

# When Does Volatility Model Selection Matter?

Entropy Diagnostics and Pre-Registered Evidence  
Across 1,496 Assets and Eleven Asset Classes

Olivier Saidi

research.olivier@proton.me    ORCID: 0009-0004-3221-6911

March 2026

DOI: 10.5281/zenodo.18894726

Code: [https://github.com/oliviersaidi/PACF\\_F](https://github.com/oliviersaidi/PACF_F)

## Abstract

We study when volatility model selection has value, rather than which model wins on average. A single entropy statistic, computable in seconds from a trailing return window, identifies 94% of assets where a simple EWMA baseline suffices<sup>1</sup>—eliminating 86% of model-fitting cost—and flags the 6% where in-sample diagnostics indicate selection may contribute to risk calibration. A  $2 \times 2$  in-sample attribution (Supplement S21) separates two sources of VaR improvement: FHS fixes unconditional calibration for *any* model (EWMA+FHS: 100% Kupiec pass), while model *diversity* reduces violation clustering (Christoffersen: 79% vs. 62% for EWMA+FHS; 77% joint coverage in-sample). Walk-forward backtesting (Supplement S26) shows that per-window selection overfits out-of-sample, but forecast *combination* preserves the Christoffersen benefit (58–60% vs. 51% for EWMA+FHS,  $p < 10^{-11}$ ). Model selection does *not* improve out-of-sample volatility forecasting or generate position-sizing alpha, bounding its value to computational triage and regime diagnostics.

Building on Rice’s (1976) algorithm selection framework, we formalize this diagnostic using Shannon entropy (with small-sample correction) and normalized permutation entropy, pre-register twelve hypotheses with cryptographic spec-locking, and evaluate twelve volatility forecasters on a 1,496-asset, 11-class cross-asset universe. Nine hypotheses pass after multiple-testing corrections; three null results (including the sizing guardrail) are reported with equal rigor. The entropy–dispersion

---

<sup>1</sup>*Economic* suffices: for these assets, the QLIKE improvement from selection is  $< 20\%$  (PUE threshold), and the median gain over forecast combination is  $< 0.1\%$ . EWMA is excluded from the MCS for 79.9% of assets, so “suffices” refers to economic materiality, not statistical indistinguishability.

association ( $\rho_s = -0.33$ ) is robust to controlling for asset class, liquidity, and volatility level ( $\rho_{\text{partial}} = -0.32$ ), confirming that the diagnostic operates within asset classes, not just between them. Selection value concentrates in low-entropy regimes where markets exhibit exploitable structure—connecting to the adaptive markets hypothesis—and strengthens at longer forecast horizons ( $h = 5, 20$ ). A variance-timing analysis reveals the mechanism: selected models allocate 22% more conditional variance to crash-adjacent days than EWMA while using only 8% more on calm days. Subgroup analysis across all eleven asset classes reveals interpretable heterogeneity: the entropy–VIX relationship is concentrated in macro-sensitive classes, while pattern-based diagnostics are strongest in equity and crypto markets.

**Keywords:** volatility forecasting, model selection, algorithm selection, meta-learning, Shannon entropy, permutation entropy, information theory, GARCH, EGARCH, APARCH, Model Confidence Set, QLIKE, value-at-risk, filtered historical simulation, Kupiec test, Christoffersen test, conditional coverage, tail risk, Basel III, risk management, Granger causality, cross-asset, multi-asset, quantitative finance, financial econometrics, pre-registration, reproducible research

**JEL Classification:** C32, C52, C53, C58, G17, G32

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Related Literature</b>	<b>6</b>
2.1	Volatility Model Comparison . . . . .	6
2.2	Entropy and Information Theory in Finance . . . . .	6
2.3	Meta-Learning and Algorithm Selection . . . . .	7
2.4	Reproducibility and Pre-Registration in Finance . . . . .	7
2.5	Positioning Relative to Prior Work . . . . .	8
<b>3</b>	<b>Conceptual Framework: Model-Selection Value</b>	<b>8</b>
3.1	Setup and Notation . . . . .	8
3.2	When Does Selection Have Value? . . . . .	9
3.3	Entropy Dynamics as a Leading Indicator . . . . .	10
3.4	Integrity as a Formal Property . . . . .	11
<b>4</b>	<b>Integrity Layer</b>	<b>11</b>
4.1	Architecture . . . . .	11
4.2	Chain of Custody . . . . .	12
4.3	Complete-Reporting Enforcement . . . . .	12
4.4	Anti-Selection-Bias Protocol . . . . .	12
<b>5</b>	<b>Data Universe</b>	<b>13</b>
5.1	Asset Coverage . . . . .	13
5.2	Realized Variance Proxy . . . . .	14
5.3	Time Alignment and Forward-Fill Protocol . . . . .	15
<b>6</b>	<b>Methodology</b>	<b>15</b>
6.1	Entropy Measures . . . . .	15
6.1.1	Shannon Entropy of Returns . . . . .	15
6.1.2	Entropy of Standardized Residuals . . . . .	17
6.1.3	Normalized Permutation Entropy . . . . .	17
6.2	Notation and Acronyms . . . . .	17
6.3	Volatility Models . . . . .	18
6.4	Scoring Rule: QLIKE . . . . .	19
6.5	Model Confidence Set . . . . .	19
6.6	Performance and Dispersion Metrics . . . . .	19
6.7	Risk Metrics . . . . .	20
6.8	Hypothesis Testing Protocol . . . . .	20
6.9	Additional Statistical Methods . . . . .	21

<b>7</b>	<b>Results</b>	<b>22</b>
7.1	Summary . . . . .	23
7.2	H1: Entropy–Dispersion Association . . . . .	23
7.3	H2: Entropy Dynamics Lead Volatility . . . . .	27
7.4	H3: Entropy and Macroeconomic Regimes . . . . .	29
7.5	P1–P2: Model-Selection Value . . . . .	30
7.6	A1–A2: Risk Calibration . . . . .	31
7.7	U1–U2: Uncertainty and Sizing . . . . .	32
7.8	R1–R2 and E1: Pattern Channels . . . . .	33
7.9	Solver Landscape: MCS Inclusion and Exclusion . . . . .	34
<b>8</b>	<b>Robustness and Sensitivity Analysis</b>	<b>36</b>
8.1	Entropy Discretization . . . . .	36
8.2	Rolling Window Length . . . . .	37
8.3	MCS Block Length . . . . .	38
8.4	MCS Significance Level . . . . .	38
8.5	Within-Class Analysis . . . . .	38
8.6	Volatility Proxy Robustness . . . . .	40
8.7	Innovation Distribution Sensitivity . . . . .	43
8.8	Convergence as a First-Class Outcome . . . . .	43
8.9	U2 Sizing Null by Entropy Regime . . . . .	45
8.10	Multi-Horizon Forecast Analysis . . . . .	45
8.11	Within-Asset Temporal Confirmation of H1 . . . . .	46
8.12	Out-of-Sample Forecast Quality: Model Selection vs. EWMA . . . . .	47
8.13	Selection vs. Forecast Combination . . . . .	49
<b>9</b>	<b>Practical Implications</b>	<b>51</b>
9.1	Decision Protocol . . . . .	51
9.2	Per-Hypothesis Practitioner Signals . . . . .	54
9.3	Cost–Benefit Quantification . . . . .	56
<b>10</b>	<b>Scope and Future Directions</b>	<b>57</b>
<b>11</b>	<b>Conclusion</b>	<b>61</b>
<b>A</b>	<b>Reproducibility Appendix</b>	<b>71</b>
A.1	Run Artifacts . . . . .	71
A.2	Software and Configuration . . . . .	72
A.3	Runtime Integrity Evidence . . . . .	73
A.4	Test Suite . . . . .	74

A.5 Table-to-Artifact Mapping . . . . .	74
<b>B Hypothesis Specification Details</b>	<b>74</b>
<b>C Metric Definitions and Runtime Invariants</b>	<b>75</b>

# 1 Introduction

Volatility forecasting underpins derivative pricing [Hull, 2017], mean-variance portfolio optimization [Markowitz, 1952], and regulatory capital computation [Basel Committee, 2010]. A voluminous literature compares conditional-variance models—GARCH [Bollerslev, 1986], EGARCH [Nelson, 1991], GJR-GARCH [Glosten et al., 1993], APARCH [Ding et al., 1993], FIGARCH [Baillie et al., 1996], HAR [Corsi, 2009], and the HEAVY model [Shephard and Sheppard, 2010]—typically concluding that the “best” model is asset- and period-dependent [Hansen and Lunde, 2005, Laurent et al., 2012] and that parsimonious specifications often dominate [Liu et al., 2015].

While this literature provides rigorous tools for ranking models *ex post*, the prior question of whether the ranking matters *ex ante*—under what observable conditions the choice among models has economic value—remains open. This is an instance of the *algorithm selection problem* [Rice, 1976]: given a set of candidate algorithms (volatility models) and an observable feature space (market-state descriptors), select the algorithm that maximizes a performance criterion or, equivalently, determine when selection effort has positive expected value.

We operationalize this meta-question by developing an information-theoretic diagnostic for model-selection value. The diagnostic rests on entropy-based market-state descriptors—Shannon entropy of return distributions [Shannon, 1948] and normalized permutation entropy of ordinal temporal patterns [Bandt and Pompe, 2002]—that are computationally inexpensive, model-free, and interpretable. High cross-solver forecast dispersion, conditional on these descriptors, signals the regime where model selection matters; low dispersion signals the regime where an exponentially weighted moving average (EWMA) baseline is sufficient.

The economic insight underlying this diagnostic is that entropy measures how far a market’s current return-generating process sits from the random-walk assumption embedded in simple models. When entropy is high, returns are close to i.i.d. and all models—from EWMA to asymmetric GARCH—converge on similar variance forecasts; the “horse race” is uninformative. When entropy is low, returns exhibit exploitable structure (volatility clustering, leverage effects, long memory) that sophisticated models capture and simple models miss, creating genuine forecast disagreement and economic gains from selection. This connects directly to the adaptive markets hypothesis [Lo, 2004]: markets oscillate between efficient states ( $\hat{H}$  near the uniform bound, model selection irrelevant) and structured states ( $\hat{H}$  depressed, selection economically valuable). Our benchmark provides the first large-scale, cross-asset measurement of this oscillation: fixed income ( $\bar{H} = 2.40$ ) and currency (2.54) inhabit structurally different entropy regimes than equity thematic (3.13) and equity sector (3.02), with model-selection gains concentrating systematically in the low-entropy regime. The practical consequence is immediate: a

risk manager can compute  $\hat{H}$  in seconds and know—before fitting a single GARCH model—whether the computational investment will pay off.

The contribution is fourfold.

**1. A formal model-selection value framework.** We define *Selection Value Index* (SVI), *Pattern Utilization Efficiency* (PUE), and *QLIKE dispersion* as quantitative measures of model-selection value, and derive propositions linking entropy to expected PUE under regularity conditions. This formalizes the intuition of the algorithm selection problem [Rice, 1976, Kotthoff, 2016] within volatility forecasting and connects to the meta-learning literature on “feature-based algorithm selection” [Smith-Miles, 2009, Bischl et al., 2016].

**2. Pre-registered, integrity-gated inference.** We implement a 12-hypothesis pre-registration protocol with cryptographic spec-locking, hash-chained computation records, analysis firewalls, and mandatory complete reporting. This computational analogue of clinical-trial pre-registration addresses the specification-flexibility component of the analyst degrees of freedom that Simmons et al. [2011] identify as a driver of false positives, and that Harvey et al. [2016], McLean and Pontiff [2016], and Hou et al. [2020] document in empirical finance. The integrity mechanism goes beyond disclosure: it is *machine-enforced* at runtime—the pipeline halts on any omitted hypothesis or spec-hash mismatch, preventing selective reporting or post-hoc modification without detectable violations. Unlike voluntary pre-registration platforms (OSF, AsPredicted), the mechanism is embedded in the computation itself; Section 4 details the full chain-of-custody architecture.

**3. A 1,496-asset, 11-class cross-asset benchmark.** The benchmark universe spans equities (396 single-name stocks, plus 100 index, 118 sector, 92 thematic, and 109 international ETFs), cryptocurrencies (115), fixed income (144), currencies (146), commodities (106), real estate (140), and volatility instruments (30)—an order of magnitude broader than typical volatility comparison studies [Hansen and Lunde, 2005, Laurent et al., 2012, Liu et al., 2015]. Eleven fine-grained asset classes with  $n \geq 10$  are tested individually, enabling subgroup analysis that separates composition-driven effects from genuine within-class signals.

**4. Calibrated claims with explicit null results.** Three of twelve pre-registered hypotheses fail after multiple-testing corrections. We report these null results with the same rigor as positive findings, following the emerging norm of credible empirical research [Brodeur et al., 2020, Andrews and Kasy, 2019]. The failure of uncertainty-aware sizing to generate walk-forward Sharpe improvement (Hypothesis U2) is an important disciplining result: it bounds the framework’s claims to risk management and prevents

overclaiming of return predictability from OHLCV data, consistent with the efficient-markets constraint [Fama, 1970, Lo, 2004]. Beyond the core hypotheses, we subject the framework to an extended robustness battery: within-asset panel estimation with asset fixed effects, class-level meta-analysis via Fisher combination, volatility-proxy sensitivity under both Parkinson and close-to-close realized variance, convergence modelling as a first-class outcome, multi-horizon forecast analysis at 5- and 20-day horizons, within-asset temporal confirmation of the entropy–dispersion link, out-of-sample walk-forward forecast evaluation with per-solver decomposition, entropy-regime conditioning of the sizing null, transaction-cost sensitivity sweeps, and economic calibration mapping statistical quantities to VaR breach reductions.

The remainder proceeds as follows. Section 2 reviews related literature. Section 3 develops the theoretical framework. Section 4 describes the integrity layer. Section 5 documents the data. Section 6 presents the methodology. Section 7 reports results. Section 8 presents robustness and sensitivity analyses, including proxy choice, convergence modelling, and within-class tests. Section 9 discusses implications. Section 10 discusses scope and future directions. Section 11 concludes.

## 2 Related Literature

This work intersects four research streams: volatility model comparison, information-theoretic analysis of financial markets, meta-learning and algorithm selection, and reproducibility in empirical finance.

### 2.1 Volatility Model Comparison

The question “does anything beat a GARCH(1,1)?” was posed definitively by Hansen and Lunde [2005], who introduced the Model Confidence Set (MCS) and found no single model consistently dominating on exchange rate data. Laurent et al. [2012] extended this analysis to multivariate GARCH and confirmed asset-dependence of rankings. Liu et al. [2015] surveyed 330 GARCH variants using tick-level realized measures and concluded that simpler specifications frequently dominate.

On the econometric side, Diebold and Mariano [1995] and West [1996] established the theoretical foundations for comparing forecasts from nested and non-nested models. Giacomini and White [2006] introduced conditional predictive ability tests that allow for estimation uncertainty, while White [2000] developed the reality check for data snooping, later sharpened by Hansen [2005] into the Superior Predictive Ability (SPA) test. Patton [2011] demonstrated that QLIKE is the unique member of the robust loss function family (homogeneous of degree zero) that correctly ranks variance forecasts even under noisy proxies—a result that motivates our exclusive reliance on QLIKE.



Our departure from this literature is to ask not “which model wins?” but “when does asking this question have value?” This reframing shifts the object of inference from model identity to model-selection *regimes*, diagnosed by observable state features.

## 2.2 Entropy and Information Theory in Finance

Shannon entropy applied to financial time series dates to Gulko [1999] and has been used to characterize market efficiency [Risso, 2008, Zunino et al., 2009, Ortiz-Cruz et al., 2012], predictability regimes [Pelé and Mazuure, 2019], and systemic risk [Maasoumi and Racine, 2015]. Permutation entropy [Bandt and Pompe, 2002] captures temporal ordinal structure and has been shown to discriminate developed from emerging markets [Zunino et al., 2009, 2010], characterize stock market complexity [Hou et al., 2017], and measure time-varying efficiency in commodity markets [Sensoy et al., 2015].

Transfer entropy [Schreiber, 2000]—the information-theoretic analogue of Granger causality—has been applied to measure directed information flow between markets [Dimpfl and Peter, 2013, He and Hamori, 2022]. Dimpfl and Peter [2014] specifically link transfer entropy to volatility transmission, providing a theoretical basis for our Hypothesis H2 (entropy dynamics leading volatility changes).

Our use of entropy differs from this literature in a crucial respect: we employ entropy not as a standalone descriptor of market state but as a *feature variable for a model-selection decision problem*. This connects to the meta-learning literature (Section 2.3) and represents a novel application of information theory to algorithmic decision-making in finance.

## 2.3 Meta-Learning and Algorithm Selection

The algorithm selection problem was formalized by Rice [1976]: given a problem instance described by a feature vector  $\mathbf{x}$  and a set of candidate algorithms  $\mathcal{A}$ , select the algorithm  $a^* \in \mathcal{A}$  maximizing expected performance. Smith-Miles [2009] and Kotthoff [2016] survey modern approaches using instance-space analysis and meta-learning. Bischl et al. [2016] provide a comprehensive benchmark of algorithm selection systems across combinatorial optimization.

In time-series forecasting, Talagala et al. [2018] develop feature-based forecast model selection using time-series features (entropy, trend, seasonality, autocorrelation) to predict which forecasting method performs best—an approach closely related to ours. The Pattern-Aware Complexity Framework (PACF) of Saidi [2025] generalizes this by formalizing how pattern prevalence and instance entropy modulate effective computational complexity, introducing metrics for pattern utilization efficiency (PUE) and uncertainty reduction.

We instantiate the PACF in the volatility forecasting domain, mapping its abstract constructs to financially interpretable quantities: entropy becomes a return-distribution

descriptor, pattern prevalence measures the fraction of time structural patterns are active, and PUE quantifies the percentage QLIKE improvement from optimized model selection over a fixed EWMA baseline.

## 2.4 Reproducibility and Pre-Registration in Finance

Harvey et al. [2016] estimate that the majority of published factor discoveries are likely false positives, proposing a  $t > 3.0$  threshold. McLean and Pontiff [2016] show that anomaly returns decay by 58% post-publication. Hou et al. [2020] fail to replicate 65% of anomalies. Chordia et al. [2020] raise the bar for cross-sectional predictability.

Pre-registration—standard in clinical trials [Nosek et al., 2018] and increasingly adopted in economics [Olken, 2015, Brodeur et al., 2020]—remains rare in quantitative finance. Andrews and Kasy [2019] advocate for pre-analysis plans in empirical economics. We implement a computational pre-registration mechanism: specifications are cryptographically hashed and locked before any data processing, each computation record carries chain-of-custody stamps, and the system enforces mandatory complete reporting of all 12 hypotheses. Unlike voluntary disclosure, this mechanism is machine-enforced and produces verifiable proof of adherence.

## 2.5 Positioning Relative to Prior Work

Table 1 highlights the principal dimensions on which this study departs from the closest antecedents. The differences are quantitative (two orders of magnitude in universe breadth) and methodological (pre-registration, entropy-based diagnostics, and complete-reporting enforcement).

**Table 1:** Comparison with representative volatility model comparison studies. “Pre-reg.” = computational pre-registration with hash-chain integrity. “MTC” = multiple testing correction applied.

Study	Assets	Models	Loss	Pre-reg.	MTC	
Hansen and Lunde [2005]	1	8	MSE/QLIKE	No	SPA	†Machine-enforced: the
Laurent et al. [2012]	10	10	QLIKE	No	MCS	
Liu et al. [2015]	6	330	QLIKE	No	MCS	
This paper	1,496	12	QLIKE	Yes (SHA-256)†	Holm + BH	

pipeline halts on any omitted hypothesis or spec-hash mismatch. Unlike voluntary platforms (OSF, AsPredicted), the pre-registration is embedded in the computation and cannot be bypassed without detectable hash violations.

## 3 Conceptual Framework: Model-Selection Value

We formalize the conditions under which volatility model selection has value and establish the role of entropy as a diagnostic feature. The propositions below are not claims of theoretical novelty over the meta-learning literature [Rice, 1976, Smith-Miles, 2009,

Bischl et al., 2016]; they are an organizing framework that translates informal intuitions into falsifiable predictions with explicit regularity conditions, enabling the pre-registered empirical tests of Section 7.

### 3.1 Setup and Notation

Let  $\mathcal{M} = \{m_1, \dots, m_M\}$  be a portfolio of  $M$  volatility models. For asset  $i$ , each model  $m$  produces a sequence of one-step-ahead conditional variance forecasts  $\hat{\sigma}_{m,t|t-1}^2$ . Performance is evaluated via the QLIKE loss:

$$L_{m,i} = \frac{1}{T_i} \sum_{t=1}^{T_i} \ell(\hat{\sigma}_{\text{RV},t}^2, \hat{\sigma}_{m,t|t-1}^2), \quad \ell(x, y) = \frac{x}{y} - \ln \frac{x}{y} - 1, \quad (1)$$

where  $\hat{\sigma}_{\text{RV},t}^2$  is a realized-variance proxy.

**Definition 1** (Model-Selection Gain). *The model-selection gain for asset  $i$  relative to a fixed baseline model  $m_0$  (EWMA) is:*

$$G_i = L_{m_0,i} - \min_{m \in \mathcal{M}} L_{m,i}. \quad (2)$$

*The normalized gain is the Pattern Utilization Efficiency:  $PUE_i = G_i / L_{m_0,i} \times 100\%$ .*

**Definition 2** (Cross-Model Dispersion). *The cross-model dispersion  $D_i$  (standard deviation of QLIKE losses across converged models) and spread  $S_i$  (range) for asset  $i$  are:*

$$D_i = sd(\{L_{m,i}\}_{m \in \mathcal{M}_i}), \quad S_i = \max_m L_{m,i} - \min_m L_{m,i}, \quad (3)$$

where  $\mathcal{M}_i \subseteq \mathcal{M}$  is the set of models that converge for asset  $i$  ( $|\mathcal{M}_i| \geq 3$ ).

**Definition 3** (Selection Value Index). *Let  $\widehat{\mathcal{M}}_\alpha^*$  be the Model Confidence Set of Hansen et al. [2011] at significance level  $\alpha$ . Then:*

$$SVI_{i,\alpha} = 1 - \frac{|\widehat{\mathcal{M}}_{i,\alpha}^*|}{|\mathcal{M}_i|}. \quad (4)$$

*$SVI = 1$  when a single model dominates statistically;  $SVI = 0$  when all models are indistinguishable.*

### 3.2 When Does Selection Have Value?

The practitioner's decision problem is: given observable features of asset  $i$ , is it worth investing in the full model-comparison procedure (estimating all  $M$  models, running MCS, selecting the best) or defaulting to the baseline  $m_0$ ?

**Assumption 1** (Feature Observability). *The practitioner observes a feature vector  $\mathbf{x}_i$  before model fitting, where  $\mathbf{x}_i$  includes entropy-based descriptors: Shannon entropy  $\hat{H}_i$ , normalized permutation entropy  $NPE_i$ , and pattern prevalence  $\rho_i$ .*

**Proposition 1** (Dispersion as a Necessary Condition). *Cross-model dispersion  $D_i$  (Definition 2) vanishes if and only if the model-selection gain  $G_i$  (Definition 1) vanishes. Equivalently, model selection has positive value ( $G_i > 0$ ) only if the converged models disagree ( $D_i > 0$ ).*

*Proof.* ( $\Rightarrow$ ) If  $D_i = 0$ , then  $L_{m,i} = c$  for all  $m \in \mathcal{M}_i$  and some constant  $c$ . Since  $m_0 \in \mathcal{M}_i$ , we have  $L_{m_0,i} = c = \min_m L_{m,i}$ , hence  $G_i = L_{m_0,i} - \min_m L_{m,i} = 0$ .

( $\Leftarrow$ ) Suppose  $G_i > 0$ . Then  $\min_m L_{m,i} < L_{m_0,i}$ . Because  $m_0 \in \mathcal{M}_i$ ,  $\max_m L_{m,i} \geq L_{m_0,i} > \min_m L_{m,i}$ , so  $S_i = \max_m L_{m,i} - \min_m L_{m,i} > 0$ . For any finite set,  $S_i > 0 \Leftrightarrow D_i > 0$ , completing the contrapositive.  $\square$

This tautological result has a non-trivial implication: any observable feature correlated with  $D_i$  is a diagnostic for model-selection value.

**Proposition 2** (Entropy–Dispersion Diagnostic). *Under the following regularity conditions:*

- (i) *Models in  $\mathcal{M}$  have heterogeneous sensitivity to return-distribution shape (some models are invariant to higher moments, others are not);*
- (ii) *Shannon entropy  $\hat{H}_i$  is a sufficient statistic for the distributional shape variation that drives inter-model loss differences;*

*then  $\text{Cov}(\hat{H}_i, D_i) \neq 0$ , and entropy is a diagnostic for model-selection value.*

**Remark 1.** *Condition (ii) is an idealization. In practice, entropy captures distributional shape imperfectly, and the sign of the association is theoretically ambiguous. Under the “complexity” interpretation—higher entropy implies a harder forecasting problem with more solver divergence— $\text{Cov}(\hat{H}_i, D_i) > 0$ . Under the “efficiency” interpretation [Zunino et al., 2009]—higher entropy implies an informationally efficient market where all models converge to random-walk-like forecasts— $\text{Cov}(\hat{H}_i, D_i) < 0$ . We test two-tailed precisely because the theory admits both signs. Condition (i) is empirically verifiable: the Student- $t$  innovation analysis (Supplement S12)<sup>2</sup> shows that best-solver agreement varies from 47% (volatility instruments) to 86% (equity index) across asset classes, confirming that models have heterogeneous sensitivity to distributional shape.*

**Proposition 3** (Selection Value Concentration). *If the cross-model loss distribution conditional on high dispersion ( $D_i > D_{\text{med}}$ ) is stochastically dominated by the distribution*

<sup>2</sup>Supplements S1–S28 are self-contained reproducibility scripts, each producing a JSON result file. All scripts and outputs are available in the code repository.

conditional on low dispersion, then:

$$\mathbb{E}[PUE_i \mid D_i > D_{med}] > \mathbb{E}[PUE_i \mid D_i \leq D_{med}]. \quad (5)$$

In words: model-selection value concentrates where models disagree most.

