

Article

Finite-Sample Precision Limits for Expected Shortfall Forecast Comparisons

Daniel Traian Pele ^{1,2,3,*}  and Miruna Mazurencu-Marinescu-Pele ¹

¹ Department of Statistics and Econometrics, Faculty of Cybernetics, Statistics and Economic Informatics, Bucharest University of Economic Studies, 010374 Bucharest, Romania; miruna@ase.ro

² Institute for Economic Forecasting, Romanian Academy, 050711 Bucharest, Romania

³ IDA Institute of Digital Assets, Bucharest University of Economic Studies, 010374 Bucharest, Romania

* Correspondence: danpele@ase.ro

Abstract

Expected shortfall (ES) is a tail functional whose estimation precision is governed by the effective tail sample size $n\alpha$ rather than by the nominal calibration size n . The resulting $(n\alpha)^{-1/2}$ information limit is well established, yet no practical framework exists for deciding whether two ES forecasts can be meaningfully distinguished over a finite calibration window. This paper converts the asymptotic rate into four operational diagnostics: a plug-in precision benchmark, a sample-size rule, a precision-fragile pairwise comparison screen, and a VaR-first diagnostic linking excess ES dispersion to first-stage quantile miscalibration. An empirical application to global financial assets and heterogeneous forecasters under standard regulatory tail parameters shows that roughly one in five pairwise ES comparisons is precision-fragile, with excess dispersion concentrated in cells with poor VaR calibration. The results suggest that ES forecast rankings at typical tail levels can be constrained by effective tail information rather than by model sophistication.

Keywords: effective tail sample size; expected shortfall; precision-fragile comparison; finite-sample precision; sample-size rule; Fissler–Ziegel score; VaR miscalibration diagnostic

MSC: 62G32; 62P05; 91G70

1. Introduction

Expected shortfall (ES) at tail level α is a coherent risk measure widely used in financial regulation, portfolio management, and quantitative risk modelling. Because ES averages losses beyond the α -quantile, its estimation draws on at most $n\alpha$ effective tail observations from a calibration sample of size n . At tail levels common in practice— $\alpha \in \{1\%, 2.5\%, 5\%\}$ —this effective count is small even when n is moderately large. The statistical difficulty of ES estimation is therefore governed by $n\alpha$, not by n .

The $(n\alpha)^{-1/2}$ scaling of ES estimation error is well established. Zwingmann and Holzmann [1] derive this rate through a central limit theorem for tail averages, Bartl and Eckstein [2] obtain a concentration lower bound for nonparametric ES estimation, and related kernel estimators [3] obey the same effective-tail-count rate. These results characterise the information limit, but they do not translate it into a practical diagnostic for deciding whether two ES forecasts can be meaningfully distinguished over a finite calibration window.

ES forecast evaluation relies on the joint elicibility of VaR and ES [4]. Two-stage regression frameworks [5] and dynamic semiparametric VaR–ES models [6] provide tools



Academic Editors: Alessandro Mazzocchi, Raffaele Mattera and J.E. Trinidad-Segovia

Received: 1 June 2026

Revised: 26 June 2026

Accepted: 27 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

for estimation and evaluation under dependence. Pele et al. [7] show that Fissler–Ziegel-based recalibration can attain the $(n\alpha)^{-1/2}$ rate under geometric mixing. That paper is concerned with achievability of the rate under temporal dependence; the present paper takes the rate as given and converts it into an operational comparison audit comprising the four diagnostics developed in Sections 3–4. ES backtesting from a regulatory perspective is studied by Acerbi and Székely [8] and Nolde and Ziegel [9]. The coherence of ES as a risk measure is established by Acerbi and Tasche [10], and a substantial recent literature develops regression- and tail-risk-based ES backtests [11,12] and semiparametric VaR–ES forecasting methods [13]; these procedures test the calibration or relative accuracy of a given forecast, whereas the present paper asks the prior question of whether the available tail data can support any such comparison at all. The question addressed here is how much the pairwise ES model comparison is statistically supportable once the effective tail count is fixed by the calibration window.

The contribution is operational rather than rate-theoretic. The paper converts the known ES information limit into a precision audit for ES model comparison. The audit has four components:

1. a plug-in precision benchmark based on the effective tail count, calibrated from tail residuals;
2. a finite-sample sample-size rule that determines the calibration window needed for a target ES precision tolerance;
3. a precision-fragile pairwise comparison screen that classifies a forecast pair as precision-fragile when the observed difference in ES recalibration corrections falls below the precision floor implied by the available effective tail count;
4. a VaR-first diagnostic for attributing excess ES recalibration dispersion to first-stage quantile miscalibration.

