venue. In the daily variance–volume scatter, each day d has x-coordinate equal to total daily BTC volume v_d and y-coordinate equal to the within-day sample variance of hourly returns. Since $\mathrm{E}[\widehat{\mathrm{Var}}_d(r_{1h})] \approx \mathrm{Var}(r_{1h})$ while $\mathrm{E}[v_d] \approx 24\ \mathrm{E}[v_{1h}]$, the model's variance–volume identity implies a predicted slope

$$\frac{\mathrm{Var}(r_{1h})}{\mathrm{E}[v_d]} \approx \frac{1}{24}\frac{\mathrm{Var}(r_{1h})}{\mathrm{E}[v_{1h}]} = \frac{R_r}{24}. \tag{24}$$

With the full-sample moments in Table 4, this yields a predicted daily scatter slope of $R_r/24 = 5.996 \times 10^{-10}$. Empirically, the day-by-day ratios $\widehat{\mathrm{Var}}_d(r_{1h})/v_d$ have mean 6.65×10^{-10}, which is our

 doi:10.20944/preprints202604.0256.v1

primary estimate under the constant-ratio hypothesis. The day-level OLS slope through the origin is lower at 4.5×10^{-10} because it effectively volume-weights. High-volume days exhibit attenuated ratios when gross two-sided trading rises faster than the net directional pressure that drives prices. That gross-versus-net reading is plausible, but it should be treated as an interpretation of the attenuation rather than as a separately established result. This attenuation is especially pronounced in the highest-volume endpoint bin (mean ≈380,505 BTC/day, $N = 101$ days), whose average ratio is about 0.38 times the overall mean. One plausible reading is that gross two-sided activity rises faster than net directional pressure on those days, but the paper treats that only as an interpretation of the attenuation rather than as a separate empirical result. The fitted reference line in Figure 1 uses the theoretical slope $R_r/24$; the empirical OLS-through-origin slope is overlaid as a secondary comparison.

3.3. Directional Impact Cross-Check

As an independent empirical cross-check on price impact, we use day-by-day minute signed-flow regressions only as descriptive triangulation and as a normalization device for the feedback proxy, not as a causal impact design or as structural identification. We estimate day-by-day minute-level regressions of one-minute returns on contemporaneous signed BTC order flow. For each day d we run

$$r_{d,m} = a_d + \kappa_d \, q_{d,m} + \varepsilon_{d,m}, \tag{25}$$

where $q_{d,m}$ is signed BTC volume (buyer-initiated minus seller-initiated) computed directly from trade records as $\sum(\texttt{qty} \times \texttt{side})$, with `qty` the Binance base-asset quantity in BTC and $\texttt{side} \in \{+1, -1\}$ indicating trade direction; hence κ_d has units BTC^{-1}. The daily intercept a_d is estimated but not used elsewhere in the paper. Across 2,017 days, the median $|\kappa_d|$ is 1.997×10^{-5} BTC^{-1} with interquartile range $[1.401 \times 10^{-5}, 2.619 \times 10^{-5}]$, and the median R^2 is 0.2946. The absolute-median summary is used only to obtain a positive directional-impact scale for the feedback proxy. In the present sample this choice is not driving the result mechanically: the median signed slope is also 1.997×10^{-5} BTC^{-1}, and negative daily slopes occur on fewer than 0.05% of days. These minute-level slopes are about $6.007\times$ smaller than the observable variance scale $\sqrt{R_r} \approx 1.2 \times 10^{-4}$, which is expected because $\sqrt{R_r}$ is a variance coefficient while κ_d is a directional slope in BTC units. The moderate median R^2 of 0.2946 suggests that minute signed flow carries economically meaningful price information, but the regression is used only to supply the directional normalization in $\beta_{\text{eff}}^{\text{proxy}}$. The separate variance–volume moment condition provides the variance-consistent scale used in the diffusion calculations.

3.4. Variance-Matched One-Hour Skellam Benchmark and Falsification Check

This subsection is intentionally a falsification check rather than a validation exercise. Figure 3 compares empirical hourly returns to a variance-matched one-hour Skellam benchmark. The object benchmarked here is not the exact increment law of the general $M/G/\infty$ framework. Rather, it is the symmetric special-case reference distribution obtained after fixing the one-hour horizon, replacing realized volume by the sample mean one-hour BTC volume, and matching the observed one-hour variance scale R_r. This benchmark intentionally switches off feedback and regime variation so the symmetric core can be judged on its own. Specifically, the theoretical quantiles are generated from

$$r_t^{\text{Sk}} = \bar{r} + \sqrt{R_r}\, Z_t, \qquad Z_t \sim \text{Skellam}(\bar{v}/2, \bar{v}/2),$$

where $\bar{r}$ is the sample mean hourly return and $\bar{v} = \text{E}[v_{1h}]$ is the sample mean hourly traded BTC volume. Figure 3 therefore benchmarks one-hour return increments under a symmetric variance-matched reference distribution; it does not fit the stationary price-level law from Section 2.2.2 directly, and it should be read as an unconditional benchmark rather than as a fully volume-conditioned fit. The sharper main-text evidence appears in Table 7, which conditions on realized one-hour volume. After standardizing returns by $\sqrt{R_r v_t}$, the model-implied central 80% and 50% bands cover 85.3% and 60.7% of standardized observations, with empirical/model width ratios 0.862 and 0.756. Ratios below

 doi:10.20944/preprints202604.0256.v1

one mean that the model-implied central bands are wider than the empirical center. These results reject the symmetric benchmark as a full descriptive model for hourly spot BTC/USDT returns, but that does not make the framework useless: the practical value of the paper lies in monitoring liquidity per traded BTC, local recovery timescales, and signed-flow pressure, while the benchmark still serves as a transparent center-of-distribution reference that shows exactly where an additional tail overlay is needed. Additional burstiness and tail diagnostics are deferred to the Online Appendix.

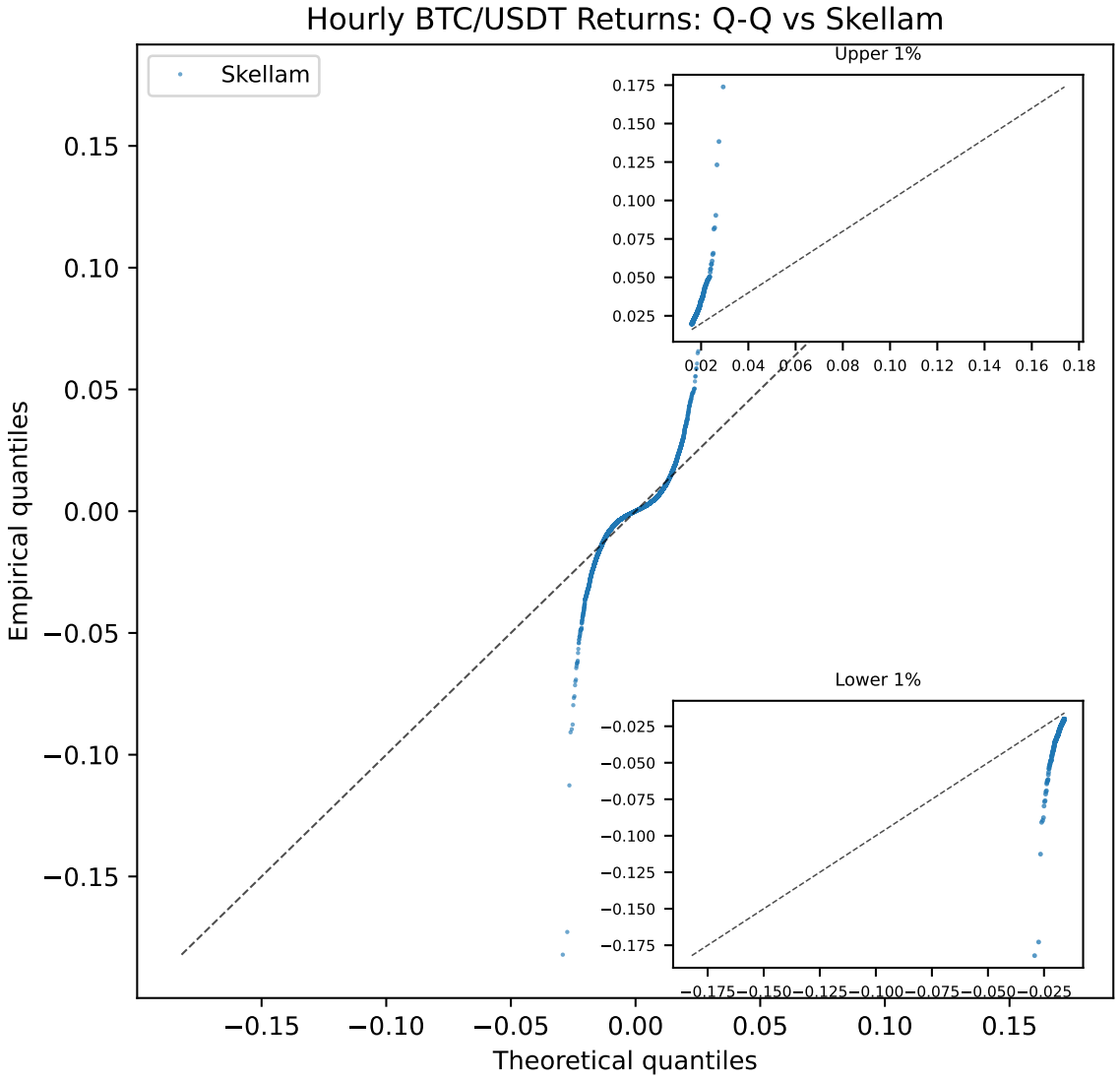

Figure 3. Q–Q comparison of empirical hourly BTC simple returns with a variance-matched Skellam benchmark. The symmetric reference provides only a coarse description of the center and understates the most extreme moves. Table 7 contains the sharper evidence after conditioning on realized one-hour volume.

Table 7. Volume-conditioned middle-quantile check for standardized one-hour returns. Conditioning on realized one-hour BTC volume improves the center-fit comparison relative to the unconditional Q–Q benchmark, but the symmetric core still remains too dispersed. Width ratios below one imply that the model-implied central bands are wider than the empirical center.

Central band	Empirical coverage / model width diagnostic	Value
80% band	Empirical coverage of model-implied band	85.3%
80% band	Empirical/model width ratio	0.862
50% band	Empirical coverage of model-implied band	60.7%
50% band	Empirical/model width ratio	0.756

 doi:10.20944/preprints202604.0256.v1

Even after conditioning on realized one-hour volume, the single-regime symmetric core remains too wide in the center and too thin in the far tails. Appendix C.2 of the Online Appendix and Appendix C.4 of the Online Appendix report the supplementary burstiness and tail diagnostics.