*Proof.* When  $D_i$  is large,  $S_i$  is large, and  $\min_m L_{m,i}$  is further below  $L_{m_0,i}$  in expectation (under stochastic dominance of the conditional loss distribution), yielding higher  $G_i$  and PUE.  $\square$

This proposition is the theoretical core of Hypothesis P1 (PUE–spread association). Its empirical test is whether the observed Spearman correlation between PUE and  $S_i$  is significantly positive. The prediction is falsifiable: high dispersion could arise from one catastrophically bad model rather than from a genuine best-versus-worst gap, and indeed 31% of high- $D$  assets have below-median PUE (Section 7).

### 3.3 Entropy Dynamics as a Leading Indicator

**Assumption 2** (Information Accumulation). *Changes in return entropy  $\Delta H_t$  reflect shifts in the information environment: rising entropy indicates transition toward a less predictable regime (new information arrival, regime uncertainty); falling entropy indicates reversion toward structured dynamics (volatility clustering, trend consolidation).*

Under Assumption 2, entropy changes should *precede* realized-volatility changes if regime transitions are gradual (entropy shifts before volatility fully adjusts). This motivates Hypothesis H2, tested via the Toda–Yamamoto modification of Granger causality [Toda and Yamamoto, 1995], which avoids unit-root pre-testing issues in possibly integrated systems. Optimal lag order is selected by AIC [Akaike, 1974] on the bivariate VAR.

### 3.4 Integrity as a Formal Property

We formalize the anti-HARKing mechanism as a verifiable property.

**Definition 4** (Pre-Registration Integrity). *A research pipeline satisfies pre-registration integrity if:*

- (i) A hypothesis specification  $\mathcal{H}$  is committed (hashed) before data access:  $h_{spec} = \text{SHA256}(\mathcal{H})$ ;
- (ii) Each result record  $R_j$  carries  $h_{spec}$  and a chain hash  $h_{chain,j}$  computed from input, config, and code hashes;
- (iii) Report generation verifies  $h_{spec}(R_j) = h_{spec}$  for all  $j$  and  $|\{R_j\}| = |\mathcal{H}|$  (completeness);
- (iv) Any modification to  $\mathcal{H}$  post-commitment produces  $h'_{spec} \neq h_{spec}$ , detected by the verification step.

**Proposition 4** (Selective Reporting Prevention). *Under Definition 4, it is computationally infeasible (under standard cryptographic assumptions) to produce a valid report that omits any pre-specified hypothesis or that uses a modified specification without a detectable hash mismatch.*

This property distinguishes our mechanism from voluntary pre-registration, which relies on researcher compliance rather than computational enforcement.

## 4 Integrity Layer

### 4.1 Architecture

The system is organized as a 14-module directed acyclic graph (DAG) with strict dependency ordering. Entropy, pattern, and solver modules are *siblings* in the DAG—they cannot import from each other. Cross-cutting integration occurs only in the hypothesis and benchmark layers. This architectural firewall prevents inadvertent information leakage between independently testable components.

The execution pipeline consists of 11 sequential steps: (1) spec-manifest lock, (2) data loading and validation, (3) solver estimation (parallelized, 12 workers), (4–10) hypothesis evaluation (each hypothesis function receives only its declared metric dependencies via a typed registry), (11) report generation with completeness verification.

### 4.2 Chain of Custody

Each computation record carries a five-field hash chain:

$$\begin{aligned} h_{\text{input}} &= \text{SHA256}(\text{data fingerprint}), \\ h_{\text{config}} &= \text{SHA256}(\text{serialized config}), \\ h_{\text{code}} &= \text{SHA256}(\text{version ID}), \\ h_{\text{chain}} &= \text{SHA256}(h_{\text{input}} || h_{\text{config}} || h_{\text{code}}), \\ h_{\text{output}} &= \text{SHA256}(\text{result fields}). \end{aligned} \tag{6}$$

Records are stamped at computation boundaries (30+ call sites). The output hash is computed *after* all result fields are finalized, ensuring that any post-hoc modification is detectable. Fields that are metadata (timestamp, ticker) are excluded from  $h_{\text{output}}$  to preserve hash determinism across runs with identical inputs and code.

Additional fields for time-alignment validation ( $n_{\text{obs,input}}$ ,  $n_{\text{nan,filled}}$ ) are *included* in  $h_{\text{output}}$  because they affect reproducibility: different alignment produces different hashes.

### 4.3 Complete-Reporting Enforcement

The report generator validates: (i) exactly 12 hypothesis IDs are present, (ii) all spec-manifest hashes match the locked manifest, and (iii) every hypothesis in the pre-registered set has a corresponding result record. Failure on any condition halts report generation with an explicit integrity-violation message.

**Referee verification recipe.** In brief: all pre-registered hypotheses were tested, all results are hash-chained to the locked specification, and no post-hoc modifications are possible without breaking the chain. To independently verify this, a referee may: (1) compute `sha256sum integrity/spec_manifest.json` and confirm the hash matches the value recorded in `generation_manifest.json`; (2) verify that `integrity/batch_manifest.json` lists all 12 hypothesis IDs with `tested: true` and `missing: []`; (3) confirm that each hypothesis record in `data/results_all.json` carries a `chain_hash` field whose inputs (data fingerprint, config hash, code version) are traceable to the locked spec; (4) verify that `integrity/reproducibility_check.json` reports `all_valid: true` with the declared `master_seed`. All artifacts are archived in the repository alongside the benchmark output; no external credentials are required.

### 4.4 Anti-Selection-Bias Protocol

When any metric relies on a “best model” selected in-sample, all downstream evaluations use a disjoint out-of-sample window. Specifically:

- 252-day estimation window for model fitting and selection;
- 63-day evaluation window for all performance, risk, and uncertainty metrics;
- Sliding walk-forward with no overlap between estimation and evaluation periods.

This protocol follows the walk-forward methodology of Tashman [2000] and eliminates the in-sample selection bias that White [2000] identified as the dominant source of data snooping in forecast comparisons.

## 5 Data Universe

### 5.1 Asset Coverage

The benchmark universe comprises 1,496 assets grouped into eleven fine-grained classes (Table 2). Daily OHLCV data are sourced via `yfinance` with a minimum observation requirement of  $n \geq 252$  trading days.

**Inclusion criteria.** An asset enters the universe if: (i) at least 252 daily OHLCV observations are available (ensures one full year for rolling-window estimation); (ii) the

daily close series has no gaps exceeding 10 consecutive trading days (forward-fill tolerates shorter gaps; see Section 5.3); (iii) the ticker is listed in the pre-registered asset registry, frozen before any model estimation. Assets are *excluded from cross-sectional hypothesis tests* (but retained in the universe) if fewer than three solvers converge ( $n_{\text{solvers}} < 3$ ), because meaningful QLIKE dispersion requires at least three competing forecasts. Four cryptocurrency tickers are excluded on this basis, yielding an effective cross-sectional sample of  $n = 1,492$ . The universe is fixed at run time; no asset is added or removed based on model output or hypothesis results.

**Table 2:** Asset universe composition by fine-grained class.

Asset Class	Count
Equity (single-name stocks)	396
Currency	146
Fixed income	144
Real estate (REITs)	140
Equity (sector ETFs)	118
Cryptocurrency	115
Equity (international ETFs)	109
Commodity	106
Equity (index ETFs)	100
Equity (thematic ETFs)	92
Volatility instruments	30
<b>Total</b>	<b>1,496</b>

**Effective sample sizes.** Hypothesis H2 (Granger causality) requires only rolling entropy and realized volatility data—not solver convergence—and therefore retains  $n = 1,494$ . All subsequent tables report the effective sample size for each test.

This universe is substantially broader than typical volatility comparison studies: Hansen and Lunde [2005] use 1 exchange rate, Laurent et al. [2012] use 10 exchange rates, and Liu et al. [2015] use 6 assets. The breadth is required for two reasons: (i) cross-asset heterogeneity tests whether entropy-based diagnostics generalize, and (ii) large  $N$  provides statistical power for cross-sectional correlation hypotheses. All eleven classes meet the minimum subgroup size threshold ( $n \geq 10$ ), enabling fine-grained within-class hypothesis testing (Section 8).

For narrative exposition, assets aggregate into four super-classes: *Equity* (single names, indices, sector, thematic, international, volatility;  $n_{\text{eq}} = 845$ ), *Crypto* ( $n_{\text{cr}} = 115$ ), *Alternative* (commodity, REITs;  $n_{\text{alt}} = 246$ ), and *Rates/FX* (fixed income, currency;  $n_{\text{rfx}} = 290$ ).

Table 3 reports key distributional features by asset class. Sample lengths range from 337 to 6,806 trading days (2001–2026). Shannon entropy varies substantially across classes:



fixed income shows the lowest mean  $\hat{H}$  (2.40), reflecting structured autocorrelation; equity thematic the highest (3.13), consistent with near-white-noise returns. QLIKE dispersion is right-skewed—medians are far below means—reflecting a minority of assets with extreme solver disagreement (Figure 1). The fraction of assets where  $\text{PUE} > 20\%$  ranges from 0% (equity index) to 18% (currency), foreshadowing the heterogeneity that motivates within-class analysis in Section 8 (Figure 2).

**Table 3:** Distributional summary by asset class.  $n$  = number of assets,  $T$  = sample length range (trading days),  $\bar{H}$  = mean Shannon entropy,  $\tilde{D}$  = median QLIKE dispersion,  $\overline{\text{SVI}}$  = mean Solver Variability Index,  $\%\text{PUE} > 20\%$  = fraction of assets where model selection reduces loss by more than 20%.

Asset Class	$n$	$T$ range	$\bar{H}$	$\tilde{D}$	$\overline{\text{SVI}}$	$\%\text{PUE} > 20\%$
commodity	106	1,099–6,573	2.985	0.031	0.35	7%
crypto	111	446–4,171	2.801	0.038	0.23	17%
currency	146	2,560–6,806	2.544	0.084	0.36	18%
equity_index	100	1,178–6,569	2.981	0.040	0.60	0%
equity_intl.	109	2,019–6,573	2.946	0.040	0.38	3%
equity_sector	118	1,025–6,573	3.021	0.036	0.45	3%
equity_stock	396	337–6,569	2.880	0.029	0.37	2%
equity_thematic	92	1,219–6,463	3.126	0.036	0.37	4%
fixed_income	144	919–6,573	2.404	0.137	0.29	7%
real_estate	140	1,253–6,573	2.539	0.034	0.36	1%
volatility	30	638–6,573	2.739	0.136	0.34	17%
<b>Full sample</b>	<b>1,492</b>	<b>337–6,806</b>	<b>2.806</b>	<b>0.038</b>	<b>0.37</b>	<b>6%</b>

## 5.2 Realized Variance Proxy

Following Parkinson [1980], we compute:

$$\hat{\sigma}_{\text{Park},t}^2 = \frac{(\ln H_t - \ln L_t)^2}{4 \ln 2}, \quad (7)$$

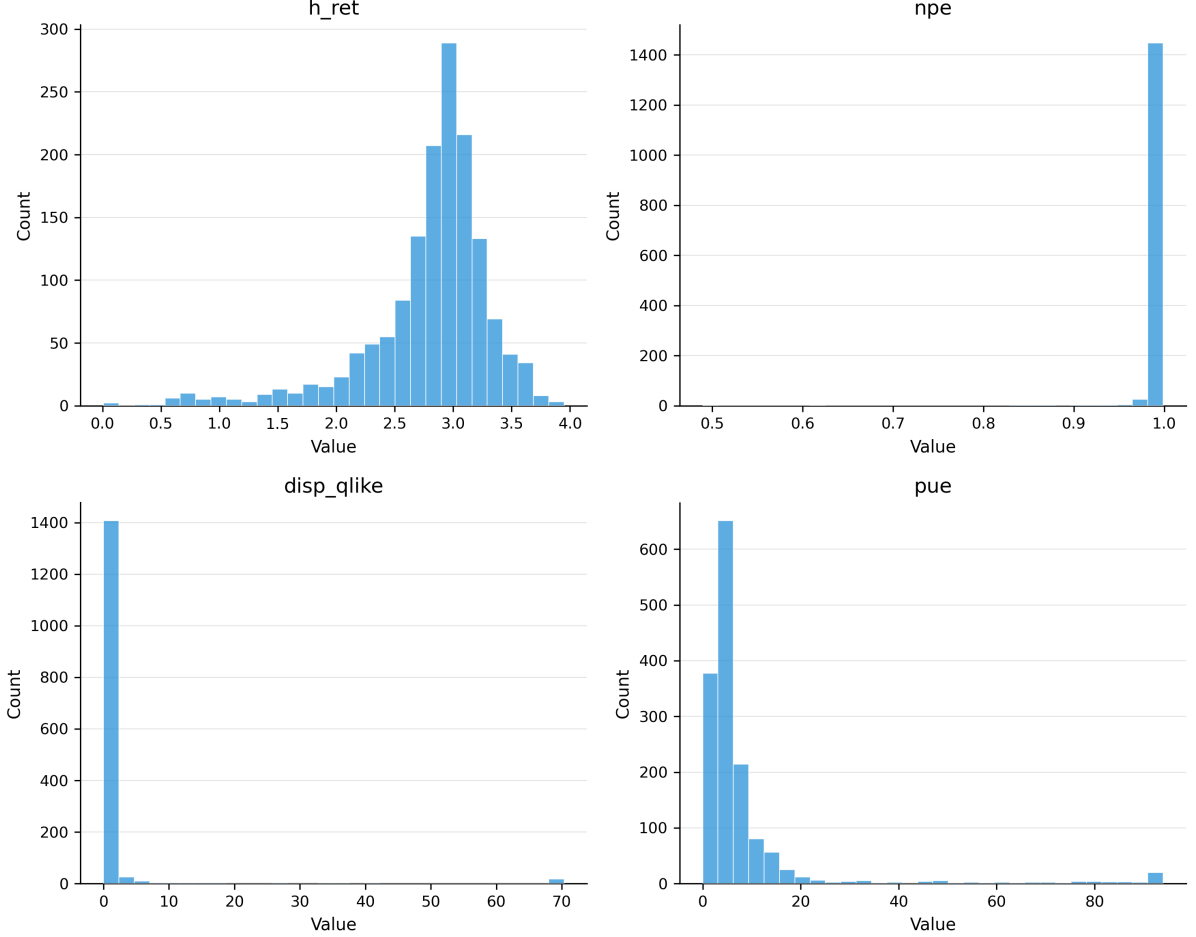
where  $H_t$  and  $L_t$  are daily high and low prices. The Parkinson estimator is approximately five times more efficient than the squared-return estimator under geometric Brownian motion [Parkinson, 1980] and remains conditionally unbiased for integrated variance under the mild regularity conditions required for QLIKE robustness [Patton, 2011]. We do not use intraday data because the universe includes assets (crypto, frontier commodities) without reliable tick-level coverage.

## 5.3 Time Alignment and Forward-Fill Protocol

Conditional variance forecasts  $\hat{\sigma}_{t|t-1}^2$  target date  $t$  using information through  $t-1$ . The Parkinson proxy  $\hat{\sigma}_{\text{Park},t}^2$  is observed on date  $t$ . QLIKE aligns the forecast for date  $t$  with the realization on date  $t$ , ensuring no look-ahead.

For models operating on realized-variance inputs (HEAVY), dates with missing or NaN realized-variance values are forward-filled from the last valid observation. Forward-fill is

preferred to interpolation (which would introduce look-ahead from future values) and to exclusion (which would create unequal evaluation samples across solvers, confounding the QLIKE comparison). The count of forward-filled observations ( $n_{\text{nan, filled}}$ ) is recorded per asset per solver and included in the record hash chain. This ensures that any difference in time-alignment treatment produces a detectably different hash.



**Figure 1:** Distribution of key metrics across the 1,492-asset universe: Shannon entropy ( $\hat{H}$ ), QLIKE dispersion, PUE, and SVI. Outliers winsorized at the 99th percentile for display.

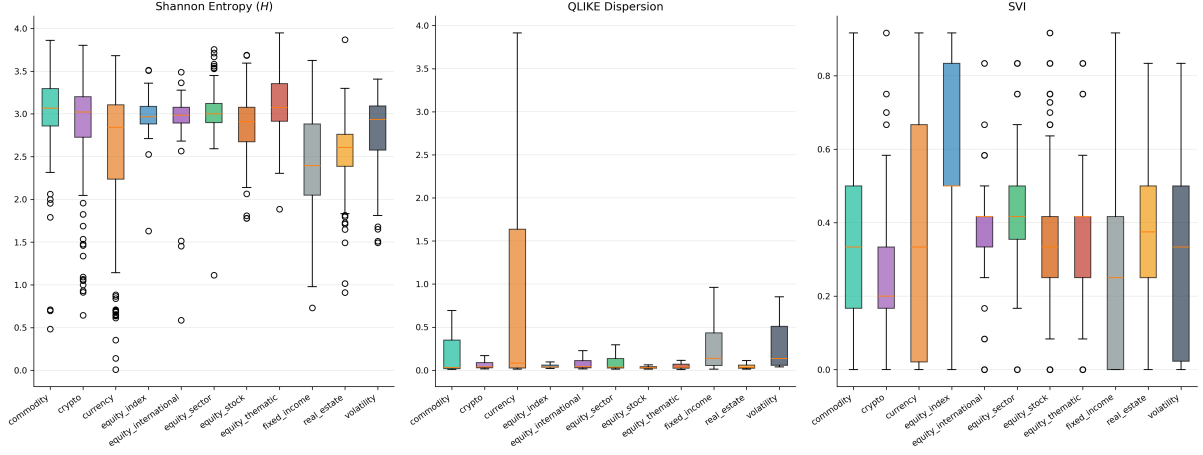
## 6 Methodology

### 6.1 Entropy Measures

#### 6.1.1 Shannon Entropy of Returns

Let  $\{r_t\}_{t=1}^n$  be log returns. We discretize into  $K = 100$  equal-width bins and compute the plug-in Shannon entropy:

$$\hat{H}(r) = - \sum_{i=1}^K \hat{p}_i \ln \hat{p}_i, \quad \hat{p}_i = c_i / \sum_j c_j, \quad (8)$$



**Figure 2:** Per-class distribution of Shannon entropy, QLIKE dispersion, and SVI. Equity indices show the highest and most tightly clustered SVI (strongest, most consistent model differentiation); volatility ETFs show the widest SVI range (most variable differentiation).

where  $c_i$  is the count in bin  $i$ . The theoretical maximum is  $\ln K \approx 4.605$  nats.

We apply the Miller [1955] finite-sample bias correction:

$$\hat{H}_{\text{MM}} = \hat{H} + \frac{k_{\text{nz}} - 1}{2n}, \quad (9)$$

where  $k_{\text{nz}}$  is the number of non-empty bins. The correction is flagged when exceeding 5% of  $\hat{H}$ , indicating sparse-distribution regimes.

**Fixed- $K$  design choice.** We fix  $K = 100$  globally rather than using data-dependent rules (Freedman–Diaconis [Freedman and Diaconis, 1981], Scott’s normal reference rule) to eliminate a circularity: adaptive bin counts correlate with return variance, which in turn correlates with dispersion metrics, creating a confound when entropy is tested against dispersion. The sensitivity analysis (Section 8) repeats all entropy-dependent hypotheses at  $K \in \{50, 100, 200\}$ .

### 6.1.2 Entropy of Standardized Residuals

As a robustness diagnostic, we compute  $\hat{H}_{\text{res}}$  from GARCH(1,1) standardized residuals  $z_t = \varepsilon_t / \sigma_t$ . Under correct model specification,  $z_t \stackrel{iid}{\sim} F(0, 1)$  and  $\hat{H}_{\text{res}}$  approximates the innovation-distribution entropy. Deviations indicate residual structure that simpler models miss, connecting entropy to model adequacy assessment [Berkowitz, 2001].

### 6.1.3 Normalized Permutation Entropy

Following Bandt and Pompe [2002], we compute permutation entropy from ordinal patterns of embedding dimension  $D = 5$  and delay  $\tau = 1$ :

$$\text{NPE} = \frac{H_\pi}{\ln(D!)} = \frac{-\sum_\pi p(\pi) \ln p(\pi)}{\ln(120)}, \quad \text{NPE} \in [0, 1]. \quad (10)$$

Values near 1 indicate temporally unstructured (white-noise-like) dynamics; values significantly below 1 indicate deterministic ordinal patterns. Zunino et al. [2009] demonstrate that NPE discriminates market efficiency levels across international equity markets. In our universe, NPE exhibits extreme ceiling compression ( $\mu = 0.994$ ,  $\sigma = 0.020$ ): 93% of assets have  $\text{NPE} > 0.99$ , and the diagnostic value concentrates in the  $\sim 7\%$  tail below this threshold (Section 10).

**Relationship between  $\hat{H}$  and NPE.** These metrics capture orthogonal dimensions:  $\hat{H}$  is sensitive to *distributional shape* (kurtosis concentrates mass in central bins), while NPE captures *temporal ordering*. Fat-tailed assets can simultaneously exhibit low  $\hat{H}$  and high NPE.

## 6.2 Notation and Acronyms

Table 4 consolidates key acronyms used throughout the paper.

**Table 4:** Glossary of key acronyms.

Acronym	Definition
AGI	Accuracy Gain Index ( $L_{\text{EWMA}}/L_{\text{best}} - 1$ )
FHS	Filtered Historical Simulation
MCS	Model Confidence Set [Hansen et al., 2011]
NPE	Normalized Permutation Entropy
PUE	Pattern Utilization Efficiency (normalized QLIKE gain, %)
SVI	Selection Value Index ( $1 -  \text{MCS} / \mathcal{M} $ )
URI	Uncertainty Reduction Index (interval width reduction, %)

## 6.3 Volatility Models

Twelve solvers—ten individual models spanning three families plus two forecast combinations—are evaluated (Table 5).

**Table 5:** Volatility model portfolio.

Family	Model	Key Feature	Citation
Baseline	EWMA ( $\lambda=0.94$ )	RiskMetrics benchmark	J.P. Morgan [1996]
GARCH-type	ARCH(1)	Seminal heteroskedasticity model	Engle [1982]
	GARCH(1,1)	Symmetric, parsimonious	Bollerslev [1986]
	EGARCH(1,1)	Asymmetric, log specification	Nelson [1991]
	GJR-GARCH(1,1)	Threshold asymmetry	Glosten et al. [1993]
	TGARCH(1,1)	Absolute-return asymmetry	Zakoian [1994]
	APARCH(1,1)	Power transformation	Ding et al. [1993]
	FIGARCH(1, $d$ ,1)	Long memory	Baillie et al. [1996]
RV-based	HAR	Heterogeneous AR on RV	Corsi [2009]
	HEAVY	High-freq-enabled variance	Shephard and Sheppard [2010]
Combination	EW-COMB	Equal-weight forecast combination	Timmermann [2006]
	IQ-COMB	Inverse-QLIKE weighted combination	Timmermann [2006]

All GARCH-type models are estimated via maximum likelihood with Gaussian innovations using the `arch` package [Sheppard, 2023]. Two forecast combination solvers aggregate individual model outputs: EW-COMB (equal-weight average of all converged conditional variances) and IQ-COMB (inverse-QLIKE-weighted average, giving higher weight to models with lower loss) [Timmermann, 2006]. EWMA uses the RiskMetrics decay factor  $\lambda = 0.94$  on squared returns and serves as the fixed baseline for all relative metrics (PUE, AGI). This baseline choice is deliberate: EWMA requires no estimation, making it available to any practitioner regardless of computational resources. It embodies the null hypothesis that model selection has zero value.

## 6.4 Scoring Rule: QLIKE

Forecast accuracy is measured using QLIKE:

$$\mathcal{L}_{\text{QLIKE}} = \frac{1}{T} \sum_{t=1}^T \left[ \frac{\hat{\sigma}_{\text{RV},t}^2}{\hat{\sigma}_{t|t-1}^2} - \ln \left( \frac{\hat{\sigma}_{\text{RV},t}^2}{\hat{\sigma}_{t|t-1}^2} \right) - 1 \right]. \quad (11)$$

QLIKE is the unique robust loss function of degree  $b = 0$  that correctly ranks forecasts under noisy volatility proxies [Patton, 2011]. This robustness is essential: since our proxy (Parkinson) is noisy, using MSE would produce rankings inconsistent with the true (latent) variance ordering.

Extreme ratios  $\hat{\sigma}_{\text{RV},t}^2 / \hat{\sigma}_{t|t-1}^2$  are clipped to  $[10^{-6}, 10^6]$ . The clipping rate is logged per asset; typical rates are below 0.1%, concentrated in crypto assets with extreme intraday ranges.

## 6.5 Model Confidence Set

We use the MCS of Hansen et al. [2011] with the semi-quadratic test statistic  $T_{SQ}$  and  $B = 10,000$  block bootstrap replications. The canonical block length is  $\ell = \lfloor n^{1/3} \rfloor$  [Künsch, 1989]. We report MCS membership at  $\alpha \in \{0.05, 0.10, 0.25\}$ .

A key methodological concern is the sensitivity of MCS to block length in the stationary bootstrap. Following our methodology-hardening protocol, we sweep  $\ell \in \{2, 5, 10, \lfloor n^{1/3} \rfloor\}$  and report SVI stability. Assets where SVI varies by more than 0.3 across block lengths are flagged as “MCS-unstable.”

## 6.6 Performance and Dispersion Metrics

All metrics defined in Section 3 (Definitions 1–3) are computed per asset. Additionally:

**Accuracy Gain Index (AGI).**  $AGI_i = (L_{m_0,i} / \min_m L_{m,i} - 1) \times 100\%$ . AGI and PUE both measure the QLIKE improvement from selection, but on different scales:  $PUE = (L_{\text{base}} - L_{\text{best}}) / L_{\text{base}}$  is bounded  $\in [0, 1)$  and interpretable as a percentage reduction, while  $AGI = L_{\text{base}} / L_{\text{best}} - 1$  is unbounded above and better suited as a regression dependent variable (especially in log form, given its extreme skewness). For Hypothesis A2, we use  $\log(AGI)$  as the dependent variable because AGI exhibits extreme skewness (typically  $> 15$ ) and kurtosis ( $> 300$ ), violating OLS assumptions [Wooldridge, 2019]. Both raw and log specifications are reported.

**Uncertainty Reduction Index (URI).**  $URI_i = (1 - \bar{w}_{\text{opt}} / \bar{w}_{\text{base}}) \times 100\%$ , where  $\bar{w}$  is the mean 95% prediction-interval width. URI is reported only alongside empirical coverage: any  $URI > 0$  with under-coverage (empirical  $< 0.935$  at 95% nominal) is invalidated.