The novelty is not the $(n\alpha)^{-1/2}$ rate itself. The novelty is the conversion of this rate into a practical precision audit for ES model comparison. The Le Cam two-point construction in Appendix A (Theorem A1) certifies that no estimator can beat the $(n\alpha)^{-1/2}$ rate over the distribution class, establishing that the plug-in benchmark $\hat{C}/\sqrt{n\alpha}$ has the correct functional form for a precision floor. The operational constant $\hat{C}$ is a CLT-motivated plug-in rather than the exact minimax constant c_L , but the rate it instantiates is sharp.

The natural application domain for this audit is the Fundamental Review of the Trading Book (FRTB, [14]), which sets ES at $\alpha = 2.5\%$ as the capital-determining market-risk measure evaluated over a 250-day calibration window. These parameters imply $n\alpha = 6.25$, placing the problem firmly in the finite-sample tail-scarcity regime. FRTB serves as the motivating application throughout this paper, but the audit framework applies to any setting in which ES forecasts are compared over a finite window at a fixed tail level.

The empirical analysis uses a parametric GJR-GARCH- t baseline [15] and three zero-shot time-series foundation models: TimesFM 2.5 [16], Chronos-Small [17], and Moirai 2.0 [18,19]. The forecasters are not ranked as a model-selection exercise. They provide heterogeneous first-stage forecast sequences against which the precision audit can be evaluated. Chronos-Small is retained as a stress case because it exhibits severe first-stage VaR miscalibration, allowing the VaR-first diagnostic to be evaluated under a clear failure mode.

Across 24 global assets and four forecasters evaluated at $\alpha = 2.5\%$, between 16% and 29% of pairwise ES rankings fall below the plug-in precision tolerance implied by the effective tail count; the paired block-bootstrap estimate is 22.9%. Excess ES recalibration dispersion is concentrated in cells with poor first-stage VaR calibration. For the median asset, the 250-day calibration window delivers approximately 37 bp of precision, but seven of 24 high-volatility assets require more than 250 days to reach a 50 bp tolerance.

From a mathematical-statistical perspective, the paper studies how an asymptotic information bound for a tail functional can be converted into a finite-sample diagnostic for forecast comparison. The empirical application to market-risk forecasts serves to illustrate the consequences of this effective-sample-size constraint in a realistic dependent time-series setting.

The remainder of the paper is organised as follows. Section 2 develops the theoretical framework, linking the effective tail sample size to ES recalibration through an oracle-equivalence argument. Section 3 defines the plug-in precision benchmark and the operational sample-size rule. Section 4 introduces the precision-fragile comparison screen and the VaR-first diagnostic. Section 5 reports a controlled VaR-miscalibration simulation. Section 6 presents the financial-risk application. Section 7 collects robustness and sensitivity results. Section 8 states the main limitations, and Section 9 concludes.

2. Expected Shortfall as a Tail Functional Under Effective Sample-Size Scarcity

This section establishes the theoretical framework. It defines the effective tail sample size, derives the CLT-based precision scaling, and states the oracle-equivalence argument that transfers the rate to additive ES recalibration.

2.1. Definitions and Effective Tail Sample Size

Let $X_1, \dots, X_n$ be observations from a distribution P on $\mathbb{R}$ with Lebesgue density f . Fix $\alpha \in (0, 1/2)$. The value-at-risk at level α is the α -quantile $\text{VaR}_\alpha(P) := \inf\{x : F(x) \geq \alpha\}$, and the expected shortfall is $\text{ES}_\alpha(P) := \alpha^{-1} \int_0^\alpha \text{VaR}_u(P) du = \mathbb{E}[X \mid X \leq \text{VaR}_\alpha(P)]$. Both VaR_α and ES_α are negative for losses throughout; the FZ₀ loss in Section 6.1 assumes $e < 0$ accordingly.