3.5. Rolling Regime Classification Under the Proxy Mapping

The full-sample one-hour signed-BTC-flow slope is $\hat{b}_F = 7250.26$, and combining it with the directional-impact scale $\hat{\kappa}_{\text{dir}} = 2 \times 10^{-5}$ yields a positive companion signed-flow proxy of baseline-splice $\beta_{\text{eff}}^{\text{pf̂oxy}} = 7.24 \times 10^{-2}\,\text{h}^{-1}$ with propagated 95% interval $[0.0613, 0.0837]$. The interval construction is simulation-based: 200,000 draws combine a HAC-normal approximation for b_F with the bootstrap distribution of $\tilde{\kappa}_{\text{dir}}$, so it should be read as sampling uncertainty conditional on the baseline splice. The effective mean-reversion rate remains positive at $\theta_{\text{eff}} = 0.0477\,\text{h}^{-1}$, with interval $[0.0292, 0.0665]$, implying a descriptive full-sample resiliency half-life of 14.54 h. A positive $\beta_{\text{eff}}^{\text{pf̂oxy}}$ by itself does not imply instability: under the local OU splice, the key question is whether mean reversion remains strong enough for $\hat{\theta}_{\text{eff}}$ to stay positive. Nearby-frequency checks leave the sign intact but move the point estimate to 0.0661 and $0.0417\,\text{h}^{-1}$, so the broad practical range across the baseline and nearby alternatives is closer to $[0.0417, 7.24 \times 10^{-2}]$ than to the conditional propagated interval alone. In that sense, the sign of $\beta_{\text{eff}}^{\text{pf̂oxy}}$ is more credible than its exact level. We therefore use the nearby-frequency checks only to show that the splice is not sign-fragile; they do not convert $\beta_{\text{eff}}^{\text{pf̂oxy}}$ into a directly identified structural parameter. Any ρ summary remains secondary and is left to the appendix.

Rolling estimation sharpens that interpretation. We use overlapping 30-day windows as a compromise between local stationarity and enough hourly observations to estimate the return autocorrelation and signed-flow slope without excessive noise. After the initial window-fill period, this yields 47,655 non-missing 30-day window estimates.[2] In the baseline overlapping 30-day scheme, 66.2% of windows have $\hat{\theta}_{\text{eff}} > 0$ by point estimate alone. That point-sign share is informative but too optimistic to stand on its own. Using a simple one-standard-error autocorrelation buffer of ± 0.0373 around zero, 39.2% of overlapping 30-day windows are clearly mean-reverting, 49.4% are ambiguous, and 11.5% are non-positive. This buffer is a heuristic monitoring rule rather than a formal confidence interval under heteroskedastic, serially dependent returns. A more conservative doubled companion buffer leaves only 17.2% of overlapping windows clearly mean-reverting and 79.9% ambiguous. Across those two heuristic buffers, the clearly mean-reverting share therefore lies in a range of roughly 17.2–39.2%, with ambiguity dominant under either rule. The daily-sampled and non-overlapping companion splits are similar: (39.2%, 49.4%, 11.4%) and (38.8%, 49.3%, 11.9%). Appendix C.6 of the Online Appendix reports the full sensitivity tables, including shorter 15-day and longer 60-day windows and the conservative companion buffer. The rolling feedback proxy remains economically meaningful, but the central regime-classification result is the sign and magnitude of $\hat{\theta}_{\text{eff}}$, not the behavior of any derived boundary ratio.

Figure 4 reports the rolling feedback and mean-reversion diagnostics. The figure should be read first as a regime-classification device based on $\hat{\theta}_{\text{eff}}$, with the rolling feedback proxy shown alongside it. Taken together, the full-sample and rolling results are consistent with return-following signed-flow pressure in BTC/USDT, but the mean-reverting component dominates only in a subset of windows once a near-zero heuristic buffer is imposed. The windows with ambiguous or non-positive $\hat{\theta}_{\text{eff}}$ are economically important precisely because they identify periods in which the baseline queueing interpretation becomes unreliable.

[2] "Non-missing" means windows for which the rolling return autocorrelation and signed-flow regression are both computable once the observed missing UTC hours are carried through the hourly grid. The count therefore differs slightly from the purely mechanical number of overlapping 30-day calendar windows. The sample contains 34 missing UTC hours overall, so a small number of windows lose one or more hourly observations even though the rolling clock continues to advance.

 doi:10.20944/preprints202604.0256.v1

3.6. Simple Next-Day Monitoring-Value Check

To check whether the rolling diagnostics do more than summarize the same-day sample, we run a deliberately simple day-ahead sort. For each UTC day d, we take the end-of-day 30-day rolling liquidity diagnostic $\widehat{R}_r(d)$ together with the end-of-day rolling regime label from Figure 4, then compare them with realized outcomes on day $d+1$. The matched sample starts only once both predictors are fully filled 30-day end-of-day diagnostics. The outcomes are next-day realized variance of hourly returns, the next-day maximum absolute hourly return, and a tail-day indicator equal to one when the next-day maximum absolute hourly return falls in the top decile of the matched sample. This is not a forecast horse race and does not claim optimal monitoring thresholds; it is a descriptive check of whether the rolling outputs sort meaningfully different next-day risk conditions.

Table 8. Simple next-day monitoring-value check based on end-of-day 30-day rolling diagnostics. Predictors are measured at the end of UTC day d, and outcomes are computed from the observed hourly returns on day $d+1$. The matched sample contains 1,987 predictor-outcome day pairs, running from 30 January 2020 through 8 July 2025. We omit the first 29 UTC days so that both end-of-day predictors are fully filled 30-day diagnostics and omit the final sample day because it has no $d+1$ outcome; 15 matched days contain fewer than 24 hourly returns because of the observed missing UTC hours. "Tail day" means that the next-day maximum absolute hourly return lies in the top decile of the matched sample, with threshold 2.89%. Because the predictors are persistent rolling states, this table is included only as a descriptive relevance check and does not establish incremental forecast evidence over generic volatility persistence.

Predictor bucket	Days	Next-day realized variance	Next-day max $\lvert r_{1h}\rvert$	Next-day tail-day rate
End-of-day rolling R_r tercile				
Low	663	0.000026	1.31%	6.5%
Mid	661	0.000044	1.56%	8.5%
High	663	0.000072	1.98%	15.1%
End-of-day rolling resiliency regime (baseline heuristic buffer)				
Clear positive	779	0.000048	1.53%	8.9%
Ambiguous	982	0.000046	1.63%	10.2%
Non-positive	226	0.000054	1.86%	13.3%

Table 8 shows that the rolling liquidity diagnostic carries the clearest next-day signal. Moving from the low to the high rolling-R_r tercile raises next-day realized variance from 0.000026 to 0.000072, increases the next-day mean maximum absolute hourly return from 1.31% to 1.98%, and raises the next-day tail-day rate from 6.5% to 15.1%. The resiliency classification adds a weaker but directionally similar signal: non-positive windows are followed by larger next-day maximum hourly moves and higher tail-day rates than clearly mean-reverting windows (13.3% versus 8.9%). Under the doubled companion buffer from Appendix C.6 of the Online Appendix, the clearly classified sample shrinks sharply (347 clear-positive days and 59 non-positive days), but the same tail-rate ordering remains directionally intact (15.3% versus 6.9%). These patterns do not validate a trading rule or establish predictive superiority over a generic volatility-persistence monitor, but they do show that the rolling diagnostics carry useful next-day monitoring information rather than functioning only as contemporaneous summaries.

3.7. Secondary One-Hour Tail Overlay

For the empirical risk application, we stay in reduced-form observable notation and at the one-hour horizon only. Let $\bar{v} = \mathrm{E}[v_{1h}]$ denote average one-hour BTC volume, and let $\widehat{V}_{1h} = \mathrm{Var}(r_{1h}) = R_r\bar{v}$ denote the observed one-hour return variance. The main-text core benchmark is therefore the one-hour Normal approximation with variance $\widehat{V}_{1h}$. To monitor the most extreme tail events, we add a threshold-split Normal-plus-Student-t overlay re-fit every 24 hours inside trailing 2160-hour windows. Over 46214 hourly forecasts, the rolling Gaussian benchmark breaches at 1.80% and the stress-state overlay at 1.13%. We include that overlay as a supporting parametric tail monitor; fuller benchmark

 doi:10.20944/preprints202604.0256.v1

comparisons and backtesting details remain in Appendix E of the Online Appendix and Appendix C.10 of the Online Appendix.

4. Discussion

4.1. Short-Horizon Price Formation and Liquidity Interpretation

The results support a market-microstructure interpretation in which inventory and trade imbalances shape short-horizon BTC/USDT price variation on Binance. The variance–volume relation shows that gross trading activity and short-horizon variance move together in a disciplined way. At the same time, the BTC-flow diagnostics are consistent with positive returns being followed, on average, by additional buyer-initiated BTC flow at the hourly horizon. In practical terms, this means that high activity can raise volatility through both gross two-sided trading and directional signed-flow pressure, even though full-sample return autocorrelation remains negative.

The queueing formulation is useful because, inside the model, arrival and holding-time assumptions organize occupancies, occupancies organize net imbalance, and net imbalance organizes the local price-deviation state. Empirically, the paper identifies monitoring diagnostics tied to that structure rather than the underlying queue primitives themselves. That decomposition explains what would be lost without queueing: the observable bundle would no longer live in one stock-flow state space, the recovery interpretation of θ_{eff} would lose its turnover counterpart, and the pooled occupancy translation would disappear. The model adds value by providing a disciplined accounting system that connects observable market states to short-horizon price formation, volatility, and resiliency.

4.2. Market Monitoring and Risk Application

From a monitoring perspective, the strongest outputs are the rolling liquidity path in Figure 2, the rolling resiliency classification in Figure 4, and the next-day descriptive sort in Table 8. A rising rolling R_r indicates more variance per traded BTC, a longer descriptive half-life indicates slower local recovery, and ambiguous or non-positive $\hat{\theta}_{\text{eff}}$ flags periods in which resiliency is weaker or harder to classify. Appendix C.7 of the Online Appendix points in the same direction for isolated large hourly shocks: clear-positive windows reverse more over the next 24 hours than ambiguous or non-positive windows.

The rolling $\beta_{\text{eff}}^{\text{proxy}}$ series is best read as a companion signed-flow gauge rather than as a standalone stability estimate. Its role is to sit beside $\hat{\theta}_{\text{eff}}$ and help distinguish momentum-heavy from more resilient market phases. Queueing matters here because it forces the rolling liquidity, resiliency, and signed-flow diagnostics to live on one occupancy-accounting scale rather than as separate alarms.