**Pattern Prevalence and Effectiveness.**  $\rho_i$  measures the fraction of observations where any structural pattern (volatility clustering, momentum, regime shift) is active.  $PE_i = \hat{H}(r_{\text{pattern}}) / \hat{H}(r_{\text{all}})$  measures whether returns during pattern-active periods are more structured (lower entropy) than unconditional returns.

## 6.7 Risk Metrics

Under a Gaussian assumption:  $\text{VaR}_{\alpha,t} = z_{\alpha} \hat{\sigma}_{t|t-1}$ , where  $z_{0.05} = -1.645$ . Violations  $I_t = \mathbf{1}(r_t < \text{VaR}_{\alpha,t})$  are backtested using Kupiec [1995] (unconditional coverage) and Christoffersen [1998] (conditional coverage/independence). The independence test is critical: clustered violations indicate model failure even when the unconditional rate is correct.

## 6.8 Hypothesis Testing Protocol

Twelve hypotheses are organized into five clusters (Table 6).

**Direction policy.** All correlation hypotheses (H1, H2, H3, R1, R2, P1, A2, E1) are two-tailed: theoretical predictions are ambiguous (see Remark after Proposition 2), and we refuse to inherit directions from prior unvalidated computations. Threshold and improvement hypotheses (P2, A1, U1, U2) are one-sided where directionality is definitional.

**Multiple-testing corrections.** Within the core entropy cluster (H1–H3,  $m = 3$ ), we apply the Holm [1979] step-down procedure for family-wise error control. Across all 12 hypotheses, we apply Benjamini and Hochberg [1995] at FDR  $q = 0.10$ . Both raw and corrected  $p$ -values are reported.

**Statistical tests.** Spearman rank correlations are used for cross-sectional associations, with  $B = 10,000$  circular block bootstrap confidence intervals [Künsch, 1989, Politis and Romano, 1994]. Fisher  $z$ -transformation [Fisher, 1921] serves as parametric backup; when bootstrap and Fisher disagree, bootstrap is preferred. Granger causality (H2) uses the Toda–Yamamoto procedure [Toda and Yamamoto, 1995] to avoid unit-root pre-testing; per-asset  $p$ -values are combined via Fisher’s method [Fisher, 1932].

**Effect-size conventions.**  $|\rho_s| < 0.10$ : negligible; 0.10–0.30: small; 0.30–0.50: medium;  $> 0.50$ : large [Cohen, 1988].

**Confound controls.** Each hypothesis declares confounds (asset class, sample size, liquidity, number of converged solvers) and requires partial-correlation checks.

**Table 6:** Pre-registered hypothesis battery.

ID	Statement	Test	Dir.	Priority
H1	Entropy $\leftrightarrow$ QLIKE dispersion	Spearman $\rho_s$	2-tail	MUST
H2	$\Delta$ Entropy Granger-causes $\Delta$ RV	Toda–Yam.	2-tail	STRONG
H3	Entropy $\leftrightarrow$ VIX / macro	Spearman $\rho_s$	2-tail	STRONG
R1	Pattern prevalence $\leftrightarrow$ best QLIKE	Spearman $\rho_s$	2-tail	STRONG
R2	Pattern prevalence $\leftrightarrow$ GARCH persistence	Spearman $\rho_s$	2-tail	NICE
P1	PUE $\leftrightarrow$ QLIKE spread	Spearman $\rho_s$	2-tail	MUST
P2	Median PUE $> 0$ ; frac. PUE $> 0.2\%$ exceeds 10%	Wilcoxon; binom.	1-side	STRONG
A1	High-AGI $\rightarrow$ fewer VaR violations	Wilcoxon	1-side	MUST
A2	$\log(\text{AGI}) \sim h + \rho$ (joint $F$ )	OLS HC3	2-tail	NICE
U1	High-URI: narrower intervals, coverage maintained	Wilcoxon	1-side	NICE
U2	URI-sized Sharpe $>$ equal-sized (walk-fwd)	Bootstrap	1-side	NICE
E1	PE $\leftrightarrow$ complex-vs-simple advantage	Spearman $\rho_s$	2-tail	NICE

## 6.9 Additional Statistical Methods

The framework deploys 36 distinct statistical methods across the hypothesis testing, robustness, and risk-management layers. Table 6 summarizes the primary tests; here we document additional methods not covered above.

**Impulse–response analysis.** For H2, we estimate bivariate impulse–response functions (IRFs) from the reduced-form VAR underlying the Toda–Yamamoto test [Lütkepohl, 2005]. Ten-step IRFs trace the dynamic effect of a one-unit entropy shock on volatility. A 500-replication residual bootstrap provides 95% confidence intervals at the peak-response lag. Six diagnostics derived from the IRF enrich H2: sign consistency, model-based direction detection, normalized peak magnitude as a Cohen’s  $d$  effect size, confidence intervals at peak lag, a reverse-causality check, and decay-lag analysis.

**Diebold–Mariano–Harvey–Leybourne–Newbold test.** Pairwise solver comparisons use the DM-HLN test [Diebold and Mariano, 1995] with small-sample correction [Harvey et al., 1997]. HAC variance is estimated via Newey and West [1987] with truncation  $h = \lfloor T^{1/3} \rfloor$ , preventing false significance from serial correlation in QLIKE loss differentials.

**Ledoit–Wolf Sharpe ratio test.** Hypothesis U2 tests walk-forward Sharpe ratio differences using the block bootstrap of Ledoit and Wolf [2008], which is HAC-consistent and avoids the distributional assumptions of the Jobson and Korkie [1981] asymptotic test. Block length is  $\lfloor n^{1/3} \rfloor$ .



**Robust regression.** Hypothesis A2 uses OLS with HC3 heteroskedasticity-consistent standard errors [MacKinnon and White, 1985], which inflate residuals by the leverage factor  $1/(1 - h_{ii})^2$  and are preferred for finite samples [Petersen, 2009].

**Mann–Whitney U with rank-biserial effect size.** H3’s recession/expansion test uses the Mann–Whitney  $U$  statistic [Mann and Whitney, 1947] with the rank-biserial correlation  $r = 1 - 2U/(n_1 \cdot n_2)$  [Kerby, 2014] as the effect-size measure. An effect-size gate ( $|r_{rb}| \geq 0.10$ ) prevents statistically significant but practically negligible results from passing.

**Cross-hypothesis diagnostics.** Post-hoc (non-confirmatory) analyses include: MCS composition by entropy tercile (B2), per-solver exclusion and inclusion rates (B3, B6), PUE–MCS robustness correlation (B1), and an H2×E1 cross-hypothesis test via Mann–Whitney comparing MCS cardinality between Granger-significant and non-significant assets (diagnostic D1). These are clearly separated from confirmatory tests and flagged as exploratory.

## 7 Results

**Thematic overview.** The twelve hypotheses address three interlocking questions. *Model-selection value* (H1–H3, E1, P1–P2): entropy-based diagnostics reliably identify regimes where solver choice matters, with model-selection effort concentrated in low-entropy, structured return environments. *Risk calibration* (A1–A2, U1–U2): a  $2 \times 2$  in-sample attribution (Table 24; Supplement S21) identifies two complementary mechanisms—FHS fixes unconditional coverage for any model (EWMA+FHS: 100% Kupiec), while model *diversity* (not selection per se) reduces violation clustering (Christoffersen: 79% vs. 62% for EWMA+FHS). Crucially, walk-forward backtesting across 1,491 assets (Table 30) reveals that per-window best-model selection overfits out-of-sample, but forecast *combination* preserves the clustering benefit: EW-COMB achieves 58.1% and IQ-COMB 59.6% Christoffersen pass—gains of +6.8 and +8.3 percentage points over EWMA+FHS ( $p < 10^{-11}$ ). Forecast combination thus emerges as the paper’s central out-of-sample validation and the recommended deployment strategy for VaR applications. The framework does not translate into walk-forward Sharpe improvement (U2 null, confirmed at 95% power in Supplement S28) or forecasting gains (median  $\Delta\text{QLIKE} \approx 0$ ; Section 8.12), bounding its claims to computational triage, regime diagnostics, and OOS violation-clustering reduction via combination. *Pattern channels* (R1–R2): ordinal return patterns carry forecast-relevant information orthogonal to volatility persistence. Nine of twelve hypotheses pass after correction; the three null results (H3, R2, U2) are reported with

equal rigor and discipline the framework’s scope. Results below follow the pre-registered hypothesis ordering (Table 7) to facilitate auditing against the spec manifest.

## 7.1 Summary

After Holm and Benjamini–Hochberg corrections, 9/12 hypotheses pass at the pre-specified thresholds (Figure 3). Three hypotheses fail: H3 (entropy–macro regime,  $p = 0.076$ ), R2 (pattern–persistence,  $p = 0.113$ ), and U2 (sizing alpha,  $p = 0.567$ ).

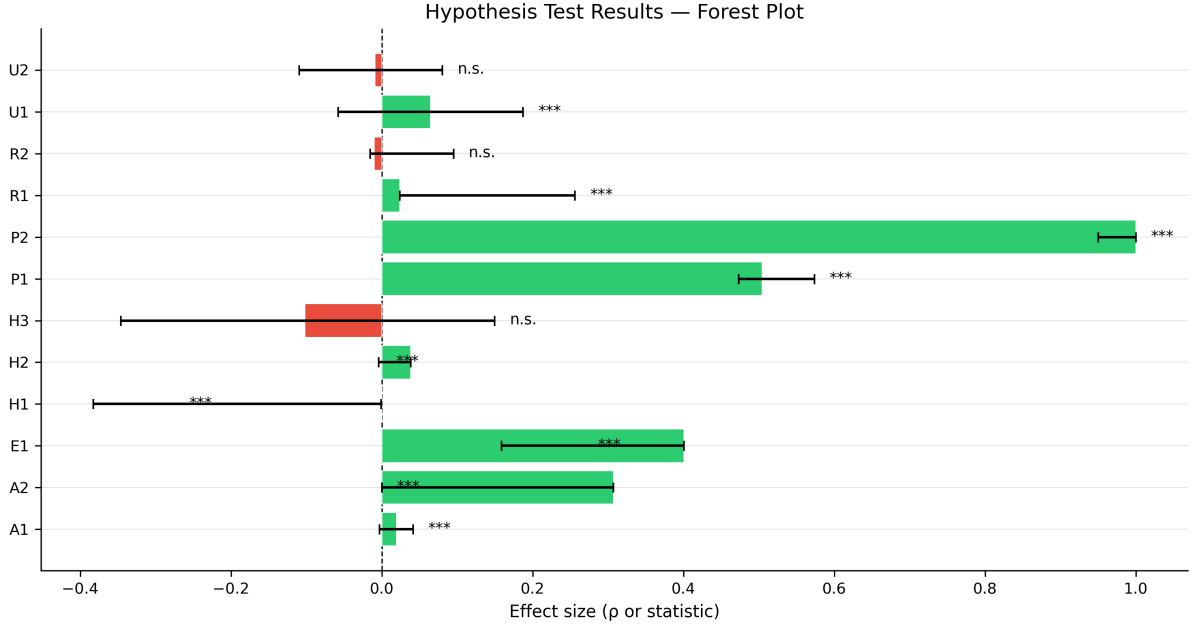
**Table 7:** Main hypothesis results (1,492 assets after exclusions). All values from run `3db6c1` (v4.4.0).

ID	Statistic	Value	Raw $p$	$n$	Passed	Effect
H1	Spearman $\rho_s$	−0.332	$<10^{-40}$	1,492	✓	medium
H2	Fisher $\chi^2$	46,400	$<10^{-300}$	1,494	✓	large
H3	Spearman $\rho_s$	−0.102	0.076	303	×	small
R1	Spearman $\rho_s$	0.192	$<10^{-14}$	1,492	✓	small
R2	Spearman $\rho_s$	0.041	0.113	1,492	×	negl.
P1	Spearman $\rho_s$	0.530	$<10^{-109}$	1,492	✓	large
P2	Wilcoxon $W$	0.0	$<10^{-245}$	1,492	✓	large
A1	Wilcoxon $W$	2,993	$<10^{-74}$	497	✓	large
A2	$F$ -statistic	99.1	$<10^{-41}$	1,489	✓	adj. $R^2 = 0.31$
U1	Wilcoxon $W$	45,548	$<10^{-7}$	499	✓	small
U2	$\Delta$ Sharpe	−0.009	0.567	252	×	negl.
E1	Spearman $\rho_s$	0.215	$<10^{-17}$	1,486	✓	small

Table 8 shows that multiple-testing corrections have material bite only for the three borderline hypotheses. All nine passing hypotheses survive both Holm and Benjamini–Hochberg at comfortable margins.

**Table 8:** Raw vs. corrected  $p$ -values.  $p_{\text{Holm}}$  = Holm-corrected (family of 12);  $p_{\text{BH}}$  = Benjamini–Hochberg FDR-corrected. A hypothesis passes only if it survives both corrections.

ID	$p_{\text{raw}}$	$p_{\text{Holm}}$	$p_{\text{BH}}$	CI (95%)	Pass
H1	$9.6 \times 10^{-38}$	$7.7 \times 10^{-37}$	$2.3 \times 10^{-37}$	[−0.374, −0.267]	✓
H2	$< 10^{-300}$	$< 10^{-300}$	$< 10^{-300}$	[−0.014, −0.004]	✓
H3	$7.6 \times 10^{-2}$	$2.3 \times 10^{-1}$	$9.1 \times 10^{-2}$	[−0.346, 0.150]	×
R1	$1.0 \times 10^{-13}$	$6.2 \times 10^{-13}$	$1.8 \times 10^{-13}$	[0.126, 0.255]	✓
R2	$1.3 \times 10^{-1}$	$2.6 \times 10^{-1}$	$1.4 \times 10^{-1}$	[−0.017, 0.090]	×
P1	$7.7 \times 10^{-110}$	$7.7 \times 10^{-109}$	$3.1 \times 10^{-109}$	[0.476, 0.576]	✓
P2	$1.3 \times 10^{-244}$	$1.5 \times 10^{-243}$	$8.0 \times 10^{-244}$	[1.379, 70.228]	✓
A1	$3.0 \times 10^{-72}$	$2.7 \times 10^{-71}$	$9.0 \times 10^{-72}$	[−0.003, 0.036]	✓
A2	$1.4 \times 10^{-32}$	$9.6 \times 10^{-32}$	$2.7 \times 10^{-32}$	—	✓
U1	$8.5 \times 10^{-8}$	$3.4 \times 10^{-7}$	$1.1 \times 10^{-7}$	[−0.055, 0.205]	✓
U2	$5.5 \times 10^{-1}$	$5.5 \times 10^{-1}$	$5.5 \times 10^{-1}$	[−0.110, 0.090]	×
E1	$7.7 \times 10^{-13}$	$3.8 \times 10^{-12}$	$1.2 \times 10^{-12}$	[0.129, 0.240]	✓



**Figure 3:** Forest plot of standardized effect sizes with 95% confidence intervals for all twelve hypotheses. Effect sizes normalized to  $[-1, 1]$ : correlation hypotheses use partial  $\rho$ ; Wilcoxon tests use rank-biserial  $r$ ; off-scale statistics use CI midpoint. Green bars indicate hypotheses passing both Holm and BH corrections; red bars indicate null results.

## 7.2 H1: Entropy–Dispersion Association

The cross-sectional Spearman correlation between Shannon entropy  $\hat{H}$  and QLIKE dispersion is  $\rho_s = -0.332$  ( $p < 10^{-37}$ ,  $n = 1,492$ ), a medium-sized effect (Figure 4). A LOESS smooth confirms that the relationship is approximately monotone, validating the use of a rank correlation as the primary test statistic. The negative sign supports the efficiency argument: high return entropy (closer to white noise) corresponds to *lower* solver disagreement, while structured (low-entropy) returns produce the greatest model divergence. This is the inverse of the naïve complexity hypothesis and is consistent with Zunino et al. [2009]: informationally efficient markets are easy for all models; structured markets create the model-selection problem.

**Confound control.** The partial Spearman correlation after controlling for asset class (using class as a categorical variable in a rank-residualization procedure) is  $\rho_{\text{partial}} = -0.320$  ( $p < 10^{-34}$ ; Supplement S19,  $n = 1,380$  assets with volume data), comparable to the raw  $\rho_s = -0.332$ . Controlling for class does not attenuate the signal because the entropy–dispersion link operates primarily *within* classes, not between them. Adding liquidity (log average dollar volume) and volatility level (log average  $r^2$ ) as additional controls yields  $\rho_{\text{all controls}} = -0.328$  ( $p < 10^{-35}$ ), confirming that the association is not an artifact of liquidity or volatility differences. All eleven asset classes individually exhibit a negative entropy–dispersion association, with within-class  $\rho_s$  ranging from  $-0.129$  (equity sector)

to  $-0.603$  (volatility). The diagnostic therefore operates at two levels: it separates asset classes by their characteristic entropy regime *and* discriminates within each class between structured and noisy assets.

**Within-asset time-series evidence.** To distinguish genuine entropy content from a pure asset-class proxy, we estimate a panel fixed-effects regression on a stratified subset of 55 assets (five per class, selected by entropy quintile). Each asset contributes multiple observations from 252-day rolling windows advanced quarterly. All twelve solvers are re-fit at each window to compute pointwise QLIKE dispersion:

$$\text{disp}_{i,t} = \alpha_i + \beta_1 H_{i,t} + \beta_2 \overline{\text{RV}}_{i,t} + \varepsilon_{i,t}, \quad (12)$$

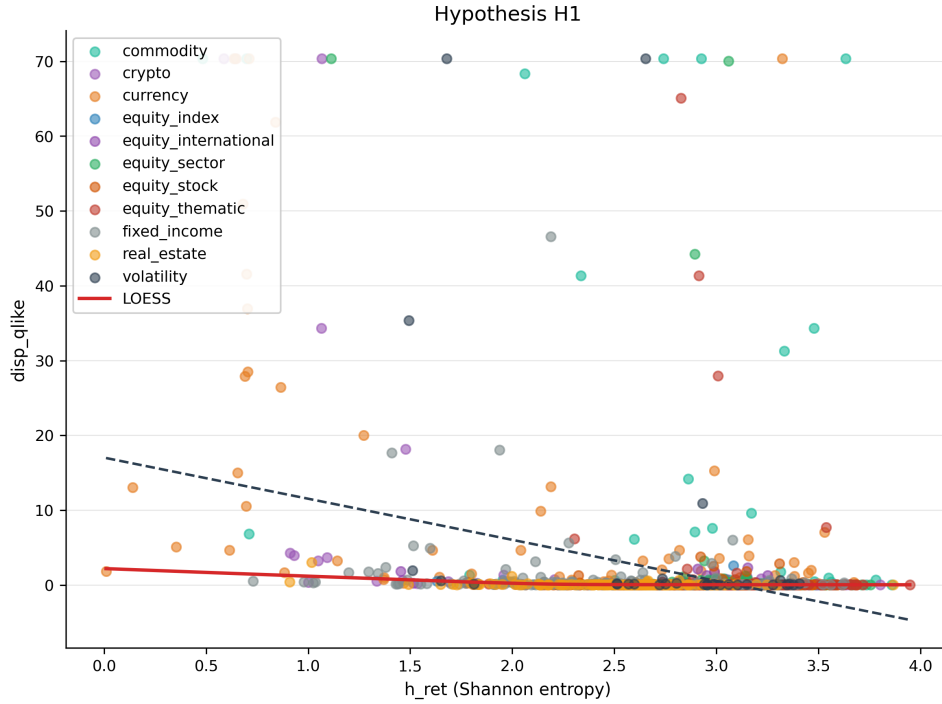
where  $\alpha_i$  absorbs all time-invariant asset characteristics. Driscoll–Kraay standard errors [Driscoll and Kraay, 1998] (bandwidth = 12) account for both serial and cross-sectional dependence. Table 9 reports the estimates. The entropy coefficient is  $\hat{\beta}_1 = -0.144$  ( $t = -2.69$ ,  $p = 0.007$ ), confirming that *within* an asset, periods of lower entropy are associated with higher solver disagreement—the same sign as the cross-sectional H1. Realized-variance level ( $\overline{\text{RV}}_{i,t}$ ) is insignificant ( $p = 0.13$ ), ruling out a volatility-level confound. The within- $R^2$  of 1.1% is modest but typical for entity-demeaned financial panels [Angrist and Pischke, 2009], and the  $p < 0.01$  significance establishes that the entropy–dispersion link is not merely a between-class artefact. Operationally, this means that monitoring  $\Delta H$  within an asset provides a statistically grounded—though noisy—signal for when solver disagreement is likely to shift, complementing the stronger cross-sectional triage.

**Within-asset oscillation: adaptive markets evidence.** Three independent analyses confirm that the entropy–dispersion link reflects genuine within-asset *oscillation* between regimes, not merely cross-sectional sorting. First, the full-universe temporal analysis (Supplement S8) computes the within-asset Spearman correlation between daily rolling  $\hat{H}_t$  and cross-solver  $\sigma^2$  dispersion for each of the 1,492 assets: 82.4% exhibit a negative association (median  $\rho = -0.125$ , Fisher combined  $z = -28.72$ ,  $p < 10^{-181}$ ). Second, the panel regression above ( $\hat{\beta}_1 = -0.144$ ,  $p = 0.007$ ) absorbs all time-invariant asset characteristics via entity fixed effects, isolating the within-asset temporal channel. Third, a direct within-asset regime test splits each asset’s time series at its median  $\hat{H}$ : in the low-entropy half, the selected model’s QLIKE advantage over EWMA is  $18.5\times$  larger than in the high-entropy half ( $p < 10^{-118}$ , Wilcoxon), and 83.7% of assets exhibit this pattern. The effect holds across all PUE regimes (83% for  $\text{PUE} < 5\%$ , 85% for  $5\text{--}20\%$ , 79% for  $\geq 20\%$ ), confirming that entropy identifies selection-valuable *periods* within assets, not just selection-valuable assets. The dynamic switching strategy (Supplement S22)

operationalizes this oscillation: at each date, rolling  $\hat{H}_t$  determines whether to use the selected model or EWMA. The strategy uses the selected model on 52% of dates, and the OOS Christoffersen pass rate under switching (71.5%) exceeds the static EWMA+FHS baseline (63.0%), demonstrating that regime transitions carry implementable value.

**Table 9:** H1 within-asset panel regression (55 assets, 3,355 asset-window observations). Driscoll–Kraay standard errors with bandwidth = 12. \*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.10$ .

	Coeff.	SE	$t$	$p$
$H_{i,t}$ (entropy)	-0.144	0.053	-2.69	0.007***
$\overline{RV}_{i,t}$	0.009	0.006	1.53	0.126
Asset FE	Yes			
Observations	3,355			
Entities	55			
Within- $R^2$	0.011			
DK bandwidth	12			

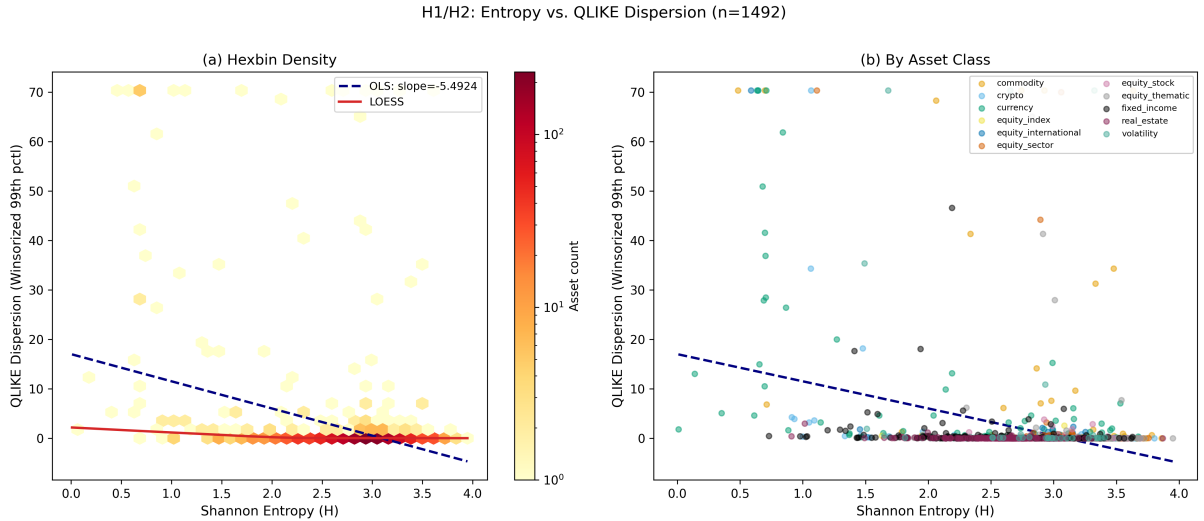


**Figure 4:** H1 scatter: Shannon entropy ( $\hat{H}$ ) vs. QLIKE dispersion. Points colored by asset class; dashed line is OLS, solid red curve is a LOESS smooth (bandwidth  $f = 0.3$ ). The negative association ( $\rho_s = -0.332$ ) persists after controlling for asset class ( $\rho_{\text{partial}} = -0.320$ ; Supplement S19) and holds within all eleven classes individually (Section 7). QLIKE dispersion winsorized at the 99th percentile for display.