The $(n\alpha)^{-1/2}$ rate for ES estimation is established by Zwingmann and Holzmann [1] via a CLT for the tail average and by Bartl and Eckstein [2] via a concentration lower bound. The mechanism is the tail identification residual

$$\phi(X) := (\text{VaR}_\alpha - X) \mathbb{1}\{X \leq \text{VaR}_\alpha\} / \alpha - (\text{ES}_\alpha - \text{VaR}_\alpha), \quad (1)$$

which satisfies $\mathbb{E}[\phi] = 0$ and $\text{Var}(\phi) = \sigma_{\text{tail}}^2 / \alpha$, where $\sigma_{\text{tail}}^2 := \mathbb{E}[(\text{VaR}_\alpha - X)^2 \mid X \leq \text{VaR}_\alpha]$. Only an α -fraction of observations falls in the tail, and each carries $O(1/\alpha)$ weight in the ES average; the CLT for the sample average of $\phi(X_1), \dots, \phi(X_n)$ gives standard error $\sigma_{\text{tail}} / \sqrt{n\alpha}$. The effective sample size for ES estimation is therefore $n\alpha$, not n .

2.2. Oracle Equivalence for Additive Recalibration

In the two-stage recalibration framework of Dimitriadis and Bayer [5], a first-stage model produces base VaR and ES forecasts, and the second stage estimates an additive ES correction $r^* := \text{ES}_\alpha(P) - \bar{E}$ by minimising a Fissler–Ziegel loss [4].

Remark 1 (Oracle equivalence). Fix a first-stage ES forecast $\bar{E}$ that does not depend on the second-stage calibration sample $X_1, \dots, X_n$. Estimating $r^*(P) := \text{ES}_\alpha(P) - \bar{E}$ is equivalent in L^1 risk to estimating $\text{ES}_\alpha(P)$: since $\bar{E}$ is a fixed constant, $\mathbb{E}_P|\hat{r}_n - r^*(P)| = \mathbb{E}_P|\hat{T}_n - \text{ES}_\alpha(P)|$ for $\hat{T}_n := \hat{r}_n + \bar{E}$. Any minimax lower bound for ES estimation therefore transfers directly to $r^*(P)$.

The conceptual chain of the paper has four links, summarised schematically in Appendix F (Figure A1): (i) the known $(n\alpha)^{-1/2}$ rate for ES estimation; (ii) oracle equivalence (Remark 1), by which additive ES recalibration inherits this rate under a fixed first-stage forecast; (iii) the plug-in precision benchmark $\hat{C} / \sqrt{n\alpha}$, which instantiates the rate empiri-

cally; and (iv) the operational diagnostics built on this benchmark (Sections 6 and 7). The novelty lies in links (iii)–(iv), not in the rate itself.

2.3. Distinction Between c_L and $\hat{C}$

An alternative Le Cam two-point construction yielding a closed-form constant c_L is provided in Appendix A for completeness. The plug-in precision benchmark $\hat{C}/\sqrt{n\alpha}$ used in the empirics is calibrated from the CLT, not from c_L . The distinction matters: c_L is the exact minimax constant from a worst-case two-point argument, whereas $\hat{C}$ is a CLT-motivated plug-in scale. Throughout, “information limit” and “precision floor” refer to the $(n\alpha)^{-1/2}$ rate; “plug-in benchmark” refers to $\hat{C}/\sqrt{n\alpha}$.

The fixed-forecast condition in Remark 1 is approximated in the empirical design: the 1000-day GJR-GARCH training window ends before the 250-day calibration window, and the foundation-model weights are not updated on the calibration sample (zero-shot inference). The benchmark is therefore an oracle-style precision diagnostic, not a formal finite-sample guarantee for the full rolling procedure. The operational benchmark $\hat{C}/\sqrt{n\alpha}$ inherits the rate from oracle equivalence; it does not inherit the exact Le Cam constant c_L . When the first stage is mis-specified, the first-stage bias η_t does not vanish and the oracle equivalence breaks down—precisely the VaR-miscalibration channel that the VaR-first diagnostic detects (Section 4.2).

3. Precision Benchmark and Sample-Size Rule

3.1. Plug-In Precision Benchmark

For each asset, forecaster, and tail level, the plug-in constant is defined as

$$\hat{C} = \hat{\sigma}_{\text{tail}}, \quad (2)$$

the sample standard deviation of the tail residuals

$$\{\hat{V}_t - X_t : X_t \leq \hat{V}_t\}. \quad (3)$$

The corresponding precision benchmark is

$$B_{i,m,\alpha} = \frac{\hat{C}_{i,m,\alpha}}{\sqrt{n\alpha}}. \quad (4)$$

This quantity instantiates the effective-tail-count rate in an empirical cell and provides the precision floor against which ES recalibration dispersion is evaluated. The benchmark is an operational diagnostic, not a formal finite-sample confidence interval.