Who uses this dashboard

The practical use is concrete. A market maker or execution desk can use R_r to see whether the venue is delivering less turnover-adjusted liquidity than usual, θ_{eff} to judge whether directional pressure appears to unwind quickly or only weakly, and $\beta_{\text{eff}}^{\text{proxy}}$ to distinguish quiet deterioration from visibly one-sided flow. An exchange risk or surveillance team can use the same trio to separate ordinary busy trading from states in which liquidity is thin, resiliency is fragile, and next-day risk outcomes are worse. That is why the framework remains useful even though the heaviest unconditional tails call for an additional overlay.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2026 doi:10.20944/preprints202604.0256.v1

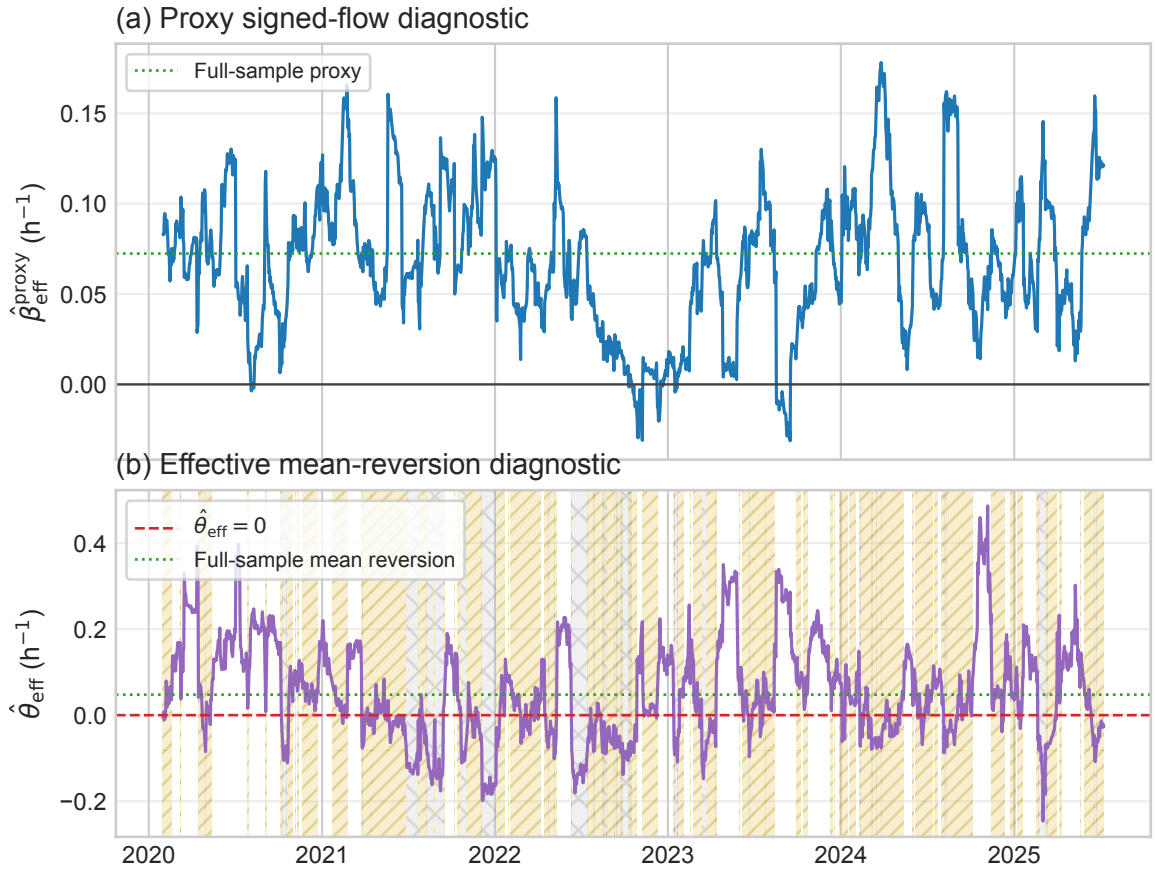

Figure 4. Rolling feedback and mean-reversion diagnostics for BTC/USDT, Binance, 2020–2025. The underlying estimates are computed in overlapping 30-day windows updated each hour after the initial 30-day fill period, but the figure samples one end-of-day point per UTC day for display clarity. Panel (a) reports the rolling companion proxy $\hat{\beta}_{\text{eff}}^{\text{proxy}}$, and panel (b) reports the rolling effective mean-reversion estimate $\hat{\theta}_{\text{eff}}$. Under the baseline heuristic buffer, the hourly-updated 30-day split is 39.2% clear positive, 49.4% ambiguous, and 11.5% non-positive. Hatched amber shading marks ambiguous windows and cross-hatched gray shading marks non-positive windows, where the local OU interpretation is not used.

The tail lesson is separate and simple: the queueing core is useful for liquidity interpretation, resiliency monitoring, and a transparent benchmark for the center of one-hour returns, while the heaviest tails require an additional overlay.

4.3. Limitations

The paper has several limitations. It is a single-venue Binance spot study even though Bitcoin trading is multi-venue; it uses a linear impact approximation rather than richer nonlinear or finite-depth dynamics; it assumes independent latent arrivals and i.i.d. holding times; and its feedback layer uses a minute-impact and hourly-flow splice as a proxy for latent directional pressure rather than directly observing customer-level queues. The stationary core is also designed for normal and moderately stressed periods rather than every crisis tail event, which is why the tail overlay remains separate. These choices keep the framework transparent and implementable, but they also bound the scope of the claims. Natural extensions include multi-venue implementations, nonlinear impact, and marked or compound-Poisson queueing generalizations.

What would weaken the framework

Three empirical patterns would materially weaken the interpretation advanced here. First, the framework would lose credibility if the variance-per-BTC relation became persistently unstable across volume bins or over rolling windows after using the same data construction. Second, it would be

 doi:10.20944/preprints202604.0256.v1

harder to maintain the resiliency interpretation if proxy-based feedback remained persistently strong while $\hat{\theta}_{\text{eff}}$ repeatedly became small or non-positive across rolling windows. Third, the queueing core would be too restrictive if it failed not only in the far tails but also in the middle quantiles of one-hour return increments once realized volume and horizon were fixed. Those are direct ways in which future evidence could narrow or overturn the usefulness of the present specification.

5. Conclusions

This paper develops a queueing-organized framework for within-venue monitoring of BTC/USDT liquidity and resiliency on Binance. The practical outputs are three linked diagnostics: the variance-per-BTC liquidity measure R_r, the effective mean-reversion rate θ_{eff}, and the companion signed-flow proxy $\beta_{\text{eff}}^{\text{proxy}}$. Using Binance trade data from 2020–2025, we find a pooled first-order variance–volume regularity away from the highest-volume tail, rolling liquidity and resiliency measures that change materially across market phases, and a next-day descriptive sort in which high-R_r days and non-positive resiliency windows are followed by worse risk outcomes than quieter and clearly mean-reverting states.

The paper's contribution is to place those diagnostics inside one stock-flow bookkeeping system. Queueing is useful here because it ties liquidity, signed-flow pressure, and recovery timescales to the same occupancy accounting logic. That common state space is what turns three separate reduced-form monitors into one interpretable venue-level monitoring framework. The intended audience is any reader who must monitor venue conditions from trade prints: market makers, execution desks, exchange surveillance teams, and short-horizon risk managers. For that audience, the payoff is a compact dashboard with a coherent economic interpretation rather than three disconnected time series. The symmetric one-hour benchmark and the tail overlay are supporting checks; the core contribution is the linked interpretation of R_r, θ_{eff}, and $\beta_{\text{eff}}^{\text{proxy}}$.

Scope and falsifiability

The paper remains deliberately focused. It is a single-venue study, it uses a linear-impact approximation, and it does not attempt an exhaustive comparison with alternative microstructure models. Within that scope, the framework should stand or fall on observable margins: a persistent breakdown of the variance-per-BTC relation, repeated failure of effective mean reversion in rolling windows, or severe failure of the queueing core even as a center benchmark once realized one-hour volume is fixed would all count against the specification rather than as minor calibration noise.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. The Online Appendix PDF provides additional derivations, model extensions, Monte Carlo validation, nearby-frequency proxy checks, rolling-window sensitivity diagnostics, and the recovery event study.

Author Contributions: Conceptualization, S.V.; Methodology, S.V.; Software, S.V.; Formal Analysis, S.V.; Investigation, S.V.; Writing—Original Draft, S.V.; Writing—Review and Editing, S.V. The author has read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw Binance BTC/USDT trade data are publicly available through the Binance historical data portal. The empirical results in this paper are generated from those public trade data using a local workflow that constructs hourly and aggressor-side summaries and then reproduces the reported panels, figures, and macro files. Two public code archives are available. The broader project code base is archived at 10.5281/zenodo.18644286. The code-only workflow snapshot used for the computations reported in this manuscript is archived at 10.5281/zenodo.19335853. That workflow archive contains the Python code and

 doi:10.20944/preprints202604.0256.v1

documentation needed to reproduce the computational pipeline for this manuscript, but it does not include raw Binance trade files or manuscript source files.

Acknowledgments: The author is grateful for computing resources provided by VS Asset Management, LLC.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Additional Discussion

Appendix A.1. Queueing Theory Primer

Queueing Theory Essentials

Queueing theory studies systems where units arrive randomly, receive service, and depart. In finance, we reinterpret these elements:

- **Units** $\rightarrow$ BTC-normalized latent occupancy units on the buy and sell sides
- **Arrivals** $\rightarrow$ New latent occupancy units opened at rate λ
- **Service time** $\rightarrow$ Holding period H until that latent occupancy exits
- **Queue length** $\rightarrow$ Outstanding latent occupancy

The $M/G/\infty$ notation describes our system:

- M = Markovian (Poisson) arrivals
- G = General service time distribution
- ∞ = Infinite servers (no capacity constraints)

Key Result: In steady state, if latent occupancy units arrive at rate λ and are held for average time $\mathbb{E}[H]$, then outstanding occupancy follows a Poisson law with mean $m = \lambda\mathbb{E}[H]$ (Palm's occupancy result for $M/G/\infty$ queues; Little's Law gives the same mean). In the manuscript, occupancies are BTC-normalized, so one latent unit corresponds by normalization to 1 BTC.

Poisson Arrivals See Time Averages (PASTA) ensures that sample-path observations of arrivals match time averages in $M/G/\infty$ systems; see Wolff (1982). This framework naturally captures how latent buy- and sell-side occupancy accumulates and dissipates in markets, providing the mathematical foundation for our pricing model.

Table A1. Assumption ladder used in the paper. Each layer adds one extra assumption and delivers a narrower object than the layer above it.

Layer	Extra assumption	What it delivers	What it does not establish
$M/G/\infty$ stationary layer	Independent Poisson arrivals and i.i.d. holding times	Stationary occupancy law, Skellam imbalance, variance–volume bookkeeping	One-hour return law or literal queue openings from public trades
Symmetric $M/M/\infty$ one-hour bridge	Symmetry plus exponential holding times	Contingent one-hour increment bridge, $\mathrm{E}[H]$, $\lambda_{\pm}$, $m_{\pm}$, and $\tilde{\kappa}$ illustration	Structural identification of those quantities from observables alone
Local OU / proxy splice	Local linearization plus minute/hour splice for signed-flow proxy	Descriptive pair $(\theta_{\text{eff}}, \beta_{\text{eff}}^{\text{proxy}})$ for rolling regime classification	A directly identified latent feedback boundary or exact queue stability law

Appendix A.2. Applicability of Queueing Assumptions to Bitcoin

Our queueing framework is most useful when four conditions are approximately present: (1) *Limited direct intraday anchoring to externally observed fundamentals*; (2) *Order-flow-dominated trading at intraday*

 doi:10.20944/preprints202604.0256.v1

horizons; (3) *A homogeneous traded asset*; (4) *Continuous high-turnover trading*. Bitcoin is a reasonable, though not unique, empirical setting for those conditions.

Appendix B. Additional Model Extensions

The extensions in this section are supplementary heuristic sketches and are not used in the main-text empirical implementation. They indicate only how the baseline queueing framework could be generalized without changing the observable monitoring strategy used in the paper.

Overview. This section first records the one-hour bridge, contingent symmetry translations, splice sensitivity, and reduced-form variance benchmark used elsewhere in the paper, and then adds three brief illustrative extensions: discrete time and price grids in B.1, together with nonlinear or finite-depth impact sketches in B.5 and B.6. All of these items are supplementary and do not add separate calibrated claims.

Appendix B.1. Discrete Time and Price Grids

Real-world trading is discrete in both time and price. The results derived in continuous time survive intact once (i) clock time is sliced into short intervals of length Δ and (ii) prices are snapped to the nearest exchange tick.