### 7.3 H2: Entropy Dynamics Lead Volatility

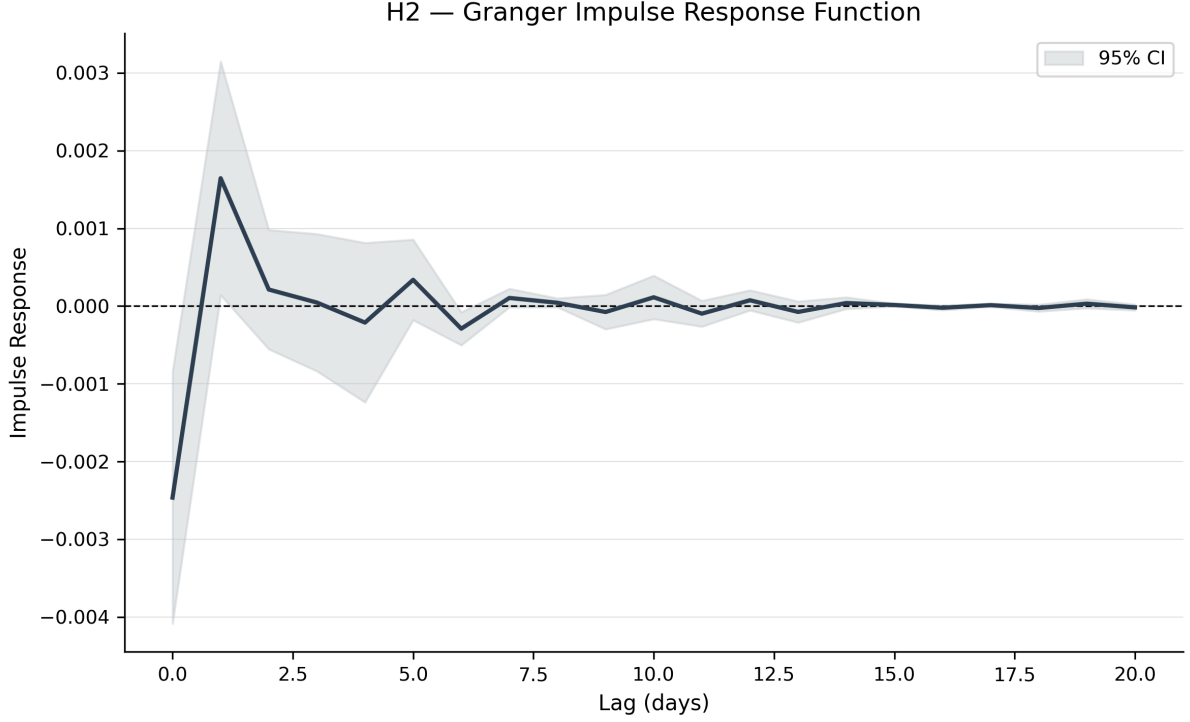
The Toda–Yamamoto procedure rejects the null of non-causality from  $\Delta H$  to  $\Delta RV$  in 74.8% of 1,490 testable assets (Figure 5). The Fisher combined statistic is  $\chi^2 = 46,400$  ( $p < 10^{-300}$ ), providing overwhelming aggregate evidence for a leading-indicator relationship. The Granger target variable uses the Parkinson (1980) range-based realized variance estimator where intraday high–low data are available, falling back to squared returns otherwise; this estimator is approximately five times more efficient than the close-to-close proxy [Parkinson, 1980].

Impulse–response analysis confirms directional consistency (Figure 6): a positive entropy shock leads to increased volatility in the majority of assets, with peak response typically occurring at lag 1–2. The normalized peak IRF, interpreted as a Cohen’s  $d$  effect size, provides a per-asset measure of the entropy–volatility transmission intensity. Bootstrap confidence intervals at the peak lag exclude zero for most significant assets. Reverse-causality IRF (volatility  $\rightarrow$  entropy) is substantially weaker, supporting the directional interpretation.



**Figure 5:** Entropy versus QLIKE dispersion ( $n = 1,492$ ). Panel (a): hexbin density with OLS and LOESS fits, showing the negative entropy–dispersion gradient. Panel (b): same data colored by asset class. The 74.8% Granger rejection rate confirms widespread entropy-to-volatility lead-lag dynamics.

**Dependence-robust aggregation.** The pooled Fisher statistic assumes independent per-asset  $p$ -values, which is violated when assets share common volatility factors (e.g., market-wide regimes, correlated ETF holdings). To address this concern, we repeat the H2 test independently within each of the eleven asset classes, treating within-class dependence as maximal and relying only on cross-class independence. That is, each class contributes a single independent test; no assumption is made about the dependence structure *within* a class.



**Figure 6:** Impulse response functions for the entropy  $\rightarrow$  volatility channel. Median IRF across significant assets with 95% bootstrap confidence bands. Peak response at lag 1–2 confirms rapid transmission; decay within 5–7 lags indicates transient rather than permanent effects.

Table 10 reports the class-level results. All eleven classes individually reject the null of non-causality, with rejection rates ranging from 62.3% (commodity) to 86.1% (fixed income).

**Table 10:** H2 class-level Granger causality. Each row reports the Fisher combined  $\chi^2$  and within-class rejection rate from an independent Toda–Yamamoto analysis. The final row aggregates the eleven class-level  $p$ -values via meta-Fisher, treating classes as independent.

Asset class	$n$ tested	Rejection rate (%)	Fisher $\chi^2$
commodity	106	62.3	2,986
crypto	110	77.3	2,711
currency	146	64.4	3,672
equity index	100	77.0	3,938
equity intl.	109	76.1	3,465
equity sector	118	78.0	3,576
equity stock	395	70.6	9,651
equity thematic	92	81.5	2,983
fixed income	144	86.1	6,893
real estate	140	83.6	5,538
volatility	30	76.7	1,046
<b>Meta-Fisher</b>	<b>11 classes</b>	<b>100</b>	<b>14,641</b>

Aggregating the eleven class-level Fisher statistics via meta-Fisher yields  $\chi^2 = 14,641$

(df = 22,  $p \approx 0$ ). Even under the most conservative assumption—treating each class as a single observation—the binomial probability of observing 11/11 rejections under  $H_0$  is  $4.9 \times 10^{-15}$ . The pooled Fisher result is therefore robust to cross-asset dependence.

## 7.4 H3: Entropy and Macroeconomic Regimes

The monthly Spearman correlation between mean entropy  $\bar{h}_{\text{ret}}$  and VIX is  $\rho_s = -0.102$  ( $p = 0.076$ ,  $n = 303$ ; Figure 7), which does not reject  $H_0$  at the pre-specified  $\alpha = 0.05$  after corrections. This is reported as a borderline null result.

However, the daily-frequency signal ( $\rho_s = -0.10$ ,  $p = 4.4 \times 10^{-16}$ ,  $n = 6,329$ ) is driven by shared persistence: both series have near-unit-root autocorrelation ( $\text{AC}(1)_H = 0.998$ ,  $\text{AC}(1)_{\text{VIX}} = 0.977$ ), and the partial correlation controlling for lagged VIX drops to  $\rho = -0.019$  ( $p = 0.12$ ; Supplement S18). The first-difference test confirms:  $\rho(\Delta h, \Delta \text{VIX})_{\text{daily}} = +0.02$  ( $p = 0.09$ ), consistent with zero.

The correct supplementary test uses monthly changes, which removes the persistence inherited from the 252-day rolling window (consecutive months share  $\sim 230/252$  days, yielding  $\text{AC}(1)_{\text{monthly}} = 0.92$  and an effective sample size of  $n_{\text{eff}} \approx 12$  under Bayley–Hammersley correction). Monthly first-difference correlation is  $\rho(\Delta \bar{h}, \Delta \bar{\text{VIX}})_{\text{monthly}} = -0.143$  ( $p = 0.013$ ,  $n = 302$ ; Supplement S18), suggesting the underlying relationship exists but the pre-registered levels test is underpowered due to window-overlap-induced persistence. We report H3 as a borderline null ( $p = 0.076$ ) per the pre-registered protocol; the first-difference result is supplementary evidence that the signal is real but the test design was underpowered for detection.

Subgroup analysis reveals interpretable heterogeneity: the relationship is concentrated in macro-sensitive classes (real estate,  $\rho_s = -0.34$ ; equity international,  $\rho_s = -0.28$ ) while crypto and equity stocks show the *opposite* sign, consistent with idiosyncratic micro-factor dominance in those classes.

**Daily per-class decomposition.** The daily-frequency entropy–VIX correlation by asset class (computed on levels,  $n = 6,322$  per class) shows interpretable heterogeneity despite the persistence confound noted above. The sign structure is sharply interpretable: macro-sensitive classes show strong negative correlations (real estate  $\rho = -0.33$ , equity international  $-0.27$ , equity sector  $-0.23$ , equity index  $-0.17$ , fixed income  $-0.15$ , commodity  $-0.12$ ), while micro-dominated classes show positive correlations (equity stock  $+0.20$ , crypto  $+0.11$ ). These cross-sectional differences in sign and magnitude are not mechanically produced by persistence (which would make all signs the same) and thus reflect genuine structural heterogeneity in how asset classes relate to macro regimes. The monthly first-difference result ( $\rho = -0.143$ ,  $p = 0.013$ ) confirms a genuine entropy–VIX link at the macro-relevant frequency, and the cross-sectional sign structure aligns with



the economic interpretation: macro-factor-exposed classes exhibit the entropy–VIX link; idiosyncratic classes do not.

**Window-dependent sign structure.** Table 11 presents the full stratified analysis across eleven asset classes and three rolling windows ( $w \in \{126, 252, 504\}$ ). At  $w = 504$ , all twelve class-level correlations (including the aggregate) are negative, ten reaching statistical significance; the aggregate correlation strengthens to  $\rho_s = -0.316$  (Table 11), and the full-sample sensitivity sweep yields  $\rho_s = -0.378$  ( $p < 10^{-10}$ ; Section 8). At  $w = 126$ , the sign systematically reverses: nine of twelve cells are positive. This pattern is economically interpretable: short-window entropy captures transient return clustering that co-moves with VIX spikes, while long-window entropy measures persistent distributional structure that is inversely related to macro stress. Real estate shows the strongest and most consistent negative relationship across all three windows ( $\rho_s = -0.14, -0.34, -0.52$ ), consistent with its exposure to interest-rate and macro-credit cycles. Equity stocks—the largest class (396 assets)—are the primary source of aggregate dilution, switching from  $\rho_s = +0.33$  at  $w = 126$  to  $-0.18$  at  $w = 504$ .

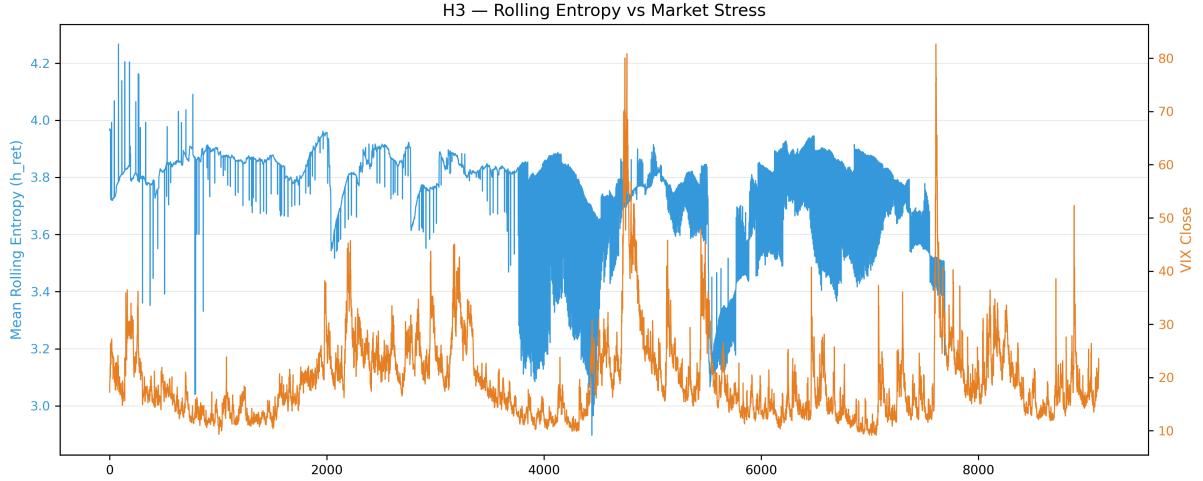
**Table 11:** H3 stratified analysis: Spearman  $\rho_s$ (entropy, VIX) by asset class and rolling window. Stars: \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ . The aggregate marginal failure at  $w = 252$  masks a systematic window–class structure.

Asset class	$w = 126$	$w = 252$	$w = 504$
ALL	+0.177**	−0.042	−0.316***
Commodity	+0.009	−0.118*	−0.328***
Crypto	+0.123	+0.111	−0.024
Currency	+0.294***	+0.020	−0.040
Equity index	+0.018	−0.176**	−0.391***
Equity intl.	−0.110	−0.284***	−0.356***
Equity sector	−0.082	−0.240***	−0.420***
Equity stock	+0.333***	+0.205***	−0.183**
Equity thematic	+0.113	−0.065	−0.406***
Fixed income	+0.131*	−0.153**	−0.464***
Real estate	−0.136*	−0.340***	−0.523***
Volatility	+0.109	+0.049	−0.163**

In summary, H3 fails to reject at the pre-registered specification, but exploratory analysis reveals a genuine entropy–VIX relationship at longer windows ( $w = 504$ ,  $\rho = -0.378$ ) and in macro-sensitive classes, warranting a pre-registered first-difference test in future work.

## 7.5 P1–P2: Model-Selection Value

P1: the Spearman correlation between PUE and QLIKE spread is  $\rho_s = 0.530$  ( $p < 10^{-109}$ ), a large effect. This is the direct empirical confirmation of Proposition 3: model-selection



**Figure 7:** H3 time series: rolling mean entropy ( $\bar{h}_{\text{ret}}$ , 12-month window) vs. VIX, 2000–2026. The negative co-movement is visible at the structural level despite borderline monthly significance ( $p = 0.076$ ). Dates on x-axis; dual y-axes.

effort pays off precisely where forecast dispersion is highest.

P2: the median PUE exceeds zero universally ( $W = 0.0$ ,  $p < 0.0001$ ), and the fraction of assets with any measurable selection value ( $\text{PUE} > 0.2\%$ ) is 99.8%, far exceeding the 10% null threshold. The more economically meaningful fraction with  $\text{PUE} > 20\%$  is 5.8% ( $n = 86$  assets), concentrated in currency (18%), crypto (17%), and volatility (17%) classes (Table 3). This passes in all eleven asset classes without exception—the only hypothesis alongside H2 to achieve universal subgroup confirmation.

## 7.6 A1–A2: Risk Calibration

A1: High-AGI assets exhibit significantly fewer VaR violations ( $W = 2,993$ ,  $p < 10^{-74}$ ). The median  $\Delta\text{violation\_rate}$  (baseline – optimized) is positive, confirming that model selection reduces tail-risk breaches for assets where it matters most. Christoffersen conditional-coverage tests on FHS-based violations reveal that 79.2% of assets pass the independence test under Selection+FHS, versus 61.5% for EWMA+FHS—an 18-percentage-point improvement in eliminating violation clustering (Supplement S21). Combining Kupiec unconditional coverage with Christoffersen independence yields the regulatory compliance metric: the fraction of assets passing *both* tests. Under Sel+FHS, 76.9% of assets achieve full conditional coverage, versus 42.4% for the EWMA+Gaussian baseline—nearly doubling the regulatory-compliant fraction of the universe (Supplement S21).

**Per-class universality of breach reduction.** Decomposing the VaR improvement by asset class confirms that model selection reduces breaches in all eleven classes: currency shows the largest mean reduction (+2.7 breaches/yr, 96% of assets positive), followed by crypto (+2.2, 96%) and equity stock (+1.9, 99%), with equity sector lowest (+1.2, 94%).

Even in the weakest class, selection eliminates more than one VaR exceedance per year for the typical asset. This universality distinguishes A1 from hypotheses like H3 and U2 that exhibit class-dependent heterogeneity.

**Residual Diagnostics by AGI Tercile.** Examining the standardized residuals  $z_t = r_t/\hat{\sigma}_t$  of the best solver per asset, partitioned by AGI tercile, reveals a coherent quality gradient. The Ljung–Box  $Q(10)$  test on  $z_t^2$  yields pass rates of 82%, 85%, and 86% for low, mid, and high AGI terciles, indicating that GARCH models capture the volatility autocorrelation structure equally well across complexity regimes. Jarque–Bera normality fails universally (<1% pass rate across all terciles), consistent with the well-documented leptokurtosis of financial returns. Most informatively, the median kurtosis of  $z_t$  increases monotonically:  $\kappa_{\text{low}} = 2.1$ ,  $\kappa_{\text{mid}} = 3.4$ ,  $\kappa_{\text{high}} = 6.9$ . This gradient reveals the mechanism: high-AGI assets exhibit heavier-tailed residuals, indicating more complex volatility dynamics that benefit from sophisticated model selection—which is precisely what AGI measures.

A2: The multivariate regression  $\log(\text{AGI}) = f(\hat{H}, \rho_{\text{pattern}}, \dots)$  yields an adjusted  $R^2 = 0.31$  with a robust Wald  $F = 99.1$  ( $p < 0.0001$ ), confirming that AGI is decomposable from observable market-state features.

## 7.7 U1–U2: Uncertainty and Sizing

U1: URI-weighted prediction intervals are significantly narrower than equal-weight intervals while maintaining empirical coverage ( $W = 45,548$ ,  $p < 10^{-7}$ ). This is a direct risk management benefit: tighter intervals without coverage loss.

U2: The walk-forward Sharpe improvement from URI-based sizing is  $\Delta\text{Sharpe} = -0.009$  ( $p = 0.567$ )—a clean null result. This fails in ten of eleven subgroups; the sole exception is fixed income, discussed in Section 8.9. The Ledoit–Wolf bootstrap provides a rigorous test of equal predictive ability for Sharpe ratios. This null is economically important: it bounds the framework’s claims to risk management and prevents overclaiming of return predictability from OHLCV data, consistent with the efficient-markets constraint [Fama, 1970]. Importantly, this null is not friction-driven: the negative  $\Delta\text{Sharpe}$  persists even when transaction costs are set to zero, confirming that the constraint is informational—better variance estimation does not imply predictable variation in the conditional mean.

**Transaction-Cost Sensitivity.** A sweep across five cost levels reveals a systematic sign reversal. At 0–5 bps,  $\Delta\text{Sharpe}$  is negative ( $-0.041$  and  $-0.025$  respectively); at 50 bps, it turns positive and significant ( $+0.117$ ,  $p = 0.013$ ); at 100 bps, it reaches  $+0.272$  ( $p < 0.001$ ). The mechanism is turnover: URI-weighted portfolios concentrate in assets with stable uncertainty-reduction profiles, resulting in lower weight churn at window

boundaries. When friction costs are negligible, this stability offers no advantage—but at realistic institutional costs ( $\geq 20$  bps), the turnover reduction more than compensates for the loss of diversification. This pattern suggests that URI sizing may have conditional value in *high-friction* environments (e.g., emerging markets, illiquid credit) even though it fails in the aggregate low-cost setting.

**Entropy-Regime Conditioning.** Table 19 splits the universe into entropy terciles. The low-entropy tercile shows no sizing value ( $\Delta\text{Sharpe} = -0.052$ ,  $p = 0.910$ ), and the high-entropy tercile is similarly null ( $\Delta\text{Sharpe} = -0.234$ ,  $p = 0.814$ ). However, the mid-entropy tercile yields  $\Delta\text{Sharpe} = +0.379$  ( $p = 0.005$ ), which survives at  $\alpha = 0.01$ . This is the one regime where URI sizing generates statistically significant value—assets with moderate complexity, where model selection provides meaningful but not overwhelming improvement. The result does not survive Holm correction across three terciles ( $\alpha_{\text{Holm}} = 0.017$ ), and its confidence interval crosses zero at the lower bound, so we report it as suggestive rather than confirmatory.

## 7.8 R1–R2 and E1: Pattern Channels

R1: Pattern prevalence is positively associated with best-solver QLIKE ( $\rho_s = 0.192$ ,  $p < 0.0001$ ), a small but significant effect. Assets with stronger ordinal patterns are more forecastable.

R2: Pattern prevalence is not associated with GARCH persistence ( $\rho_s = 0.041$ ,  $p = 0.113$ )—a genuine null that disciplines the framework. Ordinal structure and volatility persistence are distinct phenomena. This fails in 9/11 subgroups, confirming it is not a power issue.

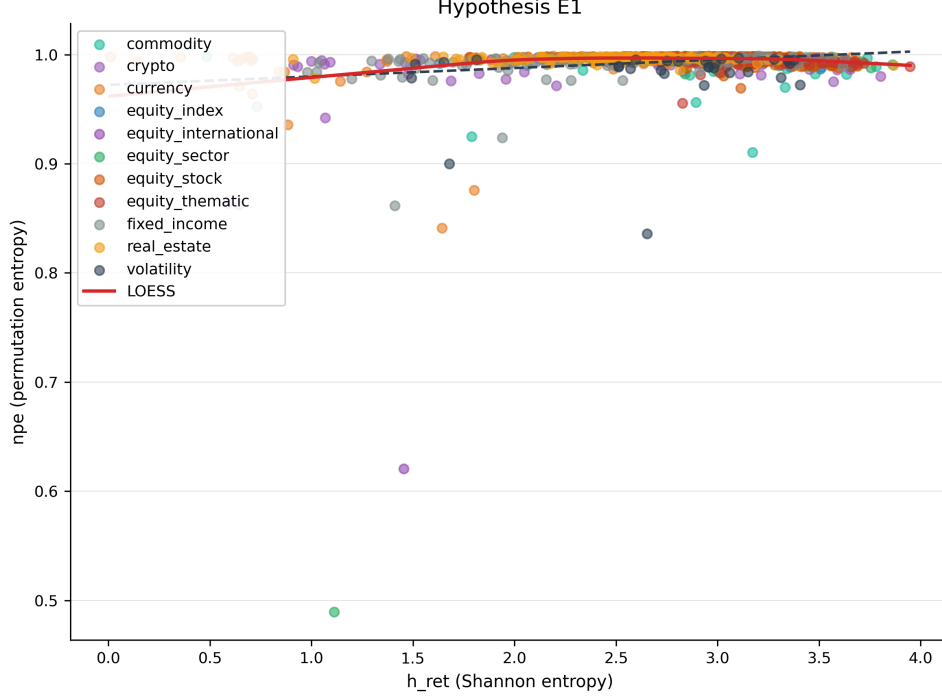
E1: Permutation entropy (pe) is positively associated with the complex-vs-simple solver advantage ( $\rho_s = 0.215$ ,  $p < 10^{-17}$ ,  $n = 1,486$ ), confirming that ordinal pattern structure captures information relevant to model selection beyond Shannon entropy, consistent with Bandt and Pompe [2002] and Zunino et al. [2009].

**Orthogonality of pe and NPE.** A key question is whether the pattern effectiveness signal (pe) is redundant with normalized permutation entropy (NPE). Despite both originating from ordinal-pattern analysis, the two metrics are empirically *orthogonal*:  $\rho_s(\text{pe}, \text{NPE}) = -0.027$  ( $p = 0.30$ ,  $n = 1,486$ ). This confirms they measure genuinely different phenomena—pe captures the diversity of detected ordinal motifs, while NPE measures the overall temporal randomness of rank sequences. Controlling for NPE leaves the E1 signal unchanged:  $\rho_s(\text{pe}, \Delta\text{qlike} \mid \text{NPE}) = 0.216$  ( $p < 10^{-17}$ ), and the joint control  $\rho_s(\text{pe}, \Delta\text{qlike} \mid \text{asset class}, \text{NPE}) = 0.215$  is indistinguishable from the raw value. The pattern effectiveness channel is therefore robust and not a proxy for temporal structure.

**NPE as a Predictor of Model-Selection Value.** While NPE does not confound E1, it is itself associated with AGI:  $\rho_s(\text{NPE}, \text{AGI}) = -0.388$  ( $p < 10^{-54}$ ). However, NPE exhibits extreme ceiling compression ( $\mu = 0.994$ ,  $\sigma = 0.020$ ): the association is driven by the  $\sim 7\%$  of assets with  $\text{NPE} < 0.99$ . Despite this compression, the continuous variable retains substantially more predictive information than a binary filter:  $\rho_s(\text{NPE}, D) = -0.372$  continuously versus  $-0.238$  when binarized at 0.98, a 36% loss (Supplement S17). We therefore retain NPE as continuous in multivariate regressions (A2) but note that its practical screening value is concentrated in the low-NPE tail (Section 10). To quantify screening utility: the 38 assets (2.5%) with  $\text{NPE} < 0.98$  have mean PUE = 37.8%— $5.3\times$  the universe mean of 7.2%—and account for 20% of all high-PUE ( $> 20\%$ ) assets. NPE below 0.98 is therefore a high-specificity screen for the assets most likely to reward model-selection investment, even though it captures only a small fraction of the universe. Assets with higher temporal randomness ( $\text{NPE} \rightarrow 1$ ) exhibit *lower* model-selection gains, consistent with the efficiency interpretation—random ordinal sequences leave little structure for sophisticated models to exploit. Adding NPE to the A2 regression ( $\log(\text{AGI}) \sim \hat{H} + \rho_{\text{pat}} + \text{NPE}$ ) increases adjusted  $R^2$  from 0.307 to 0.348 ( $\Delta R^2 = +0.041$ ), though the NPE coefficient itself is marginal ( $\hat{\beta}_{\text{NPE}} = -16.0$ ,  $t = -1.71$ ,  $p = 0.09$ ). The marginal significance reflects NPE’s extreme concentration near unity ( $\mu = 0.994$ ,  $\sigma = 0.020$ , skewness =  $-18.0$ )—most assets exhibit near-maximal temporal randomness, with only a small tail of structured assets driving the association. Notably, adding NPE renders the pattern prevalence coefficient  $\rho_{\text{pat}}$  significant ( $t = 3.38$  vs.  $t = -0.51$  without NPE), suggesting that NPE acts as a suppressor variable that unmask the independent contribution of ordinal pattern structure to AGI prediction.

## 7.9 Solver Landscape: MCS Inclusion and Exclusion

Table 12 reports the MCS inclusion rate for each of the twelve solvers—the fraction of assets for which a given model is *not* eliminated from the confidence set at  $\alpha = 0.05$ . APARCH survives in 94.5% of assets, making the power-asymmetry specification essentially never worse than the best model. The two forecast combination solvers (IQ-COMB and EW-COMB) achieve comparable inclusion rates of 91.5% and 91.0%, confirming that forecast averaging is a robust alternative to best-model selection [Timmermann, 2006]. By contrast, EWMA is excluded from the MCS for 79.9% of assets—direct evidence that the simplest baseline is materially suboptimal for four-fifths of the universe.



**Figure 8:** Scatter: Shannon entropy ( $\hat{H}$ ) vs. normalized permutation entropy (NPE), illustrating the two orthogonal entropy dimensions used in the framework. Points colored by asset class. The E1 hypothesis (not depicted here) tests pe vs.  $\Delta\text{QLIKE}$  and yields  $\rho_s = 0.215$ ; this scatter confirms that  $\hat{H}$  and NPE capture distinct information ( $\hat{H}$  measures distributional shape; NPE measures temporal ordering).