Appendix B.1.1. Discrete Time

Fix a step of $\Delta = 60\,\text{s}$ ($\Delta = 1/60\,\text{h}$).[3] At tick n, the numbers of newly opened long- and short-side latent occupancy units, $A_{n,+}$ and $A_{n,-}$, are independent Poisson variates:

$$A_{n,+} \sim \text{Poisson}(\lambda_+ \Delta) = \text{Poisson}(13.67), \tag{A1}$$

$$A_{n,-} \sim \text{Poisson}(\lambda_- \Delta) = \text{Poisson}(13.67), \tag{A2}$$

where the BTC-normalized contingent symmetry translation in Appendix B.3 implies $\lambda_+ \approx \lambda_- \approx 820\,\text{h}^{-1}$ for the 2020–2025 Binance sample Binance (2025). Because the mean holding time is $\mathbb{E}[H] \approx 5.20\,\text{h}$, an existing occupancy unit survives one minute with probability

$$p_+ = p_- = 1 - \frac{\Delta}{\mathbb{E}[H]} = 1 - \frac{1/60}{5.20} \approx 0.996792. \tag{A3}$$

These dynamics define a $GI/\text{Geo}/\infty$ queue whose steady-state occupancies equal those of the continuous-time model:

$$B_\infty \sim \text{Poisson}(m_+ \approx 4{,}262), \qquad S_\infty \sim \text{Poisson}(m_- \approx 4{,}262). \tag{A4}$$

Running example

In the BTC-normalized contingent symmetry translation used throughout the appendix, discretisation leaves $m_\pm$ unchanged, so the net inventory remains $X_\infty \sim \text{Skellam}(4{,}262, 4{,}262)$ and the same variance–volume relations continue to hold.

Appendix B.1.2. Price Grid

Let $\delta P = 0.01\,\text{USD}$ be the BTC/USDT tick size and $q_0 = 0.001\,\text{BTC}$ the minimum trade lot. Grid-aligned impact is implemented as

$$P_n = P_0 + \delta P \,\lfloor \kappa X_n / \delta P \rceil, \tag{A5}$$

[3] $\Delta = 60\,\text{s}$ keeps the Poisson means in the low teens, striking a good balance between numerical accuracy and computational speed. If finer granularity is desired, set $\Delta = 1\,\text{s}$; all formulas below continue to apply after replacing $\lambda_\pm \Delta$ accordingly.

 doi:10.20944/preprints202604.0256.v1

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer and with the structural impact slope $\tilde{\kappa} = 1.26 \times 10^{-4}\,\mathrm{BTC}^{-1}$ (so $\kappa = \bar{P}\tilde{\kappa} \approx \$6.289\ \mathrm{BTC}^{-1}$ for $\bar{P} = \$50{,}000$). At $P_0 = 50{,}000\,\mathrm{USD}$, one tick is a mere $\delta P/P_0 = 2 \times 10^{-7} = 0.002\,\mathrm{bp}$, so linearisation error is negligible: an inventory change of $\delta P/\kappa \approx 1.59 \times 10^{-3}$ BTC (using $\bar{P} = \$50{,}000$) contracts moves the mid-price by exactly one tick.

Summary

Discretising time at one-minute intervals and snapping prices to the one-cent grid leaves the Poisson–Skellam backbone largely intact. In the present calibration, the continuous-time formulas remain adequate local approximations and any grid effects are second-order.

Appendix B.2. One-Hour Increment Bridge

The main text maps the observable variance-per-BTC moment into a structural return-impact slope $\tilde{\kappa}$. This subsection gives the one-hour bridge explicitly under the symmetric $M/M/\infty$ mapping used for that contingent translation. The general $M/G/\infty$ occupancy result from Palm's theorem still underlies the stationary inventory law; the extra step here is narrower because the increment covariance formula below uses exponential holding times.

Let B_t and S_t denote stationary long and short occupancies with common mean holding time $\mathbb{E}[H]$ and common mean occupancy $m = \lambda\mathbb{E}[H]$. For an $M/M/\infty$ queue,

$$\mathrm{Cov}(B_t, B_{t+h}) = me^{-h/\mathbb{E}[H]}, \qquad \mathrm{Cov}(S_t, S_{t+h}) = me^{-h/\mathbb{E}[H]}.$$

Therefore, for the net imbalance $X_t = B_t - S_t$,

$$\mathrm{Var}(X_{t+h} - X_t) = 2\,\mathrm{Var}(X_t) - 2\,\mathrm{Cov}(X_{t+h}, X_t) = 4m\big(1 - e^{-h/\mathbb{E}[H]}\big).$$

Using $m = \lambda\mathbb{E}[H]$ and the symmetric executed-volume identity $\mathbb{E}[\nu_h] = 4\lambda h$, the return increment $r_{t,h} \approx \tilde{\kappa}(X_{t+h} - X_t)$ satisfies

$$\frac{\mathrm{Var}(r_{t,h})}{\mathbb{E}[\nu_h]} = \tilde{\kappa}^2\,\frac{\mathbb{E}[H]\big(1 - e^{-h/\mathbb{E}[H]}\big)}{h}. \tag{A6}$$

At the one-hour horizon used in the paper ($h = 1$),

$$\frac{\mathrm{Var}(r_{1h})}{\mathbb{E}[\nu_{1h}]} = \tilde{\kappa}^2\,c_H, \qquad c_H := \mathbb{E}[H]\big(1 - e^{-1/\mathbb{E}[H]}\big). \tag{A7}$$

Equation (A7) is the bridge used for the symmetry-based structural translation in the main text and in the appendix running examples. The occupancy units in this bridge are latent and BTC-normalized. The bridge should therefore not be read as a one-to-one mapping from public trades to literal queue openings or exits; executed BTC enters only through the first-moment bookkeeping identity for $\mathrm{E}[\nu_h]$.

Appendix B.3. Additional Contingent Symmetry Translations

Under the same symmetric $M/M/\infty$ special-case translation used in the main text, mean executed one-hour volume implies baseline directional intensities

$$\lambda_+ = \lambda_- = \frac{\mathrm{E}[\nu_{1h}]}{4} \approx 820\ \mathrm{BTC/h}.$$

Combining this with the translated holding time $\mathrm{E}[H] \approx 5.20\,\mathrm{h}$ yields contingent occupancies

$$m_+ = m_- = \lambda_\pm\,\mathrm{E}[H] \approx 4{,}262\ \mathrm{BTC},$$

where queue occupancies are BTC-normalized throughout, so one latent inventory unit corresponds by normalization to 1 BTC. The bridge to executed BTC volume is only a first-moment bookkeeping convention: the normalization is chosen so that expected openings and closures of latent occupancy

 doi:10.20944/preprints202604.0256.v1

units match expected executed BTC turnover in the symmetric identity above. A marked or compound-Poisson extension would preserve that first-moment bridge while changing higher moments. These values are reported here only to make the symmetry arithmetic transparent. They are not separately identified empirical estimates.

Appendix B.4. Splice Sensitivity of the Contingent Translation

Because the contingent symmetry translation inherits the minute/hour proxy splice, the implied holding times and occupancies move when nearby-frequency alternatives are substituted into $\hat{\beta}_{\text{eff}}^{\text{proxy}}$. Table A2 reports the baseline splice together with the 5-minute directional and 2-hour flow alternatives. The point is only interpretive: the contingent translation remains useful for separating turnover from recovery, but the exact queue numbers should not be read as stable estimates when the proxy splice itself shifts. Across these nearby variants, the implied holding time ranges from about 5.20 to 7.63 hours and the total latent occupancy from about 8,524 to 12,510 BTC-normalized units.

Table A2. Sensitivity of the contingent symmetry translation to nearby-frequency proxy-splice alternatives. These values are illustrative arithmetic under the symmetric bridge, not directly identified structural estimates.

Splice variant	$\hat{\beta}_{\text{eff}}^{\text{proxy}}$ (h^{-1})	Contingent $\mathbb{E}[H]$ (h)	Contingent $m_+ + m_-$
Baseline splice	0.0724	5.20	8,524
5-minute directional alternative	0.0661	5.56	9,116
2-hour flow alternative	0.0417	7.63	12,510

Appendix B.5. Illustrative Nonlinear Price Impact

The linear price impact assumption can be relaxed to accommodate realistic order book dynamics while maintaining substantial analytical tractability.

Appendix B.5.1. Exponential Impact

For multiplicative price effects, we use:

$$P_t = P_0 \exp(\tilde{\kappa} X_t) \tag{A8}$$

Since $X_\infty \sim \text{Skellam}(m_+, m_-)$, all moments of P_∞ remain closed-form. The k-th moment is:

$$\mathbb{E}[P_\infty^k] = P_0^k \exp\left[-(m_+ + m_-) + m_+ e^{k\tilde{\kappa}} + m_- e^{-k\tilde{\kappa}}\right] \tag{A9}$$

The log-price $Y_t = \ln P_t = \ln P_0 + \tilde{\kappa} X_t$ has variance $\tilde{\kappa}^2(m_+ + m_-)$, preserving the volume–volatility relationship for log-returns. *Note:* This section discusses log-*prices* only; all empirical returns used elsewhere are arithmetic.

Appendix B.5.2. Saturating Impact (Hyperbolic Tangent)

For bounded price impact that saturates at extreme inventory levels:

$$P_t = P_0 + \kappa P_0 \tanh(X_t / L) \tag{A10}$$

where $L > 0$ controls the saturation scale.

Moments exist as finite sums over the Skellam support:

$$\mathbb{E}[P_\infty] = P_0 + \kappa P_0 \sum_{k=-\infty}^{\infty} \tanh(k/L) \cdot \Pr\{X_\infty = k\} \tag{A11}$$

 doi:10.20944/preprints202604.0256.v1

For $L \gg \sqrt{m_+ + m_-}$, a Taylor expansion yields:

$$\mathbb{E}[P_\infty] \approx P_0 + \frac{\kappa P_0}{L}(m_+ - m_-) - \frac{\kappa P_0}{3L^3}[(m_+ - m_-)^3 + 3(m_+ - m_-)(m_+ + m_-)]. \tag{A12}$$

The cubic correction can introduce additional kurtosis relative to the linear model. Under the Binance full-sample calibration, these nonlinear transforms mainly matter in larger imbalance states and are therefore kept here only as illustrative alternatives rather than as separate empirical objects.

Appendix B.6. Illustrative Finite Order Book Depth

Real order books have finite depth, leading to state-dependent price impact. We model this through piecewise-linear impact functions that remain analytically tractable.

Appendix B.6.1. Piecewise Linear Impact Schedule

Define the impact function:

$$\kappa(X) = \begin{cases} \kappa_1, & |X| < X_1 \\ \kappa_2, & X_1 \le |X| < X_2 \\ \kappa_3, & |X| \ge X_2 \end{cases} \tag{A13}$$

with $0 < \kappa_1 < \kappa_2 < \kappa_3$ reflecting deeper impact at higher inventory levels.

Appendix B.6.2. Price Distribution with Depth Tiers

The price variance decomposes as:

$$\mathrm{Var}[P_\infty] = \sum_{r=1}^{3} \kappa_r^2 \sum_{k \in \mathcal{K}_r} k^2 \Pr\{X_\infty = k\} - (\mathbb{E}[P_\infty] - P_0)^2 \tag{A14}$$

where $\mathcal{K}_1 = \{k : |k| < X_1\}$, $\mathcal{K}_2 = \{k : X_1 \le |k| < X_2\}$, and $\mathcal{K}_3 = \{k : |k| \ge X_2\}$.

Each truncated sum uses the Skellam PMF. For efficient evaluation, use the stable ratio recursion

$$\frac{\Pr\{X = k+1\}}{\Pr\{X = k\}} = \sqrt{\frac{m_+}{m_-}} \frac{I_{k+1}(2\sqrt{m_+ m_-})}{I_k(2\sqrt{m_+ m_-})}, \quad k \ge 0, \tag{A15}$$

with the symmetric formula for negative k.

Under the Binance full-sample calibration, this finite-depth illustration leaves the effective linear impact close to the top-of-book slope because most stationary mass remains in the first two depth tiers. The main implication is therefore qualitative: finite depth matters primarily in larger imbalance states and is not separately estimated in the paper.

Appendix B.7. Reduced-Form One-Hour Variance Benchmark

The main text now keeps the risk application at the one-hour horizon only. The relevant observable object is therefore

$$\widehat{V}_{1h} := \mathrm{Var}(r_{1h}) = R_r\, \bar{v}, \qquad \bar{v} = \mathrm{E}[v_{1h}],$$

which is the directly observed one-hour return variance used to normalize both the Gaussian benchmark and the threshold-split stress overlay. The local OU approximation and θ_{eff} remain useful as resiliency diagnostics, but the paper no longer treats them in the appendix as a separate multi-horizon return-law derivation.

Using the Binance calibration, the one-hour benchmark standard deviation is $\sigma_r \approx 0.69\%$ (about \$344 at $P^* = \$50{,}000$), and the corresponding descriptive half-life from the local mean-reversion

 doi:10.20944/preprints202604.0256.v1

estimate remains $t_{1/2} = \ln 2/\theta_{\text{eff}} = 14.54\,\text{h}$. This is the only variance object required by the main-text VaR overlay.

Appendix C. Additional Empirical Results and Robustness

Appendix C.1. Volume-Conditioned Dispersion of the Variance-Per-BTC Ratio

Figure A1 reports the within-bin dispersion of the daily variance-per-BTC ratio across the same volume bins used in the main-text scatter. The central bins cluster around the overall mean ratio, while the highest-volume bin is visibly attenuated. This is why the main text describes the variance–volume evidence as first-order compatibility rather than as a literal invariance result.

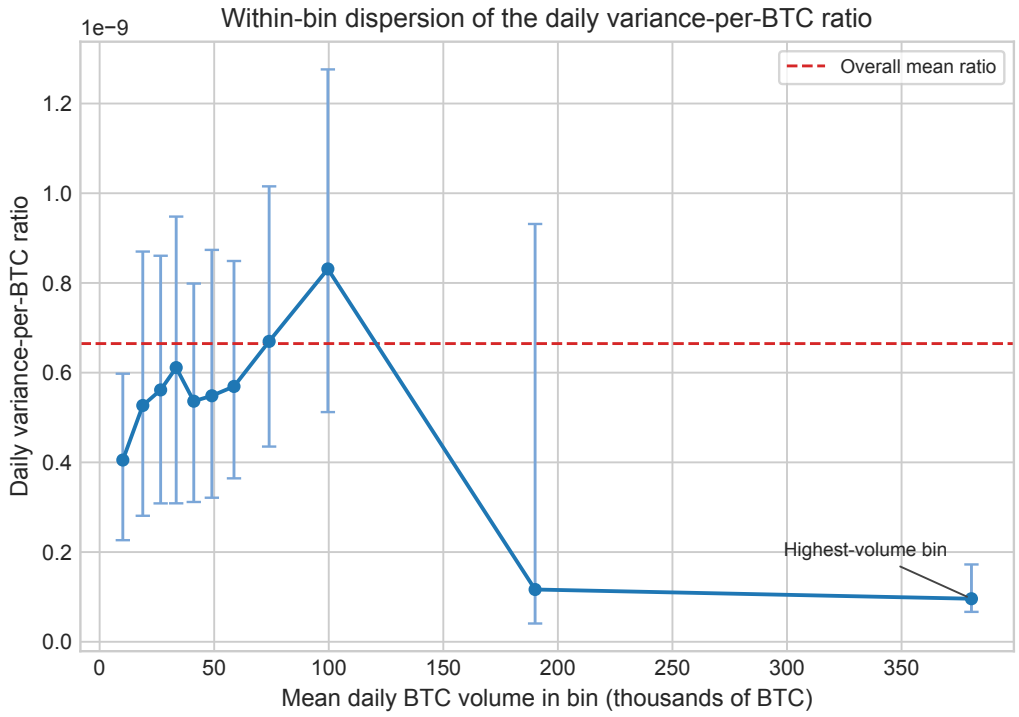

Figure A1. Within-bin dispersion of the daily variance-per-BTC ratio across volume bins for BTC/USDT, Binance, 2020–2025. Points show the within-bin median ratio and bars show the interquartile range. The dashed line is the overall mean ratio. The main takeaway is that the central bins cluster around a fairly stable ratio, while the highest-volume bin shows clear attenuation.

Appendix C.2. Volume-Conditioned Variance-Matched Increment Benchmark

The main text's Q–Q plot fixes only the horizon and the sample-average hourly volume. Here we condition more tightly on realized one-hour volume. This subsection still studies a variance-matched benchmark rather than the exact increment law of the general $M/G/\infty$ framework. For each hour t, define the standardized increment

$$z_t := \frac{r_t - \bar{r}}{\sqrt{R_r\, \nu_t}},$$

where r_t is the simple one-hour return, $\bar{r}$ is the sample mean one-hour return, R_r is the observable variance-per-BTC moment, and ν_t is realized one-hour BTC volume. Under the symmetric queueing benchmark, the corresponding conditional law is

$$z_t^* = \frac{Z_t}{\sqrt{\nu_t}}, \qquad Z_t \mid \nu_t \sim \text{Skellam}(\nu_t/2, \nu_t/2).$$

 doi:10.20944/preprints202604.0256.v1

Using the observed ν_t sequence over 48,374 hourly bars and 8 simulated draws per hour, the empirical standardized middle quantiles are

$$(q_{0.10}, q_{0.25}, q_{0.75}, q_{0.90}) = (-1.097, -0.504, 0.514, 1.114),$$

versus model-implied quantiles

$$(-1.284, -0.675, 0.671, 1.281).$$

The model-implied central 80% band therefore covers 85.3% of standardized observations, and the model-implied central 50% band covers 60.7%. The corresponding empirical/model width ratios are 0.862 for the 80% band and 0.756 for the 50% band. Ratios below one mean that the model-implied central bands are wider than the empirical center. Volume conditioning thus improves the middle-quantile comparison relative to the unconditional benchmark in the main text, but the symmetric variance-matched core still overstates central dispersion somewhat and does not eliminate tail misspecification.

Appendix C.3. Sub-Sample Reduced-Form Checks

The main paper now reports 2020–2022, 2023, and 2024–2025 side by side so the reader can see directly that the reduced-form liquidity and resiliency moments move across major market phases. This appendix keeps the earlier endpoint contrast only as a compact descriptive comparison between the pre-transition window and the later ETF-era window; 2023 is treated in the main paper as the transition year rather than excluded from the evidentiary record.

Interpretation

The earlier sub-sample windows show somewhat heavier trading intensity (3,986 BTC versus 1,305 BTC per one-hour bar), while the later window shows a larger variance-per-BTC estimate (1.4× the 2020–2022 value) together with slower effective mean reversion (0.0336 versus 0.0471). Those comparisons are descriptive rather than structural. Once the symmetry mapping is re-imposed on short sub-samples, contingent quantities such as $\mathbb{E}[H]$, $\lambda_{\pm}$, and $m_{\pm}$ can move substantially even when the directly observed moments change only moderately. For that reason, the main text keeps the structural translations at the full-sample level and uses the sub-sample checks only to show that the reduced-form diagnostics remain regime-dependent.

Appendix C.4. Addressing the Tail Risk Challenge

Scope

The Skellam (or NB–Skellam) core provides a useful baseline for normal trading periods and a coarse description of the center of the unconditional one-hour increment distribution. The far tails during stress periods necessarily require an additional component—such as a core–Normal / tail–Student-t mixture—to improve tail coverage relative to the Gaussian benchmark.

The one-hour return series has empirical excess kurtosis 47.29. By comparison, the count-space Poisson term is only 1.8×10^{-5} and the count-space NB approximation rises only to 5.2889. Rather than invalidating the baseline interpretation, this gap delineates the domain over which the core specification is useful:

What the Model Captures Well (the bulk of central observations):

- Normal market conditions with continuous order flow
- Gradual order-flow shifts across descriptive market phases
- Routine volatility from imbalanced arrival rates
- Mean-reversion from latent occupancy unwinding

What Requires Extensions (stress-period tail events):

- Exchange hacks or policy shocks

 doi:10.20944/preprints202604.0256.v1

- Coordinated whale movements
- Cascading liquidations from leveraged positions
- Technological disruptions (forks, 51% attacks)

Subsection Appendix C.4.1 shows that a two-component mixture improves unconditional tail coverage materially relative to a rolling Gaussian benchmark, but still falls short of exact 99% unconditional coverage and continues to show violation clustering during stress episodes. In that specification, the queueing-motivated core benchmark describes normal times while a heavy-tailed Student-t component captures crisis periods. The core model provides the baseline; practitioners must still add a stress-state layer for tail-risk work.

This parallels established practice in equity markets, where models capture normal volatility but require separate treatment of crashes (Bates, 2000). The paper's narrower contribution is to characterize *normal* order-flow dynamics before turning to extreme events.

Appendix C.4.1. Heavy-Tail Regime Switch

Pure Skellam dynamics capture *normal* order-flow fluctuations but miss the extreme tails seen in realised returns. We therefore model the hourly return r_t as a two-component mixture described in Appendix E. Here we use a 1-hour VaR overlay based on a core-Normal / tail-Student-t mixture with threshold at $2\sigma_{\text{emp}}$, estimated by a threshold split rather than by full EM-MLE.

$$f_r(x) = \pi_0 \underbrace{\mathcal{N}(x;\mu_0,\sigma_0^2)}_{\text{core}} + \pi_1 \underbrace{t_\nu(x;\mu_1,\sigma_1^2)}_{\text{stress}}, \quad \pi_0+\pi_1=1, \tag{A16}$$

where the core component is fit on the in-threshold observations and is only queueing-motivated rather than queueing-imposed, while the stress component adds a near-zero-mean Student-t shock with scale σ_1 and degrees of freedom ν.[4]

Parameter estimation

We fit (A16) descriptively to the 2020–2025 Binance hourly close-to-close returns. The full-sample parameter estimates are

$$\pi_0 = 0.952, \qquad \mu_0 = 1.10\times10^{-4}, \qquad \sigma_0 = 0.00436, \tag{A17}$$

$$\pi_1 = 0.048, \qquad \mu_1 = -4.44\times10^{-4}, \qquad \sigma_1 = 0.02314, \qquad \nu = 21.92. \tag{A18}$$

For risk management, however, the paper evaluates this specification in a rolling out-of-sample backtest rather than relying on the full-sample fit alone. The intent is pragmatic tail augmentation, not structural identification from the queueing model.