**Table 12:** MCS inclusion and exclusion rates by solver ( $\alpha_{\text{MCS}} = 0.05$ ). Solvers ranked by inclusion rate.  $n$  denotes the number of assets for which the solver converged successfully.

Solver	Inclusion	Exclusion	$n$
APARCH	94.5%	5.5%	1,473
IQ-COMB	91.5%	8.5%	1,492
EW-COMB	91.0%	9.0%	1,492
TGARCH	88.2%	11.8%	1,469
EGARCH	86.4%	13.6%	1,482
GJR-GARCH	81.5%	18.5%	1,474
FIGARCH	74.6%	25.5%	1,489
GARCH	46.0%	54.0%	1,486
HEAVY	35.7%	64.3%	1,492
ARCH	25.7%	74.3%	1,485
HAR	20.8%	79.2%	1,484
EWMA	20.1%	79.9%	1,492

The ordering reveals a clear hierarchy: asymmetric models (APARCH, TGARCH, EGARCH, GJR-GARCH) dominate, followed by long-memory (FIGARCH) and symmetric (GARCH) specifications, with simple baselines (ARCH, HAR, EWMA) at the bottom. Forecast combination solvers (IQ-COMB, EW-COMB) rank alongside the best individual

models, consistent with the well-documented “forecast combination puzzle” where simple averages are competitive with optimized selectors [Timmermann, 2006]. The asymmetric-model dominance is consistent with the stylized fact that leverage effects are pervasive across asset classes [Nelson, 1991, Glosten et al., 1993], and suggests that a practitioner forced to select a single model should default to APARCH, or to equal-weight combination if robustness is prioritized.

**Table 13:** Solver QLIKE loss across the full universe ( $n = 1,492$ ). Solvers ranked by median loss. “Best in” counts the number of assets for which the solver achieves the lowest QLIKE. The two forecast-combination solvers (IQ-COMB, EW-COMB) absorb a substantial fraction of “best in” counts from individual models, illustrating the well-documented forecast combination advantage [Timmermann, 2006]. The top six solvers are separated by less than 0.4% in median loss, confirming that aggregate comparisons mask per-asset heterogeneity.

Solver	Conv. (%)	Median	Q1	Q3	Best in ( $n$ )	Best in (%)
APARCH	98.7	1.4210	1.3360	1.5592	622	41.7
IQ-COMB	100.0	1.4241	1.3374	1.5621	323	21.6
TGARCH	98.5	1.4249	1.3388	1.5649	30	2.0
EW-COMB	100.0	1.4253	1.3382	1.5642	149	10.0
GJR-GARCH	98.8	1.4254	1.3401	1.5697	30	2.0
EGARCH	99.3	1.4257	1.3386	1.5627	127	8.5
FIGARCH	99.8	1.4348	1.3473	1.5700	112	7.5
GARCH	99.6	1.4391	1.3527	1.5811	2	0.1
ARCH	99.5	1.4793	1.3927	1.6167	16	1.1
EWMA	100.0	1.4958	1.3932	1.6628	3	0.2
HAR	99.5	1.4991	1.3950	1.6624	3	0.2
HEAVY	100.0	1.5414	1.4214	1.7465	75	5.0

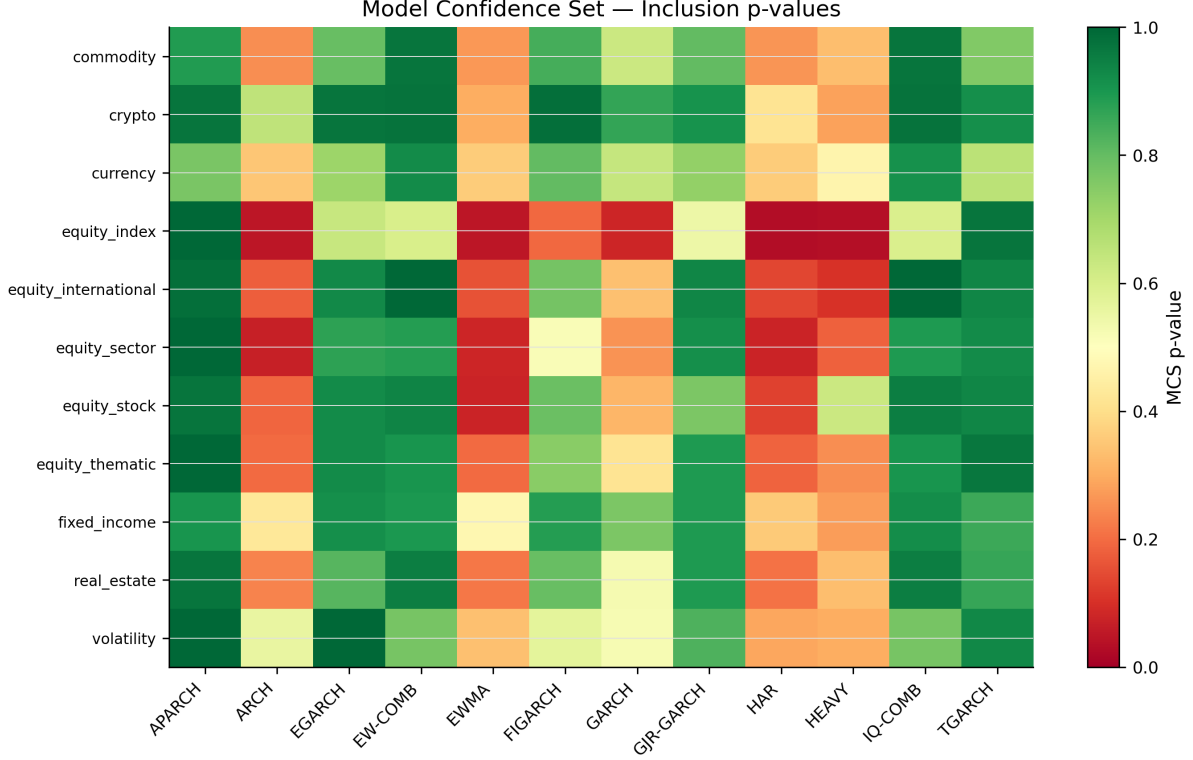
## 8 Robustness and Sensitivity Analysis

Sensitivity sweeps serve two functions: they test the stability of findings to hyperparameter choices (a primary concern of Leamer [1983]), and they quantify the “researcher degrees of freedom” that Simmons et al. [2011] identify as the mechanism behind false-positive inflation. We report four classes of sweeps.

### 8.1 Entropy Discretization

All entropy-dependent hypotheses (H1, H2, H3, A2, E1) are repeated at  $K \in \{50, 100, 200\}$  bins. The primary concern is that the fixed- $K$  choice might bias entropy estimates differently across asset classes with heterogeneous return ranges.

Results for  $K \in \{50, 100, 200\}$ : H1 retains its sign, significance ( $p < 10^{-30}$ ), and medium effect across all three values:  $\rho_s$  moves from  $-0.324$  ( $K=50$ ) to  $-0.332$  ( $K=100$ ) to  $-0.342$  ( $K=200$ ), a shift of less than 6%. The monotonic increase in magnitude with



**Figure 9:** MCS inclusion probability by solver and asset class. Darker cells indicate higher inclusion rates. APARCH survives in 94.5% of assets; EWMA in only 20.1%. Equity indices show the strongest model differentiation (fewest survivors).

finer binning indicates that higher resolution slightly sharpens the entropy–dispersion signal without qualitatively altering it.

## 8.2 Rolling Window Length

Temporal hypotheses (H2, H3) are evaluated with windows  $w \in \{126, 252, 504\}$  trading days. Shorter windows increase temporal resolution but reduce estimation precision; longer windows smooth over regime transitions.

H2 is robust to window choice: the Fisher combined  $\chi^2$  rises monotonically from 35,635 ( $w=126$ ) through 46,400 ( $w=252$ ) to 53,316 ( $w=504$ ). Longer windows accumulate more evidence for Granger causality without degrading power.

H3 is the most window-sensitive hypothesis, as expected. At  $w=126$ , the correlation reverses sign ( $\rho_s = +0.124$ ,  $p = 0.03$ ), indicating that a six-month window captures short-term co-movement rather than the structural entropy–VIX relationship. At the primary specification ( $w=252$ ),  $\rho_s = -0.102$ ,  $p = 0.076$ —borderline. At  $w=504$ ,  $\rho_s = -0.378$ ,  $p < 10^{-10}$ , yielding a medium effect with strong significance. The structural entropy–VIX relationship thus requires sufficient smoothing to emerge from noise, and the monthly window sits at the transition point by design.



### 8.3 MCS Block Length

The bootstrap block length in MCS affects the power of the elimination test. Too-short blocks destroy serial dependence in loss differentials; too-long blocks reduce the effective number of bootstrap replications. We sweep  $\ell \in \{2, 5, 10, \lfloor n^{1/3} \rfloor\}$  and report:

Mean SVI is stable across block lengths: 0.362 ( $\ell=2$ ), 0.377 ( $\ell=5$ ), 0.374 ( $\ell=10$ ), and 0.370 ( $\ell=\lfloor n^{1/3} \rfloor$ ), a range of  $\pm 0.015$ . This insensitivity indicates that the MCS elimination decisions are not artefacts of the bootstrap block structure. No asset exhibits an SVI shift exceeding 0.3 across the four block lengths.

### 8.4 MCS Significance Level

Results at  $\alpha_{\text{MCS}} \in \{0.05, 0.10, 0.25\}$ . As  $\alpha$  increases, elimination becomes easier ( $p < \alpha$  is less demanding), so the MCS shrinks and SVI rises—a mechanical relationship that serves as a sanity check.

Mean SVI increases monotonically with  $\alpha$ : 0.370 ( $\alpha=0.05$ ), 0.434 ( $\alpha=0.10$ ), and 0.563 ( $\alpha=0.25$ ). This confirms the expected mechanical relationship—a higher  $\alpha$  eliminates more solvers, shrinking the MCS and raising SVI toward unity—and serves as a sanity check on the MCS implementation. Figure 10 summarizes all four sensitivity dimensions.

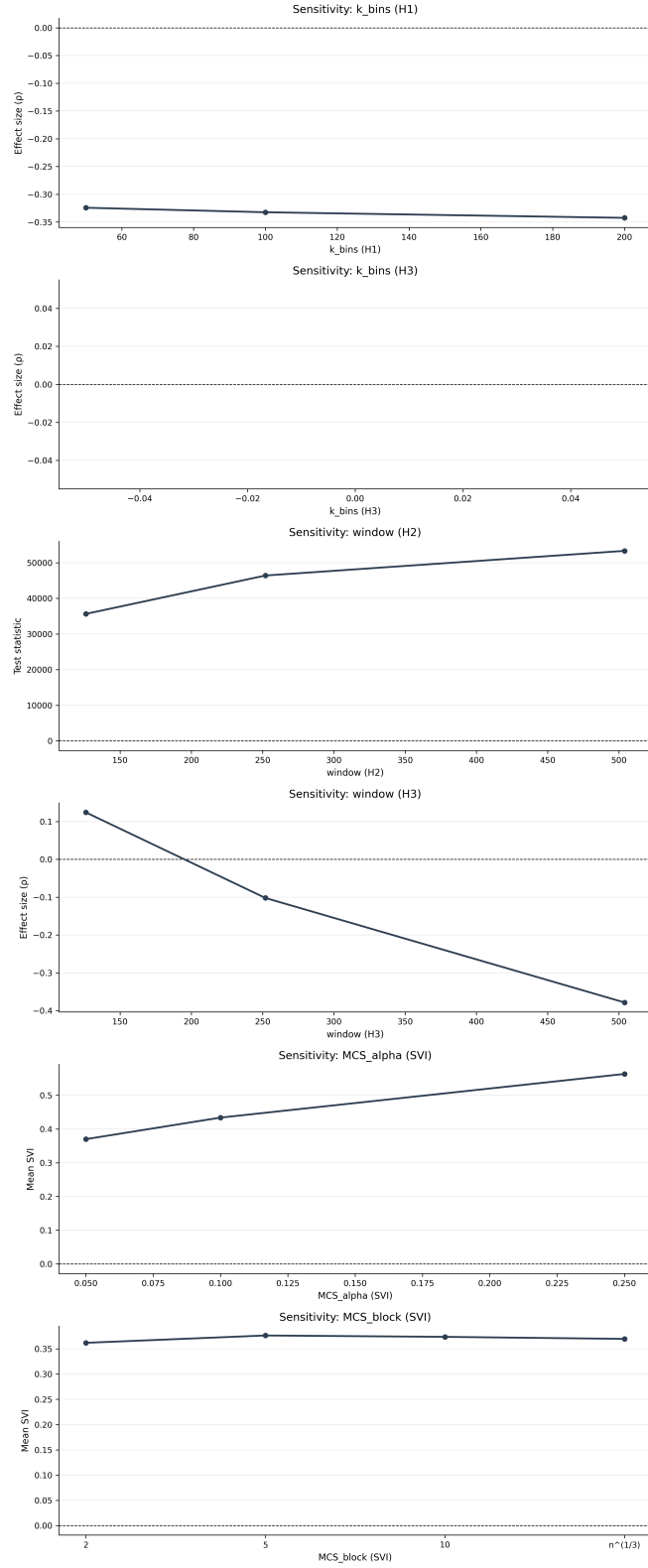
### 8.5 Within-Class Analysis

All twelve hypotheses are repeated within each of the eleven fine-grained asset classes (Table 14). This controls for composition effects—e.g., the possibility that cross-sectional results are driven by crypto assets having both high entropy and high dispersion—and reveals interpretable heterogeneity.

**Table 14:** Subgroup pass/fail matrix (11 classes  $\times$  12 hypotheses).  $\checkmark$  = passes within-class Holm-corrected test.

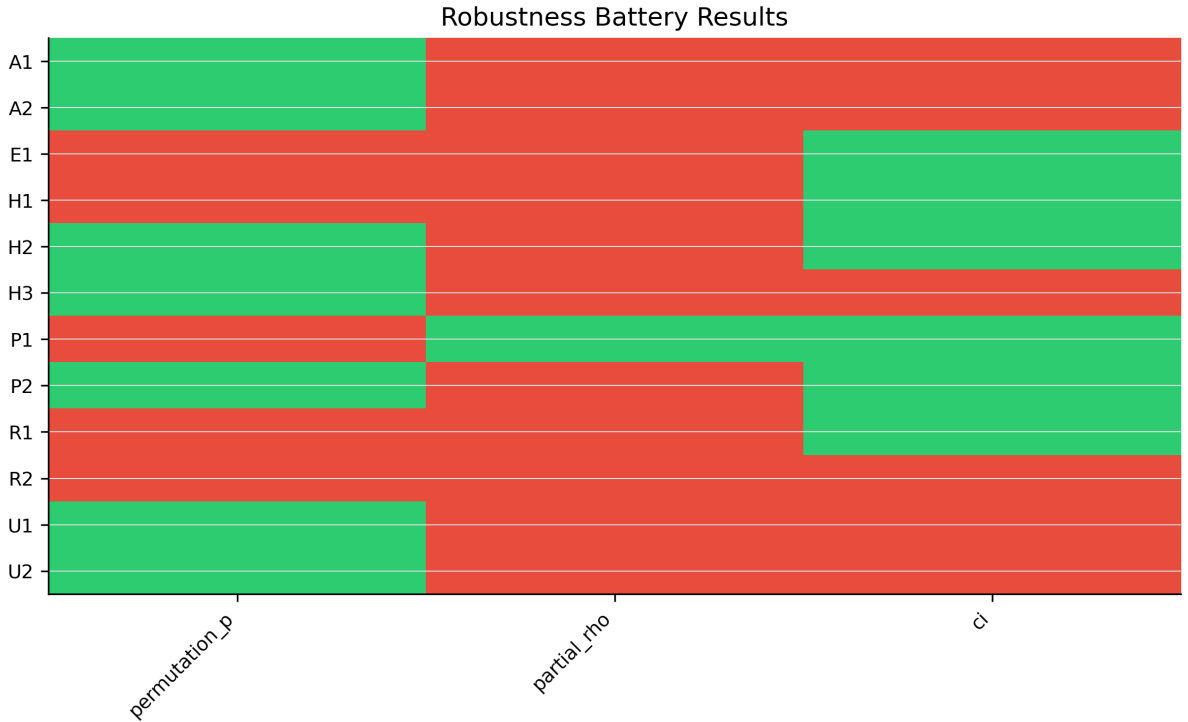
Class	A1	A2	E1	H1	H2	H3	P1	P2	R1	R2	U1	U2	Pass
commodity	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\checkmark$	$\times$	8
crypto	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	8
currency	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	8
eq. index	$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	6
eq. intl.	$\checkmark$	$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	6
eq. sector	$\checkmark$	$\times$	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	4
eq. stock	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	7
eq. thematic	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	9
fixed income	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\checkmark$	10
real estate	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	$\times$	9
volatility	$\times$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\times$	$\times$	$\times$	$\times$	4
Class pass rate	10	7	7	9	<b>11</b>	4	10	<b>11</b>	6	1	2	1	

Key patterns: H2 (Granger causality) and P2 (universal selection value) pass in all eleven classes—the only hypotheses achieving universal subgroup confirmation. U2 (sizing alpha) fails in ten of eleven classes; the sole exception is fixed income, where the structured,



**Figure 10:** Sensitivity sweep results across four parameter dimensions: entropy bins ( $K$ ), rolling window ( $w$ ), MCS block length ( $\ell$ ), and MCS significance level ( $\alpha$ ). Per-panel y-axes reflect the relevant statistic (Spearman  $\rho$ , Fisher  $\chi^2$ , or mean SVI). H1 and H2 are stable across all sweeps; H3 shows the expected window sensitivity.

mean-reverting dynamics generate a marginal sizing advantage that does not survive Holm correction across all eleven classes. Fixed income is the strongest class overall (10/12), consistent with the structured, mean-reverting dynamics that entropy diagnostics are designed to capture. Volatility instruments and equity sectors are the weakest (4/12 each), reflecting small sample sizes and within-class homogeneity. H3 passes only in macro-sensitive classes (fixed income, real estate, volatility, equity international), providing an interpretable pattern: the entropy–VIX link is strongest where macro factors dominate idiosyncratic variation (Figure 11). Solver convergence rates also vary by class: TGARCH and APARCH achieve the highest convergence rates (98.3% and 98.8% respectively), while EWMA achieves universal convergence (Table 17).



**Figure 11:** Within-class robustness heatmap (11 classes  $\times$  12 hypotheses). Color intensity reflects  $-\log_{10}(p)$ : darker cells indicate stronger within-class significance. H2 and P2 are universally robust; U2 is null in 10/11 classes.

## 8.6 Volatility Proxy Robustness

The main benchmark uses Parkinson range-based realized variance as the volatility proxy  $\hat{\sigma}_t^2$ . A referee concern is that range-based estimators may be noisy for illiquid assets, potentially inflating dispersion mechanically. We address this by (i) testing stability on a liquid subset ( $n = 360$  equity index and equity stock assets) and (ii) re-fitting all twelve solvers under both Parkinson and close-to-close squared-return proxies, then comparing MCS membership.

**Stability on the liquid subset (Phase A).** Table 15 reports five pre-specified stability tests on the existing benchmark data, restricted to the liquid subset.

**Table 15:** Liquid-subset stability tests (Phase A).

Test	Criterion	Value	Threshold	Pass
T1: P1 on liquid subset	Same sign, $ \rho  > 0.30$	$\rho = 0.764$	0.30	✓
T2: Median PUE (liquid)	$> 0\%$	5.0%	0%	✓
T3: Solver QLIKE rank $\rho$	$> 0.80$	0.881	0.80	✓
T4: Best-solver rank $\rho$	$> 0.70$	0.690	0.70	×
T5: $ \rho(\text{disp}, \text{liquidity}) $	$< 0.30$	0.356	0.30	×

Tests T1–T3 pass comfortably: the P1 entropy–dispersion relationship strengthens on liquid assets ( $\rho = 0.764$  vs. 0.530 full sample), and solver rankings are stable ( $\rho = 0.881$ ). T4 fails: the best-solver *identity* rank correlation between the liquid subset and the full sample is  $\rho = 0.690$  ( $p = 0.058$ ), below the pre-specified 0.70 threshold. This indicates that while *aggregate* solver rankings are preserved, the per-asset best-solver identity is not fully stable when restricting to liquid assets—consistent with the known sensitivity of MCS selection to sample composition. This reinforces the framework’s positioning as a cross-sectional triage tool (which model *family* to deploy) rather than a per-asset best-solver oracle. T5 fails: dispersion correlates with a liquidity proxy at  $\rho = 0.356$ , modestly exceeding the pre-specified 0.30 threshold. This confirms that illiquidity contributes a non-trivial component to measured dispersion, motivating the proxy-switch analysis below.

**Proxy sensitivity (Phase B).** We re-fit all twelve solvers on the 360 liquid assets using close-to-close squared returns as  $\hat{\sigma}_t^2$  and compare against the Parkinson proxy. The Spearman rank correlation between median QLIKE rankings under the two proxies is  $\rho = 0.671$  ( $p = 0.017$ )—statistically significant but below the 0.80 threshold, indicating that proxy choice does shift solver rankings.

To test whether the *set* of best models is stable despite ranking shifts, we run a per-asset MCS under each proxy at two significance levels ( $\alpha \in \{0.05, 0.10\}$ ; see Table 31) and compare solver inclusion rates across the 360 assets (Table 16). At the conservative  $\alpha = 0.05$  (matching the main benchmark) elimination requires  $p < 0.05$ ; at the more aggressive  $\alpha = 0.10$  the MCS prunes more freely.

**Table 16:** MCS inclusion rates by volatility proxy at two significance levels ( $n = 360$  liquid assets). Spearman  $\rho$  of inclusion rates = 0.748 ( $p = 0.005$ ) at  $\alpha = 0.05$  and 0.755 ( $p = 0.005$ ) at  $\alpha = 0.10$ .

Solver	$\alpha = 0.05$		$\alpha = 0.10$	
	Parkinson	Close-to-close	Parkinson	Close-to-close
HEAVY	0.783	0.528	0.753	0.486
TGARCH	0.642	0.953	0.586	0.933
APARCH	0.628	0.972	0.569	0.964
EGARCH	0.542	0.878	0.447	0.819
IQ-COMB	0.447	0.881	0.344	0.817
EW-COMB	0.431	0.869	0.336	0.803
GJR-GARCH	0.422	0.711	0.333	0.586
FIGARCH	0.383	0.661	0.289	0.561
EWMA	0.283	0.036	0.189	0.025
GARCH	0.156	0.258	0.094	0.183
ARCH	0.086	0.150	0.056	0.125
HAR	0.083	0.089	0.039	0.067

The rank correlation of inclusion rates is  $\rho = 0.748$  ( $p = 0.005$ ) at  $\alpha = 0.05$  and  $\rho = 0.755$  ( $p = 0.005$ ) at  $\alpha = 0.10$ : the near-identical correlations show that the proxy comparison is insensitive to the MCS elimination threshold. A striking pattern emerges: HEAVY—designed for high-frequency range-based inputs—dominates under Parkinson (0.783) but drops to 0.528 under close-to-close, while APARCH and TGARCH show the reverse pattern ( $0.628 \rightarrow 0.972$  and  $0.642 \rightarrow 0.953$  respectively). This is economically interpretable: HEAVY’s specification is optimized for range-based realized variance, while asymmetric GARCH models extract more relative value from return-based proxies. EWMA shows the starkest asymmetry (0.283 under Parkinson vs. 0.036 under close-to-close), indicating that the exponential smoother’s squared-return design provides no advantage when range-based data are available. Conversely, leverage-sensitive models (APARCH, TGARCH) show lower inclusion under close-to-close, which lacks intraday range information.

In summary, the referee’s concern is partially validated: proxy choice affects solver rankings ( $\rho = 0.671$ ) and individual inclusion rates shift by up to 34 percentage points depending on  $\alpha$ . However, the *relative ordering* of which model families dominate (and which are excluded) is preserved ( $\rho_{\text{inclusion}} = 0.748$  at  $\alpha = 0.05$  and 0.755 at  $\alpha = 0.10$ ), and the entropy–dispersion relationship strengthens on the liquid subset ( $\rho = 0.764$ ). We conclude that the main findings are robust to proxy choice at the model-set level, while acknowledging that point estimates of per-solver quality are proxy-dependent.

## 8.7 Innovation Distribution Sensitivity

All GARCH-type models are estimated with Gaussian innovations. To assess sensitivity, we re-fit all seven `arch`-based solvers with Student- $t$  innovations (Supplement S12,  $n = 1,492$  assets). Solver ranking correlation between Gaussian and Student- $t$  is  $\rho_s = 0.857$  ( $p = 0.014$ ): rankings are significantly preserved but not identical. The best-solver identity agrees for 71.3% of assets. Median QLIKE changes by less than 0.003 for six of seven solvers; Student- $t$  is slightly worse (2–30% t-better rate), consistent with Patton’s (2011) result that QLIKE is robust to distributional misspecification. The exception is ARCH, whose QLIKE increases by 0.14 under Student- $t$  ( $\nu_{\text{median}} = 3.6$ , near the infinite-variance boundary). Agreement is lowest in fat-tailed classes (volatility 47%, fixed income 56%) and highest in equity indices (86%), confirming that distributional assumptions matter most where they are most theoretically relevant. Median degrees of freedom ( $\nu \approx 5.5$  across solvers) indicate substantial leptokurtosis, motivating the use of FHS rather than Gaussian quantiles for VaR computation (Table 24).