Risk metrics

The one-hour 99 % Value-at-Risk is the 1 % quantile of the mixture. In the rolling backtest used in the main text, the average one-hour forecast is

$$\text{VaR}_{0.99}^{(1\text{h})} = F_r^{-1}(0.01) \approx 0.0184 \quad (1.84\% \text{ on average}).$$

Back-test

[4] Because the empirically estimated Negative-Binomial shape parameters are small ($\hat{\alpha}_+ = 0.4922$, $\hat{\alpha}_- = 0.6423$), their contribution to excess kurtosis is small relative to the stress-state Student-t term. Hence the queueing variance is best read here as motivation and comparison for σ_0, not as a hard estimation constraint.

 doi:10.20944/preprints202604.0256.v1

Using trailing 2160-hour estimation windows and re-fitting every 24 hours, the mixture produces $k = 520$ VaR breaches over $n = 46214$ out-of-sample hourly forecasts (1.13%). The Kupiec unconditional-coverage statistic is

$$\mathrm{LR}_{\mathrm{UC}} = -2\Big[(n-k)\ln(1-p) + k\ln p - (n-k)\ln\Big(1-\tfrac{k}{n}\Big) - k\ln\tfrac{k}{n}\Big] \approx 7.03,$$

yielding a p-value of 0.008. Relative to the rolling Gaussian benchmark (breach rate 1.80%), the mixture improves unconditional coverage materially. It nevertheless still misses exact 99% unconditional coverage at conventional levels, whereas rolling historical simulation (1.07%) is slightly closer to nominal unconditional coverage in this sample. We retain the mixture because it is a compact parametric monitoring overlay tied to the same one-hour variance normalization as the core benchmark, not because it dominates nonparametric forecasting. Violation clustering remains visible in stress episodes even after the unconditional fit improves.

Interpretation

Only 4.8 % of the time does the market enter the stress state, but in that regime the scale of shocks ($\sigma_1 \approx 0.02314$) is roughly five times the core Skellam standard deviation ($\sigma_0 \approx 0.00436$), contributing substantially to the empirical excess kurtosis of 47.29 reported in the main paper's empirical summary table.

This appendix evidence sharpens the paper's risk interpretation by:

1. separating the descriptive full-sample fit from the rolling out-of-sample evaluation,
2. showing that the stress-state layer improves on a Gaussian benchmark for unconditional 99% coverage without fully solving it, and
3. quantifying how much additional tail thickness is needed beyond the queueing core.

Appendix C.5. Parameter Evolution and Stability

The main paper reports rolling estimates of the feedback proxy $\beta_{\mathrm{eff}}^{\widehat{\mathrm{proxy}}}$ and the effective mean-reversion rate $\hat{\theta}_{\mathrm{eff}}$ across 30-day windows. The informative empirical result is the variation in those directly monitored quantities: the full-sample proxy is positive, while rolling mean reversion weakens materially and sometimes turns negative. Any accompanying symmetry-based quantities remain secondary descriptive translations rather than additional sources of identification.

Appendix C.6. Rolling-Window-Length Sensitivity

Because the 30-day rolling share of windows with $\hat{\theta}_{\mathrm{eff}} > 0$ is a visible headline number in the main paper, Table A3 checks the same descriptive statistic under shorter and longer overlapping windows and under more conservative sampling schemes for the 30-day window. The point of that table is modest: it does not create new identification, but it shows that the point-sign summary is not unique to a single exact window length or a single exact overlap convention. Table A4 then adds the uncertainty-aware 30-day classification used in the revised main text. The baseline classification uses a one-standard-error buffer in the lag-1 autocorrelation, $\hat{\rho}_1 \in [-1/\sqrt{720},\, 1/\sqrt{720}]$, which corresponds to the near-zero region ± 0.0373 around zero in the autocorrelation scale. This is a heuristic monitoring rule rather than a formal confidence interval under heteroskedastic, serially dependent returns. As a more conservative companion heuristic, doubling the buffer to ± 0.0745 leaves 17.2% of overlapping windows clear positive, 79.9% ambiguous, and 2.9% non-positive. The main qualitative message is unchanged: ambiguity dominates once one moves away from the raw point-sign share.

 doi:10.20944/preprints202604.0256.v1

Table A3. Sensitivity of the rolling sign summary for $\hat{\theta}_{\text{eff}}$ to window length. The shares remain descriptive because the windows overlap heavily.

Rolling window	Sampling scheme	$\Pr(\hat{\theta}_{\text{eff}} > 0)$	$\Pr(\hat{\theta}_{\text{eff}} \leq 0)$
15-day	Hourly-updated overlap	62.7%	37.3%
30-day	Hourly-updated overlap	66.2%	33.8%
30-day	Daily-sampled overlap	66.3%	33.7%
30-day	Non-overlapping blocks	62.7%	37.3%
60-day	Hourly-updated overlap	70.6%	29.4%

Table A4. Heuristic uncertainty-aware classification for the 30-day rolling summary. "Clear positive" means the implied lag-1 autocorrelation is below $-1/\sqrt{720}$, "non-positive" means it is above $1/\sqrt{720}$, and the middle region is classified as ambiguous. The rule is a monitoring buffer, not a formal dependence-robust interval.

30-day scheme	Clear positive	Ambiguous	Non-positive
Hourly-updated overlap	39.2%	49.4%	11.5%
Daily-sampled overlap	39.2%	49.4%	11.4%
Non-overlapping blocks	38.8%	49.3%	11.9%

Appendix C.7. Post-Shock Recovery by Rolling Resiliency Regime

The main paper interprets θ_{eff} only as a local descriptive recovery-timescale diagnostic. To connect that language to realized paths, we run one narrow event study using the existing hourly returns and the existing rolling 30-day regime labels. We identify isolated large shocks as the top 97.5% of absolute one-hour returns among hours with valid 30-day classifications, retain only local maxima within ± 6 hours, and de-overlap candidate shocks within the following 24 hours. This leaves 497 events. For an event at hour t with return r_t and sign $s_t = \text{sign}(r_t)$, define the aligned remaining displacement after h hours as

$$D_{t,h} := \frac{s_t \sum_{j=0}^{h} r_{t+j}}{|r_t|}.$$

By construction, $D_{t,0} = 1$. Lower values indicate more reversal of the initial shock, and $D_{t,h} = 0$ corresponds to a full reversal by horizon h.

Table A5 and Figure A2 show a clear ordering for these rare isolated shocks. Mean initial shock magnitudes are similar across regimes (about 3% of price), but the subsequent paths differ. After 6 hours, the mean aligned remaining displacement is 0.907 in clear-positive windows, versus 0.946 in non-positive windows and 1.025 in ambiguous windows. After 24 hours, the same ordering remains: 0.795 in clear-positive windows, 0.890 in non-positive windows, and 1.017 in ambiguous windows. The corresponding mean 24-hour reversal shares are 20.5%, 11.0%, and -1.7%. This is still only a threshold-specific descriptive check, but it gives a direct path-based complement to the next-day sorting exercise in the main text: the rolling resiliency classification is associated not only with next-day risk levels, but also with visibly different recovery paths after large isolated shocks.

Table A5. Post-shock recovery by rolling 30-day resiliency regime for isolated large hourly shocks on Binance BTC/USDT, 2020–2025. Shocks are the top 97.5% of absolute one-hour returns among hours with valid rolling classifications, restricted to local maxima within ± 6 hours and de-overlapped within the following 24 hours. Lower remaining displacement indicates faster reversal of the initial shock.

Rolling 30-day regime	Events	Mean initial $\lvert r_0 \rvert$	Remaining displacement at 6h	Remaining displacement at 24h	24h reversal share
Clear positive	161	3.11%	0.907	0.795	20.5%
Ambiguous	256	3.01%	1.025	1.017	-1.7%
Non-positive	80	3.07%	0.946	0.890	11.0%

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 April 2026 doi:10.20944/preprints202604.0256.v1

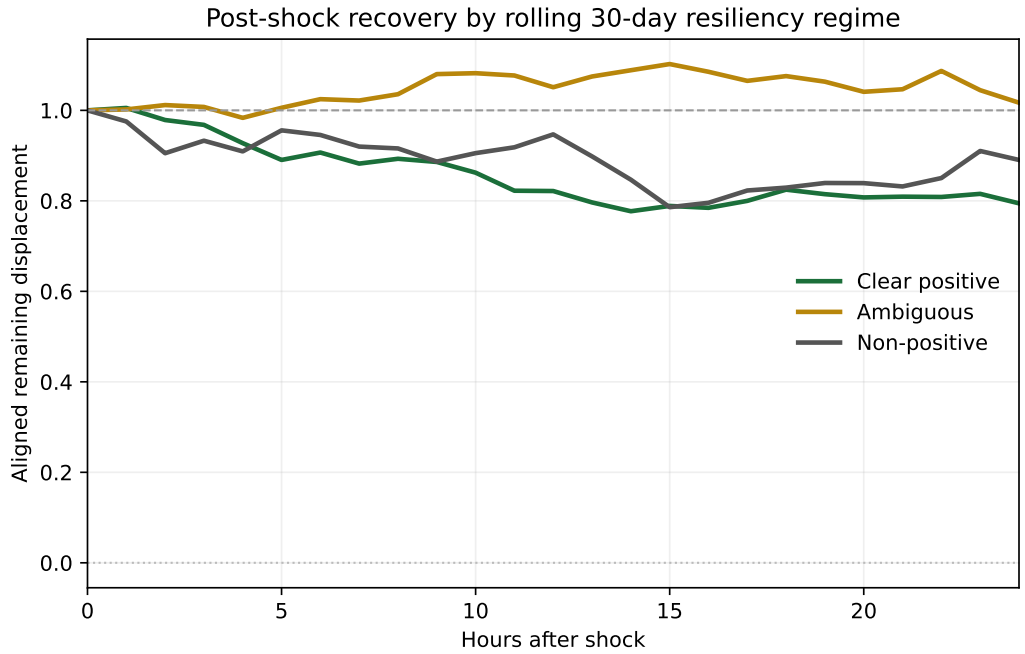

Figure A2. Aligned post-shock recovery paths by rolling 30-day resiliency regime for isolated large hourly shocks on Binance BTC/USDT, 2020–2025. Event hours are the top 97.5% of absolute hourly returns among observations with valid rolling classifications, restricted to local maxima within ± 6 hours and de-overlapped within the next 24 hours. The vertical axis reports aligned remaining displacement, normalized so the shock hour starts at 1; lower paths indicate faster reversal, and the dotted zero line corresponds to full reversal of the initial shock. Clear-positive windows recover more over the next 24 hours than ambiguous or non-positive windows in this narrow descriptive check; event counts by regime appear in Table A5.

Appendix C.8. Nearby-Frequency Robustness of the Feedback Splice

The headline proxy $\beta_{\text{eff}}^{\text{proxy}}$ combines a directional-impact scale from signed-flow regressions with a lagged signed-BTC-flow slope. Because those ingredients are estimated at different frequencies, we report two nearby alternatives here. The first replaces the baseline 1-minute directional scale with a 5-minute directional scale computed by the same day-by-day procedure on 5-minute buckets. The second replaces the baseline 1-hour flow regression with a non-overlapping 2-hour block regression and rescales the resulting cumulative signed-flow slope back to a per-hour feedback coefficient.

Table A6. Nearby-frequency robustness for the proxy feedback coefficient. All three full-sample variants remain positive and stay within the same order of magnitude as the baseline estimate.

Specification	Directional scale	Flow slope input	$\hat{\beta}_{\text{eff}}^{\text{proxy}}$ (h^{-1})
Baseline splice	1-minute $\hat{\kappa}_{\text{dir}} = 1.9971 \times 10^{-5}$	1-hour $\hat{b}_F = 7250.26$	0.0724
5-minute directional alternative	5-minute $\hat{\kappa}_{\text{dir}} = 1.8246 \times 10^{-5}$	1-hour $\hat{b}_F = 7250.26$	0.0661
2-hour flow alternative	1-minute $\hat{\kappa}_{\text{dir}} = 1.9971 \times 10^{-5}$	2-hour block slope $\hat{b}_F = 8358.10$	0.0417

The 5-minute directional alternative is 0.914 times the baseline value, while the 2-hour flow alternative is 0.576 times the baseline. These checks do not transform the splice into a directly identified structural estimate, but they do show that the positive sign and rough magnitude of the proxy are not tied to a single exact pair of sampling choices.

Appendix C.9. Secondary ρ Summary

The ratio

$$\rho \equiv \frac{4\,\tilde{\kappa}\,\beta}{\theta} = \frac{\beta_{\text{eff}}}{\theta_{\text{c}}}, \qquad \theta_{\text{c}} := \theta/4,$$

 doi:10.20944/preprints202604.0256.v1

remains available as a compact symmetry-based summary. In the full sample, the contingent structural translation gives $\rho = 7.52 \times 10^{-1}$ and $\delta = 1 - \rho = 0.247721$. We keep those quantities in the appendix rather than the main text because they do not add independent identifying information beyond the directly monitored pair $(\beta_{\text{eff}}^{\text{proxy}}, \theta_{\text{eff}})$. Near the regime boundary, small movements in $\hat{\theta}_{\text{eff}}$ can make ρ move sharply, which is why the main paper treats it as secondary.

Appendix C.10. Risk Management Performance

We backtest rolling one-hour 99% VaR forecasts rather than a fixed in-sample threshold. The pure queueing/Gaussian benchmark implies an average VaR of 1.51% and yields a breach rate of 1.80%, materially above the 1% target. The stress-state mixture raises the average forecast to 1.84% and lowers the breach rate to 1.13%, while rolling historical simulation gives 1.87% and 1.07%. The main text therefore uses the mixture only as a pragmatic parametric tail overlay: it improves materially on the Gaussian benchmark, but its Kupiec p-value of 0.008 still rejects exact unconditional 99% coverage. Historical simulation remains slightly closer to nominal unconditional coverage in this sample.

Appendix C.11. Robustness Tests

We keep the appendix robustness notes short because the main paper already reports the core reduced-form evidence. Two points matter most. First, the rolling estimates reported in the main paper show local variation rather than a single stable resiliency regime: a non-trivial minority of windows have $\theta_{\text{eff}} \leq 0$ and are better interpreted as local momentum or stress windows than as mean-reverting calibrations. Second, the auxiliary robustness material here is meant only to show that reasonable specification changes leave the paper's descriptive hierarchy intact: the main evidence remains the observable liquidity, signed-flow, and resiliency diagnostics reported in the main text.

Exchange-specific data requirement: The present implementation relies on Binance trade-print files together with aggressor-side information recoverable from the Binance schema (including `isBuyerMaker`). Extending the workflow to other venues would require rebuilding venue-specific trade-print and aggressor-side panels, so we leave multi-venue replication to future work rather than asserting cross-venue equivalence here.

Appendix C.12. Limitations and Future Data Work

A dedicated cross-venue robustness study would require harmonised fill-level data and aggressor-side reconstruction across heterogeneous APIs. We therefore leave that exercise to future work.

Appendix D. Coverage and Reconciliation of Public OHLCV Feeds

Public OHLCV aggregates (e.g. Binance "klines" and widely mirrored third-party files) do not provide the trade-print and aggressor-side fields required for the present calibration. The narrower point is that the present manuscript calibrates on full trade-print data rather than on public OHLCV aggregates, because the queueing, signed-flow, and aggressor-side diagnostics require fields that OHLCV files do not retain.

Appendix E. Estimation Details for the Core / Stress Mixture

We implement a *threshold split* rather than full EM-MLE. The threshold is $\tau = 2\sigma_{\text{emp}}$, and the mixture weights are $\hat{\pi}_0 = |\mathcal{S}_0|/T$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. The estimation procedure is:

1. Compute the empirical hourly return standard deviation σ_{emp}; set the stress threshold $\tau = 2\sigma_{\text{emp}}$.
2. **Core sample** $\mathcal{S}_0 = \{r_t : |r_t| \leq \tau\} \rightarrow$ fit $N(\mu_0, \sigma_0^2)$ via maximum-likelihood estimation.
3. **Tail sample** $\mathcal{S}_1 = \{r_t : |r_t| > \tau\} \rightarrow$ fit Student-$t(\nu_1, \mu_1, \sigma_1^2)$, same optimisation call.
4. Mixture weights $\hat{\pi}_0 = |\mathcal{S}_0|/T$, $\hat{\pi}_1 = 1 - \hat{\pi}_0$.
5. 99 % one-period VaR is $\text{VaR}_{0.99} = -F_{\text{mix}}^{-1}(0.01)$.
6. Back-test: 520 breaches out of 46214 one-hour forecasts $\Rightarrow$ unconditional coverage 1.13 %.

 doi:10.20944/preprints202604.0256.v1

The identical numerical values are reported in Appendix C.4.1 ($\hat{\pi}_0 = 0.952$, $\hat{\mu}_1 = -4.44 \times 10^{-4}$, $\hat{\sigma}_1 = 0.02314$, $\hat{\nu}_1 = 21.92$).

Appendix F. Monte-Carlo Validation of the Variance-Scale Estimator

We simulate 1000 independent $M/G/\infty$ paths with illustrative parameters $\lambda_{\pm} = 820\,\mathrm{h}^{-1}$, $\bar{H} = 2\,\mathrm{h}$ and $(\sqrt{R_r})_{\mathrm{true}} = 1.2 \times 10^{-4}$ (chosen for computational tractability; the bias test is scale-invariant). In each path we estimate the variance-scale coefficient from the same sample ratio $\sqrt{\mathrm{Var}(r)/\,\mathrm{E}[\nu]}$ used in the main paper's empirical-implementation subsection. Table A7 summarizes bias and dispersion.

Table A7. Finite-sample performance of the variance-scale estimator $\widehat{\sqrt{R_r}}$ (1000 replications).

T (h)	Bias	SD	BiasSE	RMSE
250	-2.66×10^{-7}	5.18×10^{-6}	1.64×10^{-7}	5.19×10^{-6}
500	-5.00×10^{-8}	3.67×10^{-6}	1.16×10^{-7}	3.67×10^{-6}
750	-8.74×10^{-8}	3.11×10^{-6}	9.83×10^{-8}	3.11×10^{-6}
1000	-2.79×10^{-8}	2.74×10^{-6}	8.68×10^{-8}	2.74×10^{-6}

Bias is statistically indistinguishable from zero and the RMSE follows the $\mathcal{O}(T^{-1/2})$ benchmark, confirming practical unbiasedness for sample windows of a few weeks or longer. See Figure A3.

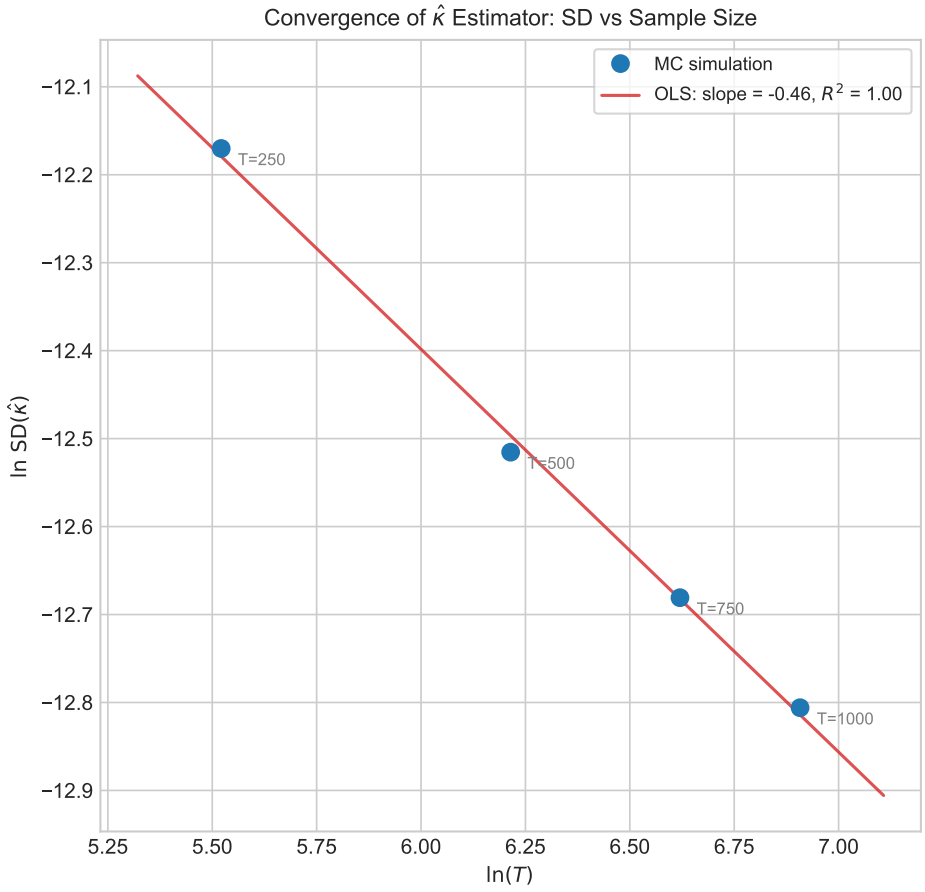

Figure A3. Finite-sample precision of the variance-scale estimator. Natural-log plot of sample length T (hours) versus the standard deviation of $\widehat{\sqrt{R_r}}$ obtained from 1 000 Monte-Carlo replications (blue dots). The red line is the OLS fit: $\ln\left(SD(\widehat{\sqrt{R_r}})\right) = -0.46\ln(T) - 9.65$, $r^2 = 1.00$, confirming the expected SD $\propto T^{-1/2}$ rate.

 doi:10.20944/preprints202604.0256.v1

Appendix G. Occupancy of a Discrete-Time GI/Geo/∞ Queue

Let $\Delta > 0$ be the time tick. In tick $n \in \mathbb{Z}_{\geq 0}$ the numbers of new long- and short-side latent occupancy units are independent

$$A_{n,+} \sim \text{Poisson}(\lambda_+ \Delta), \quad A_{n,-} \sim \text{Poisson}(\lambda_- \Delta).$$

An existing long-side occupancy unit survives to the next tick with probability $p_+ = 1 - \Delta/\mathbb{E}[H_+]$; a short-side occupancy unit survives with $p_- = 1 - \Delta/\mathbb{E}[H_-]$.

Let B_n and S_n denote the outstanding long- and short-side occupancies at the *start* of tick n and define $X_n = B_n - S_n$.

Proposition A1. *In steady state*

$$B_\infty \sim \textit{Poisson}(m_+), \quad S_\infty \sim \textit{Poisson}(m_-),$$
$$m_+ = \lambda_+ \mathbb{E}[H_+], \qquad m_- = \lambda_- \mathbb{E}[H_-].$$

Moreover, B_∞ and S_∞ are independent. Consequently, $X_\infty = B_\infty - S_\infty \sim \textit{Skellam}(m_+, m_-)$.