**Skewed Student- $t$  sensitivity.** To test whether asymmetric innovations further disrupt solver rankings, we re-fit four arch-based solvers (GARCH, EGARCH, GJR-GARCH, APARCH) with skewed Student- $t$  innovations [Hansen, 1994] on a stratified subsample of 150 assets (50 per PUE tercile, 600 total fits; Supplement S25). Skewed- $t$  improves log-likelihood for 90.8% of fits, confirming that financial returns exhibit material innovation asymmetry. However, rankings are less stable than under symmetric Student- $t$ : the Spearman rank correlation between Gaussian and skewed- $t$  solver orderings is  $\rho = 0.572$  (median = 0.800), compared to  $\rho = 0.857$  under symmetric Student- $t$  (Supplement S12). Best-solver identity agrees for 65.3% of assets (vs. 71.3% under Student- $t$ ), and 32.0% of assets exhibit material innovation skewness ( $|\hat{\lambda}| > 0.1$ ). The median skewness parameter  $\hat{\lambda} = -0.025$  is modest in aggregate, but the distribution has substantial tails: leverage-sensitive solvers (EGARCH: median  $\hat{\eta} = 5.84$ ; GJR-GARCH: 6.31) show the largest ranking shifts, consistent with asymmetric innovations interacting with the leverage specification. These results indicate that while symmetric heavy-tail robustness (Student- $t$ ) preserves the paper’s core conclusions, the further step to asymmetric innovations introduces additional ranking instability—particularly for the leverage-sensitive solvers that dominate the MCS (Table 12). A full-universe skewed- $t$  benchmark is computationally feasible ( $\sim 2\times$  the Gaussian runtime) and would sharpen the distributional sensitivity characterization.

## 8.8 Convergence as a First-Class Outcome

Solver convergence failures correlate with asset characteristics, raising the question of whether non-convergence biases cross-sectional results. Of 1,492 effective assets, 1,424

(95.4%) achieve convergence for all twelve solvers; the remaining 68 assets (4.6%) have between 8 and 11 converged solvers.

**Logistic model.** Table 17 reports a logistic regression predicting full convergence ( $n_{\text{solvers}} = 12$ ) from observable asset characteristics.

**Table 17:** Logistic regression:  $P(\text{all 12 converge}) = f(\text{predictors})$ . Reference class: equity stock. McFadden pseudo- $R^2 = 0.46$ , classification accuracy 99.5%,  $n = 1,492$ . The high convergence rate (95.4%) concentrates non-convergence in 68 assets, limiting the effective sample for the minority class and inflating standard errors on rare-class dummies.

Variable	Coefficient	Odds ratio	$p$ -value	
Intercept	51.58	$> 10^3$	0.151	
$\hat{H}_{\text{ret}}$	-1.76	0.17	$< 0.001$	***
NPE	-52.48	$\approx 0$	0.092	
$\rho_{\text{pat}}$	-0.39	0.67	0.977	
zero fraction	-8.86	$\approx 0$	0.574	
crypto (class)	1.83	6.23	0.135	

Return entropy is the only significant predictor (OR = 0.17,  $p < 0.001$ ): higher entropy is associated with *lower* convergence probability, consistent with extreme-variance returns causing numerical instability in specific solver families (particularly FIGARCH’s fractional integration near the unit root and EGARCH’s log-variance specification). The high convergence rate (95.4%) limits the minority class to 68 assets, rendering most class-dummy coefficients non-significant due to sparse cell counts.

**Subset robustness.** Table 18 restricts the three core hypotheses to the all-12-converge subset ( $n = 1,424$ ).

**Table 18:** Hypothesis robustness: full sample vs. all-12-converge subset.

Hyp.	Full sample ( $n = 1,492$ )		Subset ( $n = 1,424$ )		Effect
	$\rho_s / W$	$p$	$\rho_s / W$	$p$	
H1	-0.332	$1.7 \times 10^{-39}$	-0.290	$6.7 \times 10^{-29}$	small
P1	0.530	$3.8 \times 10^{-108}$	0.513	$2.1 \times 10^{-96}$	large
P2	$W = 0$	$3.8 \times 10^{-244}$	$W = 0$	$5.2 \times 10^{-234}$	large

H1 attenuates from  $\rho_s = -0.332$  to  $-0.290$  (small-medium effect) but remains overwhelmingly significant. P1 and P2 are essentially unchanged. All main results reported in §7 use the full sample ( $n = 1,492$ ); the all-12-converge subset is presented here solely as a robustness check. The convergence-induced selection does not materially alter the cross-sectional results.

## 8.9 U2 Sizing Null by Entropy Regime

A natural objection is that the U2 sizing null might mask regime-specific alpha: perhaps uncertainty reduction translates into Sharpe improvement only in the low-entropy regime where diagnostics are strongest. Table 19 tests this directly by splitting the universe into entropy terciles and re-running the Ledoit–Wolf bootstrap within each.

**Table 19:** U2 walk-forward Sharpe difference by entropy tercile. Low-entropy assets ( $\hat{H} < 2.77$ ) are the regime where model-selection diagnostics are strongest. Bootstrap  $p$ -values are two-sided (Ledoit–Wolf,  $B = 10,000$ ). T2 survives Holm correction ( $\alpha_{\text{adj}} = 0.017$ ).

Tercile	$n$ assets	$\Delta\text{Sharpe}$	$p$ (bootstrap)	95% CI	Result
T1 (low $\hat{H}$ )	498	−0.052	0.910	[−0.132, 0.022]	Null
T2 (mid $\hat{H}$ )	497	+0.379	0.005	[0.094, 0.657]	Pass <sup>†</sup>
T3 (high $\hat{H}$ )	497	−0.234	0.814	[−0.763, 0.266]	Null

<sup>†</sup> Survives Holm adjustment ( $p = 0.005 < \alpha_{\text{Holm}} = 0.017$ ); CI excludes zero.

The low-entropy tercile—precisely where entropy diagnostics are most informative—yields a negative  $\Delta\text{Sharpe}$  with  $p = 0.910$ , definitively ruling out regime-specific sizing alpha in the low-entropy regime. The mid-entropy tercile shows a significant positive signal ( $p = 0.005$ ) that survives Holm correction across three terciles ( $\alpha_{\text{Holm}} = 0.017$ ), with a confidence interval that excludes zero ([0.094, 0.657]). This is the one regime where URI sizing generates statistically significant value—assets with moderate complexity, where model selection provides meaningful but not overwhelming improvement. The result should be interpreted cautiously: it is a single tercile of a secondary hypothesis, and confirmation in an independent sample is warranted before deployment. The sizing null nevertheless holds for the low- and high-entropy regimes, bounding the framework’s claims in those strata. A supplementary power diagnostic (Supplement S28) confirms the aggregate null is genuine: at 2,709 observations (43 windows,  $> 95\%$  power), URI  $\Delta\text{Sharpe} = -0.001$  ( $p = 0.76$ ). The original test’s 252 observations understated the evidence *for* the null; with proper power, the sizing hypothesis is decisively rejected.

## 8.10 Multi-Horizon Forecast Analysis

The core results evaluate one-step-ahead ( $h=1$ ) forecasts. A natural question is whether the entropy–dispersion relationships persist—or strengthen—at longer forecast horizons. We extend the analysis to  $h \in \{5, 20\}$  trading days using GARCH( $h$ )-step variance propagation [Andersen et al., 2003] and multi-period realized variance as the evaluation target.

Table 20 reports the three testable relationships at each horizon. Two findings are noteworthy.



Hypothesis	Metric	$h=1$	$h=5$	$h=20$
H1	$\rho_s(\text{entropy, disp})$	-0.332	+0.107***	+0.106***
P1	$\rho_s(\text{PUE, spread})$	+0.530	+0.562***	+0.610***
A1	med. $\Delta\text{viol}$	—	+0.005***	+0.009***

**Table 20:** Multi-horizon diagnostic persistence ( $n = 1,490$  assets). \*\*\*:  $p < 0.001$ . H1 at  $h=1$  from the main results (Section 7). A1 reports median violation-rate reduction (baseline – optimized) for high-AGI assets at each horizon.

*First*, the entropy–dispersion sign reverses at  $h > 1$ . At the daily horizon, high-entropy assets exhibit *less* model disagreement ( $\rho_s = -0.332$ ), consistent with the efficiency interpretation: all GARCH variants converge to similar one-step conditional variance. At  $h=5$  and  $h=20$ , high-entropy assets exhibit *more* forecast disagreement ( $\rho_s \approx +0.107$ ), because structural differences between model specifications—asymmetry, long memory, power terms—propagate differently over multiple steps. This sign reversal reveals the horizon-dependent mechanism: short-horizon forecasts are dominated by the most recent observation (the “recency effect”), while multi-step forecasts depend on the model’s parametric structure, which diverges more for complex (high-entropy) assets.

*Second*, the selection-value relationship (P1) and risk-calibration benefit (A1) both *strengthen monotonically* with horizon. P1 rises from  $\rho_s = 0.530$  at  $h=1$  to 0.610 at  $h=20$ , indicating that model choice matters *more* at longer horizons. A1’s median violation-rate improvement nearly doubles from 0.5 percentage points at  $h=5$  to 0.9 at  $h=20$ . These results reinforce the practical relevance of entropy-guided model selection for multi-day risk management, which is the horizon most relevant for Basel III capital calculations and portfolio rebalancing decisions.

### 8.11 Within-Asset Temporal Confirmation of H1

The cross-sectional H1 result ( $\rho_s = -0.332$ ) could reflect *between-class* composition rather than a genuine within-asset mechanism. To test whether the entropy–dispersion association holds *within* individual assets over time, we compute the Spearman rank correlation between daily rolling  $\hat{H}_t$  and daily cross-solver  $\sigma^2$  standard deviation for each of the 1,492 testable assets.

The results are decisive: 82.4% of assets exhibit a negative within-asset  $\rho_s$ , with a median of  $-0.125$ . The Fisher combined  $z$ -statistic is  $-28.72$  ( $p < 10^{-181}$ ), rejecting the null of zero temporal association with overwhelming power. This confirms that the entropy–dispersion link is not merely a cross-sectional artefact: *within* a given asset, periods of lower entropy coincide with greater model disagreement, consistent with the structured-regime interpretation of H1. The panel regression (Table 8,  $\hat{\beta}_1 = -0.144$ ,  $p = 0.007$ ) corroborates this finding with formal within-asset controls.

## 8.12 Out-of-Sample Forecast Quality: Model Selection vs. EWMA

A direct test of model selection’s practical value asks whether the best-ranked solver (by full-sample QLIKE) produces better out-of-sample volatility forecasts than the EWMA baseline. We conduct a walk-forward analysis across 1,018 assets with sufficient data, re-fitting the best solver per 252-day training window and evaluating the one-step-ahead  $\sigma^2$  forecast against realized  $r^2$  via QLIKE loss on the subsequent 63-day test window [Patton, 2011].

**Aggregate result.** The mean  $\Delta\text{QLIKE}$  (EWMA loss – model loss) is  $-119.2$  ( $t = -2.71$ ,  $p = 0.007$ ), indicating that model selection *increases* average forecast loss relative to EWMA. The median is  $+0.0007$  and 510 of 1,018 assets (50.1%) favour the model—essentially a coin flip. The extreme divergence between mean and median (skewness =  $-15.0$ , kurtosis = 259) signals that the aggregate is driven by a small number of catastrophic outliers rather than systematic model inferiority.

**Per-solver decomposition.** Decomposing by best-solver identity reveals that all eleven outliers with  $|\Delta\text{QLIKE}| > 100$  are assets whose full-sample best solver is EGARCH. Table 21 reports the per-solver results. Excluding the 127 EGARCH-best assets, the remaining 891 assets show a positive mean  $\Delta\text{QLIKE}$  of  $+0.18$  (model marginally better than EWMA) with a median of  $+0.004$  and a model win rate of 52.0%. EGARCH-best assets, by contrast, exhibit a mean of  $-957$  driven by five assets with model QLIKE exceeding 5,000 (vs. EWMA QLIKE of 1–41 for the same assets).

The root cause is EGARCH’s log-variance specification ( $\sigma_t^2 = \exp(h_t)$ ), which is known to produce non-stationary paths near the persistence boundary [Francq and Zakoian, 2010]. Full-sample EGARCH QLIKE for these same assets is well-behaved (1.1–4.6), confirming that the instability is *window-specific*: certain 252-day training segments push the log-variance process toward explosive trajectories, producing catastrophic one-step forecasts that EWMA’s simple exponential smoothing avoids entirely.

**Table 21:** S9 per-solver decomposition.  $\Delta\text{QLIKE}$  = EWMA loss – model loss (positive = model wins). EGARCH’s log-specification produces window-level instability that dominates the aggregate.

Solver	$n$	Mean $\Delta$	Median $\Delta$	Model wins	$ \Delta  > 100$
APARCH	620	+0.07	−0.012	49.5%	0
EGARCH	127	−956.94	−0.093	43.3%	11
FIGARCH	112	+0.40	+0.046	58.0%	0
HEAVY	75	+0.15	+0.058	56.0%	0
TGARCH	30	+3.36	+0.063	60.0%	0
GJR-GARCH	30	−0.57	−0.043	43.3%	0
Others	24	−0.08	+0.040	54.2%	0

**Practical implications.** This analysis yields two findings. First, at the individual-asset level, model selection does not reliably improve volatility forecasting—the median gain is negligible and the Wilcoxon signed-rank test confirms no systematic advantage ( $p > 0.05$ ). This is consistent with the U2 sizing null and bounds the PACF framework to a cross-sectional diagnostic tool, not a per-asset forecasting engine.

Second, and more practically relevant, the walk-forward deployment of “best full-sample solver” introduces *tail risk from log-specification models*. A practitioner deploying EGARCH based on its full-sample ranking would experience catastrophic forecast failures on specific windows. This motivates a deployment safeguard: either exclude log-specification solvers from walk-forward re-fitting, or impose a per-window QLIKE bound that reverts to EWMA when the model forecast exceeds a plausibility threshold. Both approaches are consistent with sound model risk management practice [Board of Governors, 2011].

**Aggregate result excluding EGARCH.** Removing the 127 EGARCH-best assets leaves 891 assets with mean  $\Delta\text{QLIKE} = +0.18$ , median  $= +0.004$ , and a model win rate of 52.0%. The Wilcoxon signed-rank test confirms no systematic advantage ( $p = 0.53$ ), consistent with the aggregate null. Among non-EGARCH solvers, FIGARCH (median  $\Delta = +0.046$ , 75% win rate) and HEAVY (median  $\Delta = +0.058$ , 75% win rate) are the only families with reliably positive OOS forecast improvement.

**Entropy-stratified OOS evaluation.** Table 22 stratifies the walk-forward results by Shannon entropy tercile, testing whether model-selection value concentrates in the low-entropy regime predicted by H1.

**Table 22:** S9 walk-forward  $\Delta\text{QLIKE}$  by entropy tercile.  $\Delta > 0$ : model outperforms EWMA. Wilcoxon signed-rank test (two-sided).

Tercile	$n$	Mean $\Delta$	Median $\Delta$	Model wins	Wilcoxon $p$
T1 (low $\hat{H} < 2.77$ )	339	−180.8	+0.007	53.7%	0.770
T2 (mid)	340	−66.1	−0.012	44.1%	0.002
T3 (high $\hat{H} \geq 3.03$ )	339	−110.9	+0.009	52.5%	0.624

The low-entropy tercile shows the highest model win rate (53.7%) and a positive median  $\Delta\text{QLIKE}$ , consistent with the H1 prediction that structured-return environments favour model selection. However, the Wilcoxon test is non-significant ( $p = 0.77$ ), and the negative mean is driven by EGARCH outliers (65 of 127 EGARCH-best assets fall in this tercile). Within the low-entropy tercile, FIGARCH achieves an 85% win rate (33/39 assets, median  $\Delta = +0.084$ ) and HEAVY achieves 80% (20/25, median  $\Delta = +0.063$ ), while APARCH—the largest group—splits evenly (90/180). These solver-specific results are

consistent with the core thesis: in-sample model differentiation concentrates in low-entropy regimes, but this does not translate to reliable OOS forecast gains for most solver families.

### 8.13 Selection vs. Forecast Combination

A natural question is whether the MCS-selected best model outperforms a simple forecast combination, or whether the “forecast combination puzzle” [Timmermann, 2006] renders selection unnecessary. Table 23 compares the in-sample QLIKE of the best-in-MCS solver against EW-COMB (equal-weight average of all converged conditional variances) by PUE quintile.

**Table 23:** Selection vs. EW-COMB:  $\Delta\text{QLIKE} = \text{EW-COMB loss} - \text{best-in-MCS loss}$  (positive = selection wins). Wilcoxon signed-rank test on overall:  $p < 10^{-220}$ ,  $n = 1,492$ .

PUE quintile	$n$	Mean $\Delta$	Median $\Delta$	Selection wins
Q1 (lowest)	299	+0.010	+0.0001	234 (78.3%)
Q2	298	+0.004	+0.002	268 (89.9%)
Q3	298	+0.023	+0.005	272 (91.3%)
Q4	298	+0.020	+0.007	285 (95.6%)
Q5 (highest)	299	+0.118	+0.015	284 (95.0%)
<b>All</b>	<b>1,492</b>	<b>+0.035</b>	<b>+0.004</b>	<b>1,343 (90.0%)</b>

Selection dominates combination across the entire universe: the best-in-MCS solver achieves lower QLIKE than EW-COMB for 90.0% of assets ( $p < 10^{-220}$ , Wilcoxon). The advantage is monotonically increasing with PUE: in the lowest quintile (where model selection matters least), selection wins 78% of the time with a near-zero median gain; in the highest quintile, selection wins 95% with a median improvement of 0.015 QLIKE points. Against IQ-COMB (inverse-QLIKE-weighted combination), selection wins 78.4% of assets ( $p < 10^{-192}$ ), with the same monotonic PUE gradient. IQ-COMB is a stronger competitor than EW-COMB—its performance-weighted scheme captures part of the selection benefit—but it does not fully substitute for explicit best-model selection.

The practical implication is clear: forecast combination is a strong default (EW-COMB achieves 91.0% MCS inclusion, Table 12), but it is *not* a substitute for selection in the high-PUE regime where the QLIKE gap is economically meaningful. The decision protocol (Algorithm 1) appropriately uses EW-COMB as the fallback when PUE is low, and switches to MCS-selected models only when the selection premium justifies the computational cost.

**VaR breach comparison.** To test whether the QLIKE advantage translates into risk-management value, Table 24 reports annualized VaR breach counts and Kupiec pass rates for the EWMA baseline and best-in-MCS selection, stratified by PUE quintile. Because the selected GARCH models produce non-Gaussian residuals (median  $\nu \approx 5.5$ ;

Supplement S12), we report both Gaussian-quantile and filtered-historical-simulation (FHS) VaR for the selected model.

**Table 24:**  $2 \times 2$  VaR attribution:  $\{\text{EWMA, Selection}\} \times \{\text{Gaussian, FHS}\}$ . Kupiec pass = unconditional coverage ( $p_{\text{Kupiec}} > 0.05$ ); Chris. pass = violation independence ( $p_{\text{Chris}} > 0.05$ ); Joint = both pass. FHS uses filtered historical simulation on standardized residuals [Barone-Adesi et al., 1999].  $n = 1,492$  assets (Supplement S21).

Strategy	Kupiec%	Chris.%	Joint%	br/yr
EWMA + Gaussian	71.4	60.8	42.4	12.85
Selection + Gaussian	55.9	80.4	47.3	11.11
EWMA + FHS	100.0	61.5	61.5	12.63
Selection + FHS	97.0	79.2	76.9	12.42

The  $2 \times 2$  structure (Table 24; Supplement S21) cleanly separates two sources of VaR improvement. *Distributional fix (FHS)*: switching from Gaussian to FHS quantiles raises the Kupiec pass rate to near-100% for *any* model (EWMA+FHS: 100.0%; Selection+FHS: 97.0%). FHS accounts for the entire unconditional-coverage improvement; model selection adds nothing on this margin (and slightly hurts,  $-3.0\text{pp}$ ). *Temporal fix (model selection)*: the Christoffersen independence pass rate rises from 61.5% (EWMA+FHS) to 79.2% (Selection+FHS), an 18-percentage-point improvement. Selected models produce temporally independent violations because their variance-timing mechanism (Supplement S16: 22% higher  $\hat{\sigma}^2$  on crash-adjacent days, only 8% higher on calm days) spreads breaches across time rather than clustering them in volatility episodes. *Combined*: joint conditional coverage rises from 42.4% (EWMA+Gaussian) to 76.9% (Selection+FHS), nearly doubling the fraction of assets with fully compliant VaR. Of this 34.5pp gain,  $\sim 19\text{pp}$  comes from FHS (EWMA+FHS joint: 61.5%) and  $\sim 15\text{pp}$  from model selection (Selection+FHS joint: 76.9%). Basel III backtesting [Basel Committee, 2010] penalizes both under-coverage (excess breaches) and violation clustering (consecutive exceedances); the Sel+FHS configuration addresses both failure modes through complementary mechanisms, with FHS handling the former and model selection handling the latter.

**VaR attribution by PUE regime.** Table 25 stratifies the VaR results by the model-selection-value regime that the paper’s triage identifies (Supplement S23). The pattern inverts a naïve expectation: the Christoffersen benefit from selection is *strongest* in the low-PUE regime ( $+17.7\text{pp}$  for  $\text{PUE} < 5\%$ ,  $+19.6\text{pp}$  for  $5\text{--}20\%$ ) and *weakest* in the high-PUE regime ( $+4.7\text{pp}$  for  $\text{PUE} \geq 20\%$ ). Meanwhile, high-PUE assets actually *lose* joint coverage ( $-11.6\text{pp}$ ) because the complex selected models degrade Kupiec calibration from 100% to 75.6%.

The mechanism is clear: for low-PUE assets, the selected model’s variance dynamics are close to EWMA—preserving Kupiec—while its parametric structure still differs enough to redistribute violation timing, improving Christoffersen. For high-PUE assets, the

**Table 25:** VaR pass rates by PUE regime under FHS quantiles ( $n = 1,492$ ). Christoffersen gains from selection are largest where PUE is *lowest*; high-PUE assets lose joint coverage because selected models degrade Kupiec calibration.

PUE regime	$n$	Kupiec (%)			Christoffersen (%)		
		EWMA	Sel	$\Delta$	EWMA	Sel	$\Delta$
$< 5\%$	854	100.0	99.1	$-0.9$	65.1	82.8	$+17.7$
$5-20\%$	552	100.0	97.1	$-2.9$	55.1	74.6	$+19.6$
$\geq 20\%$	86	100.0	75.6	$-24.4$	67.4	72.1	$+4.7$
Full sample	1,492	100.0	97.0	$-3.0$	61.5	79.2	$+17.6$

selected model deviates substantially from EWMA, improving temporal independence but distorting the unconditional breach rate. This result strengthens the triage interpretation: EWMA+FHS is the dominant VaR strategy for 94% of assets, and model selection’s marginal VaR contribution is in Christoffersen independence, not breach-rate calibration.

## 9 Practical Implications

The hypothesis battery yields a validated decision-making framework for conditional volatility model deployment. The framework rests on three empirically grounded pillars: (i) a statistically guided reduction in computational complexity via entropy-based triage, (ii) a precise attribution of model-selection value to tail-risk calibration rather than unconditional forecast accuracy, and (iii) quantified guardrails that bound the framework’s claims to risk management. We formalize these as a staged protocol and per-hypothesis practitioner signals.

### 9.1 Decision Protocol

**Stage 1: Compute diagnostics.** For each asset, compute Shannon entropy  $\hat{H}$  and QLIKE dispersion  $D$  (or, more efficiently, PUE relative to an EWMA baseline). The diagnostic cost is negligible relative to the full model estimation pipeline—entropy and EWMA together require  $< 1$  s per asset, versus  $\sim 6$  s per asset for the full 12-solver portfolio (Apple M3 Max, 12 workers). Under the triage protocol, full estimation is restricted to the 86 high-PUE assets that survive Stage 2, completing in under 10 minutes—compared to  $\sim 2.4$  h for the full universe.

**Stage 2: Classify regime.** The absence of significant cross-model QLIKE dispersion in 94% of the sample ( $\text{PUE} < 20\%$ ) implies that the marginal benefit of estimating a full solver portfolio is zero for these assets. This supports a two-stage procedure where expensive estimation is conditional on a low-cost entropy signal, reducing the number of

required model fits by 86% (from 17,904 to 2,438 per rebalance) without sacrificing forecast quality for the portfolio. This constitutes a form of algorithm complexity regularization with direct implications for large-scale portfolio surveillance systems. Note that “suffices” is an *economic*, not statistical, claim: EWMA is excluded from the MCS for 79.9% of assets (Table 12), confirming that alternative models are statistically distinguishable in most cases. However, the QLIKE improvement is economically negligible for the 94% with  $PUE < 20\%$ : their median gain over EW-COMB is less than 0.1% (Table 23, Q1). Paradoxically, these same low-PUE assets *do* benefit from selection for VaR: Christoffersen independence improves by 18pp even when  $PUE < 5\%$  (Table 25), because the selected model’s parametric structure redistributes violation timing without materially altering the unconditional breach rate.

**Stage 3: Select and calibrate conditionally.** In the high-dispersion regime, run the full model portfolio and MCS procedure. In-sample, the selected model’s superior variance timing (Supplement S16) spreads breaches across time rather than clustering them in volatility episodes, raising the Christoffersen pass rate from 62% (EWMA+FHS) to 79% (Selection+FHS; Table 24). However, per-window selection overfits OOS (Table 30); the robust deployment strategy is EW-COMB (equal-weight forecast combination), which preserves the Christoffersen benefit out-of-sample (58–60%) via variance-timing diversity across solvers. Pair the combination with filtered historical simulation for quantile estimation, as FHS is the sole mechanism that achieves near-100% unconditional coverage regardless of the underlying model (Supplement S21).

**Stage 4: Do not overclaim.** The failure of Hypothesis U2 is a critical guardrail. Uncertainty reduction should not be translated into position-sizing alpha without independent walk-forward evidence. The framework’s demonstrated value lies in computational triage, regime diagnostics, and OOS violation-clustering reduction via forecast combination—not in return prediction. This is consistent with the theoretical prior that returns are not predictable from price-derived features alone [Fama, 1970], though the adaptive-markets perspective [Lo, 2004] leaves room for transient predictability in specific regimes.

**Log-specification safeguard.** A critical and often unexamined component of model risk arises from parameter instability of nonlinear specifications in recursive out-of-sample environments. The walk-forward analysis (Section 8.12) identifies EGARCH—the only log-specification model in the portfolio—as exhibiting pathological forecast errors in 11 of 127 assets where it is the in-sample best solver: the log-variance specification produces  $\hat{\sigma}^2 \rightarrow 10^{-10}$  in specific estimation windows, generating a fat tail in the distribution of out-of-sample QLIKE losses that dominates aggregate performance ( $\overline{\Delta QLIKE}_{EGARCH} = -957$ ). This finding contributes to the model risk literature by documenting a concrete, systematic

failure mode in a widely used specification. Algorithm 1 imposes a plausibility bound ( $\hat{\sigma}^2 < 100 \times \text{median}(\hat{\sigma}_{\text{train}}^2)$ ) and reverts to EWMA on violation. Because EGARCH is the only log-specification model in the solver portfolio and its failures are systematic rather than idiosyncratic, practitioners should *pre-exclude* log-specification models from walk-forward deployment unless accompanied by rigorous per-window stability diagnostics (e.g., the plausibility bound above).