Proof. Consider the long-side queue; the short-side queue follows identically. Given B_n, the number of long-side occupancy units that survive to the next tick is Binomial(B_n, p_+). Thus

$$B_{n+1} = C_n + A_{n,+}, \quad \mathbb{E}[z^{C_n} \mid B_n] = \big(1 - p_+ + p_+ z\big)^{B_n}.$$

With $G_n(z) = \mathbb{E}[z^{B_n}]$, independence of $A_{n,+}$ gives the recursion $G_{n+1}(z) = \exp\big[\lambda_+ \Delta\,(z-1)\big]\, G_n\big(1 - p_+ + p_+ z\big)$. A stationary solution must satisfy $G(z) = \exp[\lambda_+ \Delta (z-1)] G(1 - p_+ + p_+ z)$. The unique pgf with $G(1) = 1$ is $G(z) = \exp[m_+(z-1)]$ provided $m_+ = \lambda_+ \Delta/(1-p_+) = \lambda_+ \mathbb{E}[H_+]$, establishing the Poisson law.

Independence of B_∞ and S_∞ follows from independent arrivals and services together with the PASTA property of Poisson processes Wolff (1982). The Skellam result is the difference of two independent Poisson variates. □

Appendix H. Piecewise Linear Depth and Price Moments

Partition the inventory space into three symmetric regions

$$\mathcal{K}_1 = \{k \in \mathbb{Z} : |k| < K_1\},$$
$$\mathcal{K}_2 = \{k : K_1 \leq |k| < K_2\},$$
$$\mathcal{K}_3 = \{k : |k| \geq K_2\},$$

and define the impact schedule

$$\kappa(k) = \begin{cases} \kappa_1, & k \in \mathcal{K}_1, \\ \kappa_2, & k \in \mathcal{K}_2, \\ \kappa_3, & k \in \mathcal{K}_3, \end{cases} \quad 0 < \kappa_1 < \kappa_2 < \kappa_3.$$

Let $X_\infty \sim \text{Skellam}(m_+, m_-)$ and define $P_\infty = P_0 + \kappa(X_\infty) X_\infty$.

 doi:10.20944/preprints202604.0256.v1

Proposition A2. *The first two unconditional moments of P_∞ are*

$$\mathbb{E}[P_\infty] = P_0 + \sum_{r=1}^{3} \kappa_r \mu_r, \qquad \mu_r = \sum_{k \in \mathcal{K}_r} k \Pr(X_\infty = k), \tag{A19}$$

$$\mathrm{Var}[P_\infty] = \sum_{r=1}^{3} \kappa_r^2 \sigma_r^2 - \Big(\mathbb{E}[P_\infty] - P_0\Big)^2, \qquad \sigma_r^2 = \sum_{k \in \mathcal{K}_r} k^2 \Pr(X_\infty = k). \tag{A20}$$

Each truncated sum involves a finite number of Skellam probabilities

$$\Pr(X_\infty = k) = e^{-(m_+ + m_-)} \left(\frac{m_+}{m_-}\right)^{k/2} I_{|k|}\big(2\sqrt{m_+ m_-}\big).$$

Proof. Write $Y = \kappa(X_\infty) X_\infty = \sum_{r=1}^{3} \kappa_r X_\infty \mathbf{1}_{\{X_\infty \in \mathcal{K}_r\}}$. Linearity of expectation yields (A19). For the second moment, $\mathbb{E}[Y^2] = \sum_r \kappa_r^2 \mathbb{E}\big[X_\infty^2 \mathbf{1}_{\mathcal{K}_r}\big]$ because the regions are disjoint. Subtracting the square of the mean gives (A20). The explicit Skellam form follows by inserting the pmf. □

Appendix H.1. Derivation of the Count-Space Burstiness Approximation

The main text's count-space burstiness formula is an approximation for the excess kurtosis of the hourly increment $\Delta X_t = N_t^+ - N_t^-$. This subsection first records the exact common-rate Gamma–Poisson cumulant ratio and then shows the simpler proxy reported in the main text.

Let

$$N_t^\pm \mid \Lambda_\pm \sim \mathrm{Poisson}(\Lambda_\pm t), \qquad \Lambda_\pm \sim \mathrm{Gamma}(\alpha_\pm, \vartheta),$$

where the buy and sell counts may have different shape parameters $\alpha_\pm$ but share a common Gamma *rate* ϑ, matching the main text's Negative-Binomial parameterization $p_\pm = \vartheta_\pm/(\vartheta_\pm + t)$. Write $\phi := t/\vartheta$, so $\bar{n}_\pm = \mathbb{E}[N_t^\pm] = \alpha_\pm \phi$. Under this single shape–rate parameterization, the cumulant generating function of $N_t^\pm$ is

$$K_\pm(u) = -\alpha_\pm \log\big(1 - \phi(e^u - 1)\big).$$

For the difference $\Delta X_t = N_t^+ - N_t^-$,

$$K_{\Delta X}(u) = K_+(u) + K_-(-u).$$

Because odd cumulants change sign under $u \mapsto -u$ while even cumulants do not, the second and fourth cumulants of ΔX_t are just the sums of the side-specific even cumulants:

$$\kappa_2(\Delta X_t) = (\alpha_+ + \alpha_-)\,\phi(1+\phi), \tag{A21}$$

$$\kappa_4(\Delta X_t) = (\alpha_+ + \alpha_-)\,\phi(1+\phi)\big(1 + 6\phi + 6\phi^2\big). \tag{A22}$$

Writing $\bar{n}_\Sigma := \bar{n}_+ + \bar{n}_- = (\alpha_+ + \alpha_-)\phi$, the exact common-rate excess kurtosis becomes

$$\mathrm{ExKurt}(\Delta X_t) = \frac{\kappa_4(\Delta X_t)}{\kappa_2(\Delta X_t)^2} = \frac{1 + 6\phi + 6\phi^2}{(\alpha_+ + \alpha_-)\phi(1+\phi)} \tag{A23}$$

$$= \frac{1}{\bar{n}_\Sigma(1+\phi)} + \frac{6}{\alpha_+ + \alpha_-}. \tag{A24}$$

The main text uses the simpler proxy

$$\mathrm{ExKurt}(\Delta X_t) \approx \frac{1}{\bar{n}_\Sigma} + \frac{6}{\alpha_+ + \alpha_-},$$

which preserves the familiar Poisson benchmark term $1/\bar{n}_\Sigma$ plus the over-dispersion term $6/(\alpha_+ + \alpha_-)$. The approximation is therefore not an exact common-rate identity; it replaces the smaller exact remainder $1/[\bar{n}_\Sigma(1+\phi)]$ by the Poisson benchmark $1/\bar{n}_\Sigma$ to keep the decomposition transparent. In

 doi:10.20944/preprints202604.0256.v1

the paper we use this only as a burstiness proxy for hourly buy/sell trade counts. It is not presented as a fully general return-law derivation.

Appendix I. NB–Skellam Distribution

When long- and short-side occupancy counts follow negative binomial distributions (arising from Gamma-mixed Poisson processes), their difference follows a NB–Skellam distribution.

Note on empirical relevance. With the empirical dispersion parameters $\alpha_+ = 0.4922$ and $\alpha_- = 0.6423$ (Gamma–Poisson mixing), the NB–Skellam *does not* collapse to the standard Skellam. The arrival–rate coefficients of variation are $1/\sqrt{\alpha_\pm} \approx 1.43$ and 1.25, and the corresponding burstiness proxy contributes about 5.3 to the excess-kurtosis decomposition under the approximation above, both materially different from the Poisson–Skellam benchmark. In our data, however, NB–Skellam still understates the extreme tails in hourly returns, which is why this count-space extension is retained only as a supplementary burstiness note alongside the separate stress overlay.

Data Availability Statement

See the main paper's Data Availability Statement for the full description of the three access layers: public raw Binance files, the broader public general code archive, and the 'V64' code-only workflow snapshot used for the empirical outputs in this manuscript. The 'V64' workflow snapshot is publicly archived at 10.5281/zenodo.19335853. That archive contains the workflow scripts and documentation needed to reproduce the computational pipeline for this manuscript, but it does not include the manuscript files, compiled PDFs, generated outputs, or raw Binance trade files.

References

Alexander, C., Deng, J., Feng, J., & Wan, H. (2023). Net buying pressure and the information in bitcoin option trades. *Journal of Financial Markets*, *63*, 100764. https://doi.org/10.1016/j.finmar.2022.100764.

Amihud, Y., & Mendelson, H. (1980). Dealership Market: Market-Making with Inventory. *Journal of Financial Economics*, *8*(1), 31–53. https://doi.org/10.1016/0304-405X(80)90020-3.

Anastasopoulos, A., Gradojevic, N., Liu, F., Maynard, A., & Tsiakas, I. (2026). Order flow and cryptocurrency returns. *Journal of Financial Markets*, 101047. (Advance online publication) https://doi.org/10.1016/j.finmar.2026.101047.

Bates, D.S., 2000. Post-'87 crash fears in the s&p 500 futures option market. Journal of Econometrics 94, 181–238. https://doi.org/10.1016/S0304-4076(99)00021-4.

Binance, 2025. Historical spot market data: Btcusdt daily trades. https://data.binance.vision/?prefix=data/spot/daily/trades/BTCUSDT/. Accessed: July 28, 2025.

Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). How Markets Slowly Digest Changes in Supply and Demand. In T. Hens & K. R. Schenk-Hoppe (Eds.), *Handbook of financial markets: Dynamics and evolution* (pp. 57–160). Elsevier. https://doi.org/10.1016/B978-012374258-2.50006-3.

Cont, R., & de Larrard, A. (2013). Price Dynamics in a Markovian Limit Order Market. *SIAM Journal on Financial Mathematics*, *4*(1), 1–25. https://doi.org/10.1137/110856605.

Dimpfl, T. (2017). Bitcoin Market Microstructure. *SSRN Electronic Journal*. (Available at SSRN 2949807) https://doi.org/10.2139/ssrn.2949807.

Foucault, T., Pagano, M., & Röell, A. (2013). *Market liquidity: Theory, evidence, and policy*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199936243.001.0001.

Garriott, C., van Kervel, V., & Zoican, M. (2025). Queuing and inventories in limit order markets. *Journal of Financial Markets*, *75*, 100982. https://doi.org/10.1016/j.finmar.2025.100982.

Ho, T., & Stoll, H. R. (1981). Optimal Dealer Pricing Under Transactions and Return Uncertainty. *Journal of Financial Economics*, *9*(1), 47–73. https://doi.org/10.1016/0304-405X(81)90020-9.

Kim, S. T., & Stoll, H. R. (2014). Are trading imbalances indicative of private information? *Journal of Financial Markets*, *20*, 151–174. https://doi.org/10.1016/j.finmar.2014.03.003.

Koopman, S. J., Lit, R., & Lucas, A. (2017). Intraday Stochastic Volatility in Discrete Price Changes: the Dynamic Skellam Model. *Journal of the American Statistical Association*, *112*(520), 1490–1503. https://doi.org/10.1080/01621459.2017.1302878.

 doi:10.20944/preprints202604.0256.v1

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica*, *53*(6), 1315–1336. https://doi.org/10.2307/1913210.

Palm, C. (1943). Intensitätsschwankungen im Fernsprechverkehr. *Ericsson Technics*(44), 1–189.

Stoll, H. R. (1978). The Supply of Dealer Services in Securities Markets. *Journal of Finance*, *33*(4), 1133–1151. https://doi.org/10.1111/j.1540-6261.1978.tb02053.x.

Wolff, R.W., 1982. Poisson arrivals see time averages. Operations Research 30, 223–231. https://doi.org/10.1287/opre.30.2.223.