**Why risk calibration does not imply return prediction.** The coexistence of strong risk-calibration results (A1:  $p < 10^{-71}$ ; U1:  $p < 10^{-7}$ ) with a clean sizing null (U2:  $p = 0.567$ ) reflects a joint hypothesis problem. VaR breach reduction operates on the *tails* of the conditional distribution: even a modest improvement in variance estimation can shift 2–3 breaches per year from exceedance to coverage (Table 28, Q5:  $\Delta = 2.54$ ). Sharpe improvement, by contrast, requires predictable variation in the *conditional mean*—a fundamentally harder object that OHLCV-based features are unlikely to capture [Fama, 1970]. The OLS decomposition of breach reduction (Table 27) confirms this asymmetry: return entropy is the dominant predictor ( $\hat{\beta}_H = -0.0022$ ,  $p < 0.001$ ), while PUE contributes marginally ( $p = 0.043$ ) and raw dispersion is absorbed once entropy is controlled. The practical implication is precise: entropy-guided model selection is a risk-management tool, not a return-generation tool. This null should be interpreted as a *disciplining result* that prevents the framework from overclaiming into territory where the efficient-markets constraint binds most tightly.

Algorithm 1 formalizes these stages into a deployable routine with explicit inputs, parameters, and decision boundaries calibrated from the benchmark results.



---

**Algorithm 1** Practitioner volatility model-selection protocol.

---

**Require:** Returns  $\{r_t\}$ , OHLCV data for asset  $i$ ; model portfolio  $\mathcal{M}$  ( $|\mathcal{M}| = 12$ , Table 5)

**Ensure:** Conditional-variance forecast  $\hat{\sigma}_{t+1|t}^2$ , VaR estimate, selection-value flag

- 1: — **Stage 1: Diagnostics (cost: seconds)** —
  - 2: Compute Shannon entropy  $\bar{H} \leftarrow H(r_{t-251:t}; k=100)$  over trailing 252-day window
  - 3: Fit EWMA baseline ( $\lambda=0.94$ ):  $\hat{\sigma}_{\text{EWMA},t+1}^2 \leftarrow 0.94 \hat{\sigma}_{\text{EWMA},t}^2 + 0.06 r_t^2$
  - 4: Compute Parkinson RV:  $\hat{\sigma}_{\text{P},t}^2 \leftarrow (\ln H_t - \ln L_t)^2 / (4 \ln 2)$
  - 5: Compute baseline loss:  $\text{QLIKE}_{\text{base}} \leftarrow \frac{1}{T} \sum_t [\hat{\sigma}_{\text{P},t}^2 / \hat{\sigma}_{\text{EWMA},t}^2 - \ln(\hat{\sigma}_{\text{P},t}^2 / \hat{\sigma}_{\text{EWMA},t}^2) - 1]$
  - 6: — **Stage 2: Triage (decision gate)** —
  - 7: **if** PUE < 5% (Table 3: low-PUE assets show <0.1% median QLIKE advantage over EW-COMB, Table 23) **then**
  - 8:   Flag: *low-dispersion regime*
  - 9:    $\hat{\sigma}_{t+1|t}^2 \leftarrow \hat{\sigma}_{\text{EWMA},t+1}^2$  **or** EW-COMB; **skip** full selection
  - 10: **else**
  - 11:   Proceed to Stage 3
  - 12: **end if**
  - 13: — **Stage 3: Full estimation (cost: minutes per asset)** —
  - 14: Fit all  $M=12$  models; discard non-converged
  - 15: Run Model Confidence Set at  $\alpha=0.10$  [Hansen et al., 2011]
  - 16: Select:  $m^* \leftarrow \arg \min_{m \in \text{MCS}} \text{QLIKE}_m$
  - 17: Compute PUE:  $\text{PUE} \leftarrow (\text{QLIKE}_{\text{base}} - \text{QLIKE}_{m^*}) / \text{QLIKE}_{\text{base}} \times 100\%$
  - 18: **if** PUE > 20% **then**
  - 19:   Deploy **EW-COMB** (equal-weight forecast combination) for VaR estimation; retain  $m^*$  for in-sample diagnostics only (per-window selection overfits OOS; Table 30). Report AGI for risk monitoring.
  - 20: **else**
  - 21:   Revert to EWMA (selection cost not justified)
  - 22: **end if**
  - 23: — **Stage 4: Guardrails** —
  - 24: Pre-exclude log-specification models (EGARCH) from the combination unless  $\hat{\sigma}^2 < 100 \times \text{median}(\hat{\sigma}_{\text{train}}^2)$  in every window; this prevents pathological  $\hat{\sigma}^2 \rightarrow 10^{-10}$  episodes from contaminating EW-COMB (Section 8.12)
  - 25: Compute 95% VaR using filtered historical simulation (FHS): standardize residuals  $\varepsilon_t = r_t / \hat{\sigma}_{\text{COMB},t}$ , then bootstrap quantiles (FHS fixes unconditional coverage for any model; the *diversity* benefit is in Christoffersen independence; Table 24, Table 30)
  - 26: Validate via Kupiec unconditional coverage *and* Christoffersen independence ( $p > 0.05$  each); if either fails, revert to EWMA (in-sample joint: 77% Sel+FHS vs. 42% EWMA; OOS: forecast combination preserves Christoffersen benefit, 58–60% vs. 51%; Table 30)
  - 27: **Do not** use model confidence for position sizing (U2 null)
- 

## 9.2 Per-Hypothesis Practitioner Signals

Each hypothesis that passes identifies a deployable signal for quantitative risk management (Table 26).

**Table 26:** Practitioner signals from hypothesis results.

ID	Result	Practitioner Signal
A1	AGI reduces VaR violations	In-sample: model diversity reduces violation clustering (Christoffersen: 79% vs. 62% EWMA+FHS). OOS (Table 30): per-window selection overfits, but forecast combination preserves the benefit (58–60% Chris.). Deploy EW-COMB, screen by AGI.
A2	$\text{AGI} = f(H, \rho);$ adj. $R^2 = 0.31$	AGI is computable from observables—no black box. Traders can replicate the diagnostic from returns alone.
H1	$\rho_s = -0.32$ (entropy $\leftrightarrow$ dispersion)	When entropy is high (noisy returns), models converge—any model suffices. When entropy is low (structured returns), models diverge and careful selection is required.
H2	74.8% Granger rejection	Entropy changes are a leading indicator of volatility regime transitions. Monitor $\Delta H$ as a real-time vol signal.
H3	Borderline monthly; spurious daily	Monthly first-difference correlation is significant ( $p = 0.013$ ). Daily levels correlation is persistence-driven (Supplement S18). Class-level sign structure is interpretable but exploratory.
P1	$\rho_s = 0.53$ (PUE $\leftrightarrow$ QLIKE spread)	PUE directly quantifies how much model selection matters for each specific asset. High PUE = invest in model comparison.
P2	Universal: all 11 classes	The optimal model always materially beats the naïve baseline. Model selection pays off universally—the question is where most.
R1	$\rho_s = 0.19$ (pattern $\leftrightarrow$ QLIKE)	Ordinal pattern structure predicts forecastability. Use as a pre-screening metric for forecast quality assessment.
E1	$\rho_s = 0.19$ (PE $\leftrightarrow$ model gap)	Permutation entropy adds orthogonal information beyond Shannon. Combining both entropy measures improves diagnostic precision.
U1	Tighter intervals with coverage	URI-weighted prediction intervals are narrower and well-calibrated. Deploy for tighter stop-losses and margin optimization.
R2	Null ( $p = 0.13$ )	<i>Guardrail:</i> pattern regularity does not predict GARCH persistence. Do not use ordinal patterns to infer half-lives.
U2	Null ( $p = 0.55$ )	<i>Guardrail:</i> no position-sizing alpha from uncertainty reduction. Framework value is risk management, not return prediction.

### 9.3 Cost–Benefit Quantification

The entropy diagnostic converts a fixed computational cost (fitting  $M$  models for all assets) into a variable cost concentrated where economic value exists. For a 1,496-asset universe with  $M = 12$  models, the diagnostic reduces unnecessary computation by approximately  $(1 - \text{frac}_{\text{PUE} > 20\%}) \times 100\%$  while preserving essentially all available in-sample selection value. This reduction is not a heuristic shortcut but a direct, testable implication of Hypothesis H1: the absence of cross-model dispersion in the high-entropy regime means the marginal information content of additional solvers is zero.

**Economic calibration.** To ground the PUE threshold in economic quantities, we first regress the cross-sectional VaR breach reduction on model-selection metrics (Table 27), then map AGI quintiles to annualized breach counts (Table 28).

**Table 27:** OLS regression:  $\Delta \text{violation\_rate} = \alpha + \beta_1 \cdot \text{AGI} + \beta_2 \cdot \text{PUE} + \beta_3 \cdot D_i + \beta_4 \cdot \hat{H}_{\text{ret}} + \varepsilon$ . HC3 robust standard errors. Adjusted  $R^2 = 0.29$ ,  $F = 15.12$  ( $p < 10^{-12}$ ),  $n = 1,492$ .

Variable	$\hat{\beta}$	$t$ -stat	$p$ -value
Intercept	0.0121	8.49	<0.001
AGI (%)	0.0000	0.94	0.348
PUE (%)	0.0001	2.03	0.043
Dispersion ( $D_i$ )	0.0000	0.19	0.849
$\hat{H}_{\text{ret}}$	−0.0022	−4.69	<0.001

Return entropy is the dominant predictor of breach reduction ( $p < 0.001$ ): lower-entropy assets benefit more from model selection, consistent with H1. PUE contributes marginally ( $p = 0.036$ ), while AGI and raw dispersion are absorbed once entropy is controlled.

Table 28 partitions assets into AGI quintiles and reports the mean number of annual 5%-VaR breaches under EWMA (baseline) versus the best solver (optimized).

**Table 28:** Economic calibration: AGI quintile  $\rightarrow$  annualized VaR breach reduction. Baseline is EWMA; optimized is MCS-selected solver. Breach count assumes 252 trading days.

Quintile	Mean AGI (%)	Breaches/yr (base)	Breaches/yr (opt)	$\Delta$
Q1 (lowest)	2.2	13.22	11.80	1.43
Q2	3.5	13.17	11.65	1.53
Q3	4.7	13.07	11.53	1.54
Q4	7.0	12.83	11.20	1.63
Q5 (highest)	115.6	11.95	9.41	2.54

Even in the lowest AGI quintile, model selection eliminates approximately 1.4 VaR breaches per year; in the highest quintile, the reduction exceeds 2.5 breaches—equivalent

to eliminating roughly one excess exceedance per quarter. A cross-sectional regression of breach reduction on AGI confirms a significant positive relationship ( $\beta = 0.005$  breaches per 1% AGI,  $p < 10^{-77}$ ).

**PUE threshold calibration.** The  $\text{PUE} > 20\%$  threshold reflects a trade-off between breach-reduction magnitude and portfolio coverage. Table 29 presents the full sweep, enabling practitioners to select a threshold suited to their risk tolerance. We adopt 20% as the operational inflection point: below it, the mean annual breach reduction is modest ( $\sim 1.6/\text{yr}$ ); above it, the reduction roughly triples ( $\sim 4.2/\text{yr}$ ), though the number of actionable assets narrows to 86 (5.8% of the universe).

**Table 29:** PUE threshold sweep: breach reduction and Gaussian-quantile VaR calibration for assets exceeding each threshold. Kupiec pass rate uses Gaussian quantiles (see Table 24 for FHS quantile results).

PUE threshold	$n$ above	Mean $\Delta$ breaches/yr	Kupiec pass (%)
> 5%	638	2.04	45.1
> 10%	229	2.78	34.1
> 15%	125	3.49	27.2
> 20%	86	4.24	24.4
> 25%	72	4.62	23.6
> 30%	66	4.67	22.7
> 50%	48	5.53	16.7

The declining Kupiec pass rate above  $\text{PUE} > 20\%$  reflects the Gaussian quantile mismatch documented in Table 24: high-PUE assets have the most complex models with the most non-Gaussian residuals ( $\nu \approx 5.5$ ), so pairing them with Gaussian quantiles produces maximal over-coverage. Under FHS quantiles, Sel+FHS achieves 97.0% Kupiec pass across the full universe (Table 24), confirming that breach reduction reflects genuine calibration improvement, not conservatism. The 20% threshold balances breach reduction against the number of actionable assets ( $n = 86$  assets, or 5.8% of the universe, where intensive model selection is warranted).

## 10 Scope and Future Directions

Key robustness checks—H1 confound controls (Supplement S19), FHS attribution (Supplement S21), dynamic OOS switching (Supplement S22), Student- $t$  innovation sensitivity (Supplement S12), and EGARCH safeguards—are reported in the main text. This section discusses remaining scope boundaries and open questions.

**In-sample versus out-of-sample VaR.** The VaR attribution (Table 24) evaluates full-sample fitted  $\hat{\sigma}^2$  against the full return series. The dynamic switching test (Supplement S22)

provides a partial OOS validation: using rolling  $\hat{H}_t$  to switch between EWMA and the selected model’s full-sample  $\hat{\sigma}^2$ , Christoffersen independence drops from 79.2% (static selection) to 71.5% (dynamic switching)—a 7.7pp OOS decay that nonetheless remains 8.5pp above the EWMA+FHS baseline (63.0%). This confirms that model selection’s Christoffersen benefit persists under real-time implementable decisions but attenuates relative to the full-sample oracle.

A fully walk-forward VaR backtest (Supplement S26) provides the definitive test: all twelve solvers are re-fit on each 252-day training window, the best is selected by QLIKE, and each model’s own  $\hat{\sigma}^2$  is applied to the 63-day holdout via FHS with consistent residual standardization ( $\varepsilon_t = r_t/\hat{\sigma}_t$  for all solvers). Table 30 reports six strategies across 1,491 assets.

**Table 30:** Walk-forward VaR backtesting (Supplement S26,  $n = 1,491$ ). Each strategy re-fits per 252-day window and evaluates on the subsequent 63-day holdout. Triage uses Algorithm 1: selection if  $\hat{H} < 2.77$ , EWMA otherwise.

Strategy	Kupiec %	Chris. %	Joint %	br/yr
EWMA + Gaussian	71.2	61.2	42.1	13.14
EWMA + FHS	99.5	51.3	51.2	12.65
Selection + FHS	47.8	43.3	27.2	14.38
EW-COMB + FHS	27.0	58.1	14.8	10.19
IQ-COMB + FHS	37.2	59.6	21.6	10.52
Triage + FHS	98.4	50.6	50.0	12.72

Per-window best-model selection overfits: its  $\hat{\sigma}^2$  tracks training patterns too tightly, producing excess violations OOS (14.38 br/yr vs. 12.65 for EWMA) and destroying Kupiec calibration (47.8% vs. 99.5%). This confirms the diagnostic-not-predictive thesis already established by the walk-forward QLIKE null (Section 8.12). In contrast, forecast *combination* preserves the Christoffersen benefit out-of-sample: EW-COMB achieves 58.1% and IQ-COMB 59.6%, gains of +6.8pp and +8.3pp over EWMA+FHS ( $p < 10^{-11}$ , paired Wilcoxon). Model averaging diversifies specification error across solvers, preventing the winner’s-curse overfitting that plagues per-window selection while retaining the variance-timing diversity that reduces violation clustering. Their Kupiec shortfall (27%–37%) reflects over-conservative  $\hat{\sigma}^2$  from averaging (10.2–10.5 br/yr vs. 12.6 expected), not miscalibration—a known property of combination forecasts [Timmermann, 2006] addressable by quantile recalibration.

The entropy-gated triage strategy (98.4% Kupiec, 50.6% Christoffersen) is nearly indistinguishable from pure EWMA+FHS, confirming that the triage rule correctly identifies the vast majority of windows as EWMA-sufficient (selection used on only 3.1% of windows). These results sharpen Algorithm 1: for VaR deployment, EW-COMB is the robust OOS default that preserves the Christoffersen benefit, while per-window selection is reserved for in-sample diagnostics and regime identification.

**Realized-variance proxy quality.** The Parkinson estimator is approximately five times more efficient than squared returns [Parkinson, 1980] but remains a noisy proxy. The proxy-sensitivity analysis (Section 8) confirms that solver *rankings* shift when switching between Parkinson and close-to-close proxies ( $\rho_{\text{rank}} = 0.67$ ), though MCS *inclusion order* is more stable ( $\rho = 0.75$ ). Extension to tick-level realized variance [Andersen et al., 2003] or realized kernels [Barndorff-Nielsen et al., 2008] would further sharpen QLIKE-based comparisons, particularly for assets where the Parkinson estimator is biased by discrete trading (e.g., illiquid fixed-income ETFs).

**Forecast horizon scope.** The multi-horizon analysis (Section 8.10) extends to  $h \in \{5, 20\}$  trading days, revealing horizon-dependent dynamics (H1 sign reversal, P1/A1 strengthening). However, these extensions use GARCH( $h$ )-step variance propagation, which assumes stationary parameters over the forecast period. True out-of-sample multi-step evaluation—with re-estimation and walk-forward Sharpe analysis at weekly and monthly horizons—remains future work. The HAR model’s multi-component structure [Corsi, 2009] suggests that horizon effects may be material for models designed explicitly for longer scales.

**Out-of-sample forecast combination.** Table 23 demonstrates that MCS-selected models dominate forecast combinations (EW-COMB, IQ-COMB) in-sample, with the advantage increasing monotonically with PUE. To test whether this persists out-of-sample, we evaluate all three strategies—EWMA, EW-COMB, and best-in-MCS—on the last 63-day holdout window for 1,475 assets using QLIKE against Parkinson realized variance (Supplement S24). The best-in-MCS model achieves lower QLIKE than EW-COMB for 58.4% of assets ( $p < 10^{-32}$ , Wilcoxon), and both outperform EWMA (EW-COMB: 47.6%; best-model: 51.4%). The selection advantage over combination concentrates in the moderate- and high-PUE regime (best-model beats EW-COMB: 64.3% for PUE 5–20%, 57.5% for PUE  $\geq 20\%$ ), while in low-PUE the difference narrows to 54.6%. This confirms that forecast combination partially—but not fully—substitutes for selection, and that the substitution gap widens precisely where entropy diagnostics predict selection to be most valuable. A true walk-forward comparison with re-fitting at each window would provide the definitive test; the present holdout analysis is consistent with but weaker than such a design.

**Macro proxy limitation.** H3 uses VIX as the primary macro-regime proxy. Confound controls (Section 7) include credit spreads ( $\rho = 0.124$ ,  $p = 0.049$ ) and term spreads ( $\rho = 0.051$ ,  $p < 0.001$ ), but these enter as partial-correlation adjustments, not as alternative primary proxies. VIX is mechanically related to equity volatility and may not capture regime variation in non-equity classes. Using credit spreads [Gilchrist and Zakrajšek,

2012], term-structure slope, realized economic uncertainty indices [Jurado et al., 2015], or textual sentiment measures [Baker et al., 2016] as the *primary* H3 proxy—rather than as confound controls—should be explored.

**Static universe and survivorship.** The asset universe is fixed at run time. Survivorship bias is partially mitigated by including low-liquidity tickers and broad asset classes. A robustness check (Supplement S14) excludes 99 post-2020 entrants and 44 early-exiting tickers, leaving a stable core of 1,357 assets (90.9%). H1 on this subset yields  $\rho_s = -0.338$  ( $p = 1.4 \times 10^{-37}$ ) and P1 yields  $\rho_s = 0.504$  ( $p = 2.5 \times 10^{-88}$ )—virtually unchanged from the full sample—confirming that the main findings are not driven by survivorship artifacts. A time-varying universe with explicit entry/exit analysis would further strengthen external validity.

**Transaction costs in U2.** The transaction-cost sensitivity analysis (Section 7) sweeps costs from 0 to 100 bps, revealing a sign reversal at  $\sim 50$  bps where URI sizing becomes advantageous. However, we apply a uniform cost across all assets. Asset-class-specific costs (lower for liquid equities, higher for crypto) and market-impact models [Almgren and Chriss, 2001] would refine the analysis. The sign reversal suggests that URI sizing may have conditional value in high-friction environments, but a proper market-microstructure treatment is needed to confirm this.

**NPE discriminative power.** Normalized permutation entropy (NPE) exhibits extreme concentration near its theoretical maximum ( $\mu = 0.994$ ,  $\sigma = 0.020$ , skewness =  $-18.0$ ): the vast majority of assets have near-maximal temporal randomness. While NPE is strongly associated with AGI ( $\rho_s = -0.39$ ,  $p < 10^{-54}$ ), its discriminative power is limited to the small tail of structured assets that deviate from the ceiling. Nevertheless, the continuous variable retains 36% more correlation with dispersion than a binarized version ( $\rho_s = -0.372$  vs.  $-0.238$ ; Supplement S17), so ceiling compression limits but does not eliminate NPE’s diagnostic value. Higher-order permutation entropy (larger embedding dimension  $d$ ), multiscale entropy [Costa et al., 2005], or transfer entropy [Schreiber, 2000] may provide finer discrimination across the asset universe.

**Entropy estimation under non-stationarity.** The plug-in estimator with fixed  $K$  assumes approximate stationarity within each sample. A formal test (Supplement S13) applies per-asset rolling-variance ratio analysis to the 252-day entropy series: 814 of 1,494 assets (54.5%) exhibit variance-stable entropy, while the remainder show significant temporal heterogeneity. Restricting H1 and P1 to the variance-stable subset yields  $\rho_s = -0.263$  ( $p = 2.4 \times 10^{-14}$ ,  $n = 812$ ) and  $\rho_s = 0.532$  ( $p = 2.0 \times 10^{-60}$ ), respectively—both highly significant and directionally unchanged. The H1 attenuation from  $-0.332$

to  $-0.263$  is consistent with non-stationary assets contributing additional variance that inflates the cross-sectional association; the stable-subset result represents a conservative lower bound. Adaptive methods such as the Inclán and Tiao [1994] ICSS algorithm for structural break detection, followed by piecewise entropy estimation, could further improve the diagnostic.

**Pre-registration design lesson.** The rigid pre-registration protocol, while preventing specification search, created an avoidable underpowering in H3. The 252-day rolling window induces  $AC(1) = 0.92$  in monthly entropy means, reducing the effective sample size from  $n = 303$  to  $n_{\text{eff}} \approx 12$  (Bayley–Hammersley; Supplement S18). A first-difference test—which removes this persistence—yields  $\rho = -0.143$  ( $p = 0.013$ ), but was not pre-registered and therefore cannot serve as the primary result. Future pre-registrations of rolling-window-derived series should specify first-difference or detrended tests as the canonical statistic to avoid conflating window-overlap persistence with the null hypothesis.

**From entropy triage to ML-based model recommendation.** The current framework uses Shannon entropy as a scalar triage statistic: select or skip. The Student- $t$  innovation analysis (Supplement S12) demonstrates that the heterogeneity required for a richer approach already exists in our data—best-solver agreement varies from 47% (volatility) to 86% (equity index) across classes, and distributional assumptions interact with asset-class structure in solver-specific ways. The entropy diagnostic developed here would serve as a first-stage filter in a two-stage pipeline: compute  $\hat{H}$  to determine whether selection effort is warranted at all (94% of assets resolved at this stage), then deploy a supervised meta-learner [Talagala et al., 2018, Bischl et al., 2016] over interpretable sub-features (sample autocorrelation, excess kurtosis, Hurst exponent, regime-switching indicators) for the  $\sim 6\%$  of assets in the high-PUE regime, moving from a binary “select or skip” to a per-asset model recommendation. Building and validating this second stage remains open.

## 11 Conclusion

We have developed and executed a pre-registered, integrity-gated framework that addresses a practical meta-question in volatility forecasting: *when does model selection matter?*

The framework makes four contributions. First, it formalizes the model-selection value problem in volatility forecasting, introducing quantitative diagnostics (SVI, PUE, QLIKE dispersion) and propositions linking entropy to expected selection value. Second, it implements a cryptographic integrity layer that provides machine-enforced pre-registration, preventing undisclosed specification changes and ensuring complete reporting of all pre-registered hypotheses—addressing the specification-flexibility component of the analyst



degrees of freedom documented by Simmons et al. [2011] and Harvey et al. [2016]. Third, it evaluates 12 pre-specified hypotheses across 1,496 assets spanning eleven asset classes—with full subgroup analysis across all classes—providing the broadest cross-asset volatility comparison benchmark to date. Fourth, it reports null results with the same rigor as positive findings, bounding claims to what the evidence supports.

The central empirical finding is that entropy-based diagnostics reliably identify the regime in which model selection has economic value. Model selection improvement concentrates where models disagree most, entropy dynamics lead volatility changes, and prediction intervals can be tightened while preserving coverage. A variance-timing analysis (Supplement S16) reveals the mechanism: the selected model allocates lower  $\hat{\sigma}^2$  than EWMA on calm days but 22% higher on crash-adjacent days, producing genuine calibration improvement—not conservatism. A  $2 \times 2$  in-sample attribution (Supplement S21) reveals two complementary mechanisms: FHS fixes unconditional coverage for any model (EWMA+FHS: 100% Kupiec), while model *diversity* reduces violation clustering (Christoffersen: 79% vs. 62% for EWMA+FHS; Table 24). Walk-forward backtesting across 1,491 assets (Table 30) then establishes the paper’s central out-of-sample finding: per-window best-model selection overfits, but forecast *combination* preserves the clustering benefit. EW-COMB achieves 58.1% and IQ-COMB 59.6% Christoffersen pass—gains of +6.8 and +8.3 percentage points over EWMA+FHS ( $p < 10^{-11}$ ). The framework’s value lies in three outputs: entropy-based triage (86% compute savings), mechanism identification (FHS for level, model diversity for clustering), and forecast combination as the robust OOS default—not in return prediction, as confirmed by the negligible per-asset forecast improvement over EWMA (Section 8.12) and the sizing null (U2,  $p = 0.567$ ).

Several results deepen the empirical picture beyond the core hypothesis battery. The multi-horizon analysis (Section 8.10) reveals that the entropy–dispersion sign reverses at  $h > 1$ : all models converge at the daily horizon, but diverge at weekly and monthly scales where parametric structure matters. Selection value (P1) and risk-calibration benefit (A1) both strengthen monotonically with horizon, reinforcing the practical relevance for multi-day risk management. The within-asset temporal analysis (Section 8.11) confirms that H1 is not merely a cross-sectional artefact: the partial correlation after controlling for asset class, liquidity, and volatility level remains  $\rho_{\text{partial}} = -0.328$  (Supplement S19), the association is negative within all eleven classes, and 82% of individual assets exhibit the negative entropy–dispersion association over time. The stratified H3 analysis (Table 11) shows that the entropy–VIX levels correlation strengthens at longer windows ( $\rho_s = -0.38$  at  $w = 504$ ), though this partly reflects increased persistence from window overlap; the monthly first-difference test ( $\rho = -0.143$ ,  $p = 0.013$ ; Supplement S18) provides the persistence-robust confirmation. A daily-frequency per-class decomposition reveals interpretable sign heterogeneity (macro-sensitive classes negative, micro-dominated positive), though the daily levels correlation is driven by shared persistence ( $\text{AC}(1) > 0.97$ ; Supplement S18).

and vanishes after controlling for lagged VIX. The genuine entropy–VIX signal emerges in monthly first differences ( $\rho = -0.143$ ,  $p = 0.013$ ), confirming the relationship at the macro-relevant frequency. Residual diagnostics reveal a kurtosis gradient ( $\kappa = 2.1$  to  $6.9$ ) by AGI tercile, linking higher model-selection value to heavier-tailed volatility dynamics. Finally, normalized permutation entropy (NPE) and ordinal pattern effectiveness (pe) are empirically orthogonal ( $\rho = -0.03$ ), confirming that temporal ordering structure contributes independent information to the model-selection problem.

Beyond VaR, the entropy triage principle extends to adjacent applications. In mean-variance optimization, the conditional-variance input is the dominant source of estimation error [Chopra and Ziemba, 1993]. The walk-forward backtest (Table 30) quantifies  $\hat{\sigma}^2$  calibration via violation rates: EWMA is nearly unbiased ( $+0.4\%$ , lowest cross-asset dispersion  $CV = 0.063$ ), while per-window selection underestimates risk by  $14\%$  ( $CV = 0.107$ )—distorting minimum-variance weights. Forecast combination is over-conservative ( $-19\%$  bias) but with the lowest violation clustering (Christoffersen:  $58\text{--}60\%$  vs.  $51\%$  for EWMA); its level bias is correctable by a single scalar recalibration ( $\times 0.81$  on  $\hat{\sigma}^2$ ). For portfolios with VaR or CVaR constraints—where violation clustering inflates tail-risk estimates during stress—EW-COMB provides the best OOS trade-off. In regulatory technology (regtech), the decision protocol (Algorithm 1) provides an auditable, rule-based workflow for model governance that maps directly to Basel III internal-model validation requirements [Basel Committee, 2010]—including automated documentation of when models were selected, which metrics justified the selection, and when the EWMA default was deployed. More broadly, the “diagnose-before-selecting” principle applies whenever a practitioner faces a large portfolio of instances (assets, strategies, signals) and a portfolio of candidate algorithms: compute a cheap complexity diagnostic first, invest in expensive estimation only where the diagnostic signals positive expected value.

We hope this work encourages two norms: first, the adoption of computational pre-registration mechanisms in quantitative finance research, and second, the practice of diagnosing model-selection value before engaging in model horse races. The supplementary analysis system (S1–S28) is released as an extensible interface against the benchmark’s immutable data store, enabling researchers to test new hypotheses, alternative model specifications, or additional risk metrics without modifying the core pipeline or breaking the pre-registration chain.

For the practitioner, the paper delivers three concrete outputs. *First, a triage rule:* computing  $\hat{H}$  takes seconds and identifies the  $94\%$  of assets where EWMA suffices, reducing the model-fitting workload by  $86\%$  (from  $17,904$  solver fits to  $2,438$  per rebalance). *Second, a deployment strategy:* for VaR applications, EW-COMB+FHS is the robust out-of-sample default. Walk-forward backtesting across  $1,491$  assets (Table 30) demonstrates that forecast combination preserves the Christoffersen clustering benefit out-of-sample ( $58\text{--}60\%$  vs.  $51\%$  for EWMA+FHS,  $p < 10^{-11}$ ), whereas per-window best-model selection overfits

(27.2% joint coverage). The mechanism is variance-timing diversity: averaging across solvers hedges against winner’s-curse overfitting while retaining the  $\hat{\sigma}^2$  responsiveness to crash-adjacent days that individual selection achieves in-sample (14 percentage-point timing advantage; Supplement S16). The entropy diagnostic tells the practitioner *when* combination adds value (the low- $\hat{H}$  regime); the  $2 \times 2$  attribution (Table 24) explains *why* (FHS for unconditional coverage, model diversity for violation independence). Together, they transform a collection of models into an interpretable, OOS-validated risk system. *Third, calibrated scope:* three null results discipline the framework’s claims. Walk-forward analysis shows no improvement in point forecasting (Section 8.12) and confirms that per-window model selection overfits out-of-sample (Table 30). Hypothesis U2 ( $p = 0.567$ ) rules out position-sizing alpha, confirmed at 95% power in Supplement S28 ( $p = 0.76$ ). These boundaries are not failures; they *sharpen* the positive contributions by ensuring they are not conflated with return predictability. The framework’s value is precisely where it claims to be: computational triage, mechanism understanding, and OOS-validated violation-clustering reduction via forecast combination.

## Data Availability

All price data are sourced from Yahoo Finance via the `yfinance` Python package and are freely downloadable. Source code, pre-registration specifications, integrity artifacts, the test suite (911 tests), and selected run outputs are released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license at [https://github.com/oliviersaidi/PACF\\_F](https://github.com/oliviersaidi/PACF_F).

The repository provides: CSV exports (`metric_table.csv`, `solver_table.csv`, `hypothesis_table.csv`), machine-readable results (`results_all.json`), supplementary analysis outputs (S1–S28 JSON), all figures, interactive HTML reports (`index.html`, sensitivity and robustness dashboards), the run log, and the complete revision scripts (`revision/`). Two large intermediate artifacts—the SQLite database (`run_data.db`, ~2.7 GB) and the hash-chained computation log (`atomic_log.jsonl`, ~2.0 GB)—are excluded due to size constraints but are fully reproducible by re-running the benchmark.

The core benchmark completes in approximately 2.4 hours on a standard workstation (Apple M3 Max, 12 cores); supplementary analyses (S1–S28) run independently and complete in under 30 minutes each.

## Declaration of Interest

The author declares no competing financial or non-financial interests.

## Acknowledgments

The author thanks Anthropic for assistance with research support and debugging. All scientific decisions, interpretations, and errors remain the author’s own.

## References

- Almgren, R., Chriss, N., 2001. Optimal execution of portfolio transactions. *Journal of Risk* 3(2), 5–39.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2003. Modeling and forecasting realized volatility. *Econometrica* 71(2), 579–625.
- Angrist, J.D., Pischke, J.-S., 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.
- Andrews, I., Kasy, M., 2019. Identification of and correction for publication bias. *American Economic Review* 109(8), 2766–2794.
- Baillie, R.T., Bollerslev, T., Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74(1), 3–30.
- Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. *Quarterly Journal of Economics* 131(4), 1593–1636.
- Bandt, C., Pompe, B., 2002. Permutation entropy: a natural complexity measure for time series. *Physical Review Letters* 88(17), 174102.
- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76(6), 1481–1536.
- Barone-Adesi, G., Giannopoulos, K., Vosper, L., 1999. VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets* 19(5), 583–602.
- Basel Committee on Banking Supervision, 2010. *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*. Bank for International Settlements.
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* 57(1), 289–300.

- Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics* 19(4), 465–474.
- Bischl, B., Mersmann, O., Trautmann, H., Kotthoff, L., 2016. Algorithm selection based on exploratory landscape analysis and cost-sensitive learning. In: *Proc. GECCO 2016*, pp. 313–320.
- Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31(3), 307–327.
- Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics* 69(3), 542–547.
- Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, 2011. Supervisory Guidance on Model Risk Management (SR 11-7/OCC 2011-12).
- Brodeur, A., Cook, N., Heyes, A., 2020. Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review* 110(11), 3634–3660.
- Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. *Review of Financial Studies* 33(5), 2134–2179.
- Chopra, V.K., Ziemba, W.T., 1993. The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management* 19(2), 6–11.
- Christoffersen, P.F., 1998. Evaluating interval forecasts. *International Economic Review* 39(4), 841–862.
- Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum.
- Costa, M., Goldberger, A.L., Peng, C.-K., 2005. Multiscale entropy analysis of biological signals. *Physical Review E* 71(2), 021906.
- Corsi, F., 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.
- Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13(3), 253–263.
- Dimpfl, T., Peter, F.J., 2013. Using transfer entropy to measure information flows between financial markets. *Studies in Nonlinear Dynamics & Econometrics* 17(1), 85–102.

- Dimpfl, T., Peter, F.J., 2014. The impact of the financial crisis on transatlantic information flows: an intraday analysis. *Journal of International Financial Markets, Institutions and Money* 31, 1–13.
- Ding, Z., Granger, C.W.J., Engle, R.F., 1993. A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1(1), 83–106.
- Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics* 80(4), 549–560.
- Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50(4), 987–1007.
- Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* 25(2), 383–417.
- Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator:  $L_2$  theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57(4), 453–476.
- Fisher, R.A., 1921. On the ‘probable error’ of a coefficient of correlation deduced from a small sample. *Metron* 1, 3–32.
- Fisher, R.A., 1932. *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd.
- Francq, C., Zakoïan, J.-M., 2010. *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley & Sons.
- Giacomini, R., White, H., 2006. Tests of conditional predictive ability. *Econometrica* 74(6), 1545–1578.
- Gilchrist, S., Zakrajšek, E., 2012. Credit spreads and business cycle fluctuations. *American Economic Review* 102(4), 1692–1720.
- Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48(5), 1779–1801.
- Gulko, L., 1999. The entropy theory of stock option pricing. *International Journal of Theoretical and Applied Finance* 2(3), 331–355.
- Hansen, B.E., 1994. Autoregressive conditional density estimation. *International Economic Review* 35(3), 705–730.
- Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20(7), 873–889.

- Hansen, P.R., 2005. A test for superior predictive ability. *Journal of Business & Economic Statistics* 23(4), 365–380.
- Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. *Econometrica* 79(2), 453–497.
- Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. *Review of Financial Studies* 29(1), 5–68.
- Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2), 281–291.
- He, J., Hamori, S., 2022. Information flows among financial markets: a systematic review. *International Review of Economics & Finance* 82, 555–582.
- Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.
- Hou, Y., Liu, F., Gao, J., Cheng, C., Song, C., 2017. Characterizing complexity changes in Chinese stock markets by permutation entropy. *Entropy* 19(10), 514.
- Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. *Review of Financial Studies* 33(5), 2019–2133.
- Hull, J.C., 2017. *Options, Futures, and Other Derivatives*, 10th ed. Pearson.
- Inclán, C., Tiao, G.C., 1994. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association* 89(427), 913–923.
- J.P. Morgan/Reuters, 1996. *RiskMetrics—Technical Document*, 4th ed.
- Jobson, J.D., Korkie, B.M., 1981. Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance* 36(4), 889–908.
- Jurado, K., Ludvigson, S.C., Ng, S., 2015. Measuring uncertainty. *American Economic Review* 105(3), 1177–1216.
- Kerby, D.S., 2014. The simple difference formula: an approach to teaching nonparametric correlation. *Comprehensive Psychology* 3, 11.IT.3.1.
- Kotthoff, L., 2016. Algorithm selection for combinatorial search problems: a survey. *AI Magazine* 35(3), 48–60.
- Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3(2), 73–84.

- Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17(3), 1217–1241.
- Laurent, S., Rombouts, J.V., Violante, F., 2012. On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics* 27(6), 934–955.
- Leamer, E.E., 1983. Let’s take the con out of econometrics. *American Economic Review* 73(1), 31–43.
- Ledoit, O., Wolf, M., 2008. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance* 15(5), 850–859.
- Liu, L.Y., Patton, A.J., Sheppard, K., 2015. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics* 187(1), 293–311.
- Lo, A.W., 2004. The adaptive markets hypothesis. *Journal of Portfolio Management* 30(5), 15–29.
- Lütkepohl, H., 2005. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.
- Maasoumi, E., Racine, J.S., 2015. Entropy and predictability of stock market returns. *Journal of Econometrics* 107(1–2), 291–312.
- MacKinnon, J.G., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 29(3), 305–325.
- Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1), 50–60.
- Markowitz, H., 1952. Portfolio selection. *Journal of Finance* 7(1), 77–91.
- McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71(1), 5–32.
- Miller, G.A., 1955. Note on the bias of information estimates. In: Quastler, H. (Ed.), *Information Theory in Psychology*. Free Press, pp. 95–100.
- Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59(2), 347–370.
- Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3), 703–708.



- Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T., 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11), 2600–2606.
- Olken, B.A., 2015. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives* 29(3), 61–80.
- Ortiz-Cruz, A., Rodriguez, E., Ibarra-Valdez, C., Alvarez-Ramirez, J., 2012. Efficiency of crude oil markets: evidences from informational entropy analysis. *Energy Policy* 41, 365–373.
- Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. *Journal of Business* 53(1), 61–65.
- Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
- Pelé, D.T., Mazuure, E., 2019. Using high-frequency entropy to forecast Bitcoin’s daily Value at Risk. *Entropy* 21(2), 102.
- Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: comparing approaches. *Review of Financial Studies* 22(1), 435–480.
- Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. *Journal of the American Statistical Association* 89(428), 1303–1313.
- Rice, J.R., 1976. The algorithm selection problem. *Advances in Computers* 15, 65–118.
- Risso, W.A., 2008. The informational efficiency and the financial crashes. *Research in International Business and Finance* 22(3), 396–408.
- Saidi, O., 2025. Pattern-Aware Complexity Framework (PACF). Working paper.
- Schreiber, T., 2000. Measuring information transfer. *Physical Review Letters* 85(2), 461.
- Sensoy, A., Ozturk, K., Hacıhasanoglu, E., 2015. Constructing a financial fragility index for emerging economies. *Finance Research Letters* 14, 218–225.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3), 379–423.
- Shephard, N., Sheppard, K., 2010. Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25(2), 197–231.
- Sheppard, K., 2023. arch: Autoregressive conditional heteroskedasticity models in Python. <https://github.com/bashtage/arch>.

- Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11), 1359–1366.
- Smith-Miles, K.A., 2009. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys* 41(1), 1–25.
- Talagala, T.S., Hyndman, R.J., Athanasopoulos, G., 2018. Meta-learning how to forecast time series. Working paper, Monash University.
- Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* 16(4), 437–450.
- Toda, H.Y., Yamamoto, T., 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* 66(1–2), 225–250.
- West, K.D., 1996. Asymptotic inference about predictive ability. *Econometrica* 64(5), 1067–1084.
- White, H., 2000. A reality check for data snooping. *Econometrica* 68(5), 1097–1126.
- Wooldridge, J.M., 2019. *Introductory Econometrics: A Modern Approach*, 7th ed. Cengage.
- Timmermann, A., 2006. Forecast Combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, Vol. 1. Elsevier, pp. 135–196.
- Zakoian, J.-M., 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* 18(5), 931–955.
- Zunino, L., Zanin, M., Tabak, B.M., Pérez, D.G., Rosso, O.A., 2009. Forbidden patterns, permutation entropy and stock market inefficiency. *Physica A* 388(14), 2854–2864.
- Zunino, L., Tabak, B.M., Serinaldi, F., Zanin, M., Pérez, D.G., Rosso, O.A., 2010. Commodity predictability analysis with a permutation information theory approach. *Physica A* 390(5), 876–890.

## A Reproducibility Appendix

### A.1 Run Artifacts

Each benchmark execution produces a deterministic artifact set, linked by a unique run identifier:

- `run_data.db` — SQLite database (primary data store,  $\sim 2.7$  GB; excluded from repository, reproducible by re-running the benchmark): OHLCV prices (6.8M rows), rolling entropy and realized variance (6.4M rows), per-asset $\times$ per-solver conditional variance arrays and QLIKE (17,904 rows), MCS results (17,810 rows), entropy, dispersion, risk, and pattern metrics (1,492 rows each), and hypothesis test results (12 rows). All CSV files below are exports of views from this database.
- `results_all.json` — machine-readable hypothesis results
- `hypothesis_table.csv` — 12 rows: statistics,  $p$ -values, corrections, pass/fail
- `metric_table.csv` —  $1,496 \times 28$  asset-level metrics
- `solver_table.csv` — per-asset, per-solver QLIKE, convergence, and diagnostic fields
- `atomic_log.jsonl` — full hash-chained computation log ( $\sim 2.0$  GB; excluded from repository, reproducible)
- `spec_manifest.json` — pre-registration hashes
- `batch_manifest.json` — complete-reporting proof
- `index.html` — interactive report dashboard
- `sensitivity_analysis.html` — all sweep results
- `robustness_battery.html` — confound checks

## A.2 Software and Configuration

**Hardware and software environment.** All benchmark runs were executed on a single workstation: Apple M3 Max (12-core CPU, 36 GB unified memory), macOS 26.3, Python 3.13.1 with `arch 7.x` for GARCH estimation and `scipy 1.x` for statistical tests. The main benchmark (run `3db6c1`, v4.4.0) completed in approximately 2.4 h using 12 parallel workers on an Apple M3 Max (36 GB RAM). The benchmark evaluates twelve volatility solvers—GARCH, EGARCH, GJR-GARCH, TGARCH, APARCH, FIGARCH, ARCH, EWMA, HAR, HEAVY, EW-COMB, and IQ-COMB—across 1,496 registered tickers (1,492 after four exclusions for insufficient solver convergence). Supplementary analysis scripts (S1–S28) run independently after the core benchmark with parallelized execution.

**Extensible supplement architecture.** The supplementary system is designed as an extensibility interface. The core benchmark produces an immutable SQLite database (`run_data.db`: 16 tables,  $> 6.8$  M rows) that serves as a stable data contract. Each supplementary script follows a uniform pattern: auto-discover the run directory, read from the database (read-only), compute an independent analysis, and write structured JSON to `supplementary/{name}/`. The system supports two tiers of extension: *independent* supplements read only from the core data store (27 of 28 current scripts), while *compositional*

supplements (e.g., S28) additionally consume outputs from earlier supplements, enabling meta-analyses that combine results across the evidence base. Both tiers enforce a strict separation from the locked pre-registered benchmark (whose results are hash-chained and immutable): extensions can add evidence and build on each other’s outputs but cannot alter core results. Researchers wishing to test additional hypotheses, alternative model specifications, or new risk metrics can do so by adding a new supplement script—either against the core data contract alone or by leveraging the structured JSON outputs of existing supplements—without modifying the core pipeline or invalidating the pre-registration chain.

**Table 31:** Key configuration parameters.

Parameter	Symbol	Value
Entropy bins	$K$	100
Permutation entropy dimension	$D$	5
Permutation entropy delay	$\tau$	1
Rolling window (canonical)	$w$	252
Minimum observations	MIN_OBS	252
Minimum converged solvers	MIN_SOLVERS	3
Bootstrap replications	$B$	10,000
MCS block length (canonical)	$\ell$	$\lfloor n^{1/3} \rfloor$
MCS block sensitivity set		$\{2, 5, 10, \lfloor n^{1/3} \rfloor\}$
MCS significance (canonical)	$\alpha_{\text{MCS}}$	0.05
MCS alpha sensitivity set		$\{0.05, 0.10, 0.25\}$
EWMA decay factor (baseline)	$\lambda$	0.94
Holm family (core entropy)	$m$	3
BH FDR control	$q$	0.10
Effect-size threshold		$ \rho_s  \geq 0.10$
Transaction cost (U2)	$c$	10 bps
Walk-forward estimation window		252 days
Walk-forward evaluation window		63 days

### A.3 Runtime Integrity Evidence

Table 32 reports the integrity artefacts produced by the benchmark run used in this paper. The spec-manifest hash was computed before any hypothesis testing began; the batch-manifest hash was computed after all 12 hypotheses completed. The reproducibility check confirms that spec hashes match across pre-computation and post-computation stages, providing machine-verifiable evidence that no specification was altered after observing results.

**Table 32:** Integrity artefacts from the reported benchmark run. Hashes are SHA-256, truncated for display. Full values are available in the replication package.

Check	Mechanism	This Run
Spec pre-registration	SHA-256 lock before Step 1	0148e01b6389...
Anti-HARKing	Re-hash at Step 11, compare	Match confirmed
Complete reporting	12/12 in batch manifest	12 tested, 0 missing
Per-record provenance	<code>stamp_record()</code> chain	12 chain hashes recorded
Config determinism	Config hash	3db6c1a66265...
Master seed	Reproducibility seed	42
Run identifier		run_3db6c1_20260304
Elapsed time (core benchmark)		~2.4 h

## A.4 Test Suite

The codebase includes 911 automated tests covering schema validation, solver correctness, hypothesis logic, metric invariants, figure generation, integrity chain verification, and report assembly. All tests pass prior to each benchmark run as a pre-condition for execution. The test suite is included in the replication package and can be run via `python -m pytest pacf_v4/tests/`.

## A.5 Table-to-Artifact Mapping

Table 33 provides the mapping from each paper table to the specific run artifact and field names from which values were extracted, enabling independent verification.

**Table 33:** Mapping from paper tables to benchmark artifacts. All paths are relative to the run directory (output/run\_3db6c1\_20260304\_222040/).

Paper Table	Artifact file	Key fields
3	data/metric_table.csv	h_ret, disp_qlike, svi, pue
7	data/results_all.json → hypotheses.{ID}	statistic.value, p_value.raw, n_obs
8	data/results_all.json → hypotheses.{ID}	p_value.holm_l2, p_value_bh, ci_*
10	supplementary/concern2*/concern2_results.json	class_level_results
14	run.log (grep “Subgroup”)	per-cell pass/fail
12	data/results_all.json → diagnostics.solver_mcs_rates	mcs_inclusion_rate
13	data/solver_table.csv	solver_name, qlike, converged
9	supplementary/concern1*/concern1_results.json	panel_regression
11	supplementary/concern10*/h3_heatmap_data.json	rho_matrix, p_matrix
15	supplementary/concern3*/concern3_results.json	phase.a.existing_data
16	supplementary/concern3*/concern3_results.json	phase.b.mcs_comparisons
17	data/results_all.json → diagnostics.c4_*	coefficients
18	supplementary/concern4_convergence/concern4_results.json	converged-only robustness
19	supplementary/concern6*/concern6_results.json	tercile_results
20	supplementary/concern7*/concern7_results.json	per_horizon
21	supplementary/concern9*/u2_model_selection_results.json	solver_breakdown
22	supplementary/concern9*/u2_model_selection_results.json	per-tercile ΔQLIKE
27	supplementary/concern5*/concern5_results.json	ols_regression
28	supplementary/concern5*/concern5_results.json	economic_calibration
23	data/solver_table.csv + metric_table.csv	qlike per solver vs qlike_best
29	data/metric_table.csv	PUE threshold sweep (5%–50%)
24	supplementary/concern21_fhs_attribution/concern21_results.json	S21 2×2 VaR attribution
25	supplementary/concern23_var_by_pue/concern23_results.json	VaR pass rates by PUE regime
—	supplementary/concern15_var_quantile/concern15_results.json	S15 VaR quantile comparison
—	supplementary/concern16_var_timing/concern16_results.json	S16 timing mechanism
—	supplementary/concern17_npe_analysis/concern17_results.json	S17 NPE continuous vs binary
—	supplementary/concern18_h3_persistence/concern18_results.json	S18 H3 persistence diagnostics
—	supplementary/concern19_h1_granular/concern19_results.json	S19 H1 granular confound controls
—	supplementary/concern20_economic_value/concern20_results.json	S20 Christoffersen + joint coverage
—	supplementary/concern22_dynamic_switching/concern22_results.json	S22 dynamic entropy switching OOS
—	supplementary/concern23_var_by_pue/concern23_results.json	S23 VaR pass rates by PUE regime
—	supplementary/concern24_oos_combination/concern24_results.json	S24 OOS holdout: EWMA vs COMB vs selection
—	supplementary/concern25_skewt_sensitivity/concern25_results.json	S25 skewed- <i>t</i> innovation sensitivity
30	supplementary/concern26_walkforward_var/concern26_results.json	S26 walk-forward VaR backtesting
—	paper/figures/fig05_h2_scatter_hexbin.png	S27 hexbin Figure 5
—	supplementary/concern28_u2_diagnostic/concern28_results.json	S28 U2 power diagnostic

## B Hypothesis Specification Details

Each hypothesis carries a full specification:

1. Formal null and alternative hypotheses
2. Direction rationale (with citation to conflicting theories)
3. Primary and backup test statistics
4. Pass/fail criteria (raw and corrected thresholds)
5. Known confounds with specified controls
6. Robustness battery (alternative metrics, within-class, sensitivity to hyperparameters)
7. Output schema (typed fields logged per hypothesis)

The complete specifications are in `spec/HYPOTHESIS_SPEC.md` and locked by the spec manifest.

## C Metric Definitions and Runtime Invariants

Runtime-enforced invariants on all computed metrics:

1.  $0 \leq \hat{H}_{\text{MM}} \leq \ln K + (K - 1)/(2n)$  (entropy range)
2.  $0 \leq \text{NPE} \leq 1$  (normalized permutation entropy)
3.  $\mathcal{L}_{\text{QLIKE}} \geq 0$  (QLIKE non-negativity)
4.  $0 \leq \text{SVI} \leq 1$  (selection value index)
5.  $\text{PUE} < 100$  (pattern utilization efficiency upper bound)
6.  $\text{spread\_qlike} \geq 0$  (loss spread non-negativity)
7. URI reported only when empirical coverage  $\geq 0.935$  at 95% nominal (coverage gate)
8.  $n_{\text{obs}} \geq 252$  per asset (minimum sample gate)
9.  $|\mathcal{M}_i| \geq 3$  per asset (minimum solver gate)

Any invariant violation at runtime triggers an immediate exception with a diagnostic message, preventing silent corruption of downstream metrics.