In the main analysis, $\hat{C}$ is computed once over the full evaluation sample for each cell. This yields an ex-post diagnostic benchmark, not a real-time estimate. In a live implementation, $\hat{C}$ would instead be estimated on each rolling calibration window. Appendix C reports this rolling-window variant and shows that the qualitative precision-fragile comparison results are preserved, although the absolute benchmark ratios change.

Because $\hat{C}$ is a sample standard deviation of tail residuals, it carries its own sampling variance, and the precision floor $B = \hat{C}/\sqrt{n\alpha}$ inherits this uncertainty. The concern is most acute in the rolling-window variant, where $\hat{C}_t$ is re-estimated on each 250-day window from roughly $n\alpha \approx 6.25$ tail points. The main analysis mitigates the problem by computing $\hat{C}$ once over the full evaluation sample, which aggregates far more than 6.25 tail observations per cell. The rolling-window results in Appendix C (Table A4) confirm that the qualitative conclusions survive when $\hat{C}$ is instead estimated window-by-window, even though individual benchmark ratios shift—consistent with the additional noise in the floor

estimate. A Monte Carlo study in Appendix B (Table A1) confirms that the coefficient of variation of $\hat{C}$ is 0.76 under Student- t_5 at FRTB parameters, declining to 0.43 at $n = 1000$. Where the floor itself is uncertain, marginal precision-fragile classifications should be treated as tentative rather than decisive, reinforcing the conservative reading of the screen.

3.2. Empirical Dispersion Measure

The raw rolling-window ES corrections $\hat{r}_n$ contain both local estimation noise and slow-moving volatility-regime shifts. The precision benchmark concerns the former. To isolate within-regime estimation variability, we subtract a 252-day moving average from the correction path before computing its standard deviation; Appendix G (Figure A2) illustrates this decomposition for a representative asset. Robustness checks with 126- and 504-day filters in Appendix C show that the forecaster ranking is stable, although absolute ratios vary with the detrending horizon.

For each asset i , forecaster m , and tail level α , the empirical benchmark ratio is

$$R_{i,m,\alpha} = \frac{\widehat{SD}^{\det}(\hat{r}_{i,m,\alpha})}{\widehat{B}_{i,m,\alpha}}, \quad (5)$$

where $\widehat{B}_{i,m,\alpha}$ is the plug-in precision benchmark from Equation (4), and $\widehat{SD}^{\det}$ denotes the standard deviation of the detrended correction path. Values $R < 1$ indicate that the plug-in scale is conservative for that cell; this does not contradict the Le Cam lower bound, which uses a different constant c_L . Values $R > 1$ indicate excess dispersion, potentially due to VaR misalignment, non-stationarity, or forecast-extraction noise.

3.3. Finite-Sample Correction

At FRTB parameters ($\alpha = 2.5\%$, $n = 250$), the effective tail count is $k := n\alpha = 6.25$. Because the number of tail observations is random—binomial(n, α) under correct VaR calibration—the CLT scale requires an inflation factor:

$$f(n, \alpha) = \sqrt{1 + \frac{1 - \alpha}{n\alpha}}. \quad (6)$$

At FRTB parameters, $f = 1.075$, raising the required window by approximately 15%. Monte Carlo calibration (Table 1, 50,000 replications under Student- t_5 and GARCH(1,1)- t_5 DGPs) shows that the CLT plug-in is accurate to roughly 10–15% at $k = 6.25$ under i.i.d. sampling. Under GARCH dynamics, the unconditional σ_{tail} can exceed the within-regime value at short windows, producing $\hat{f} < 1$; this motivates the detrending step in Section 3.2. Full details are in Appendix B.

Table 1. Finite-sample inflation factor $\hat{f} = \widehat{SD}(\bar{X}_{\text{tail}}) / (\sigma_{\text{tail}} / \sqrt{n\alpha})$ from 50,000 Monte Carlo replications. †: $\hat{f} > 1.20$, indicating serial-dependence inflation beyond the i.i.d. finite-sample effect.

α	n	$k = n\alpha$	Student- t_5	Skewed- t_5	GARCH- t_5
1.0%	250	2.50	1.129	1.134	0.653
	500	5.00	1.124	1.119	0.918
	1000	10.00	1.063	1.044	1.080
	2000	20.00	1.032	1.022	1.217 †
2.5%	250	6.25	1.101	1.117	1.006
	500	12.50	1.050	1.043	1.150
	1000	25.00	1.028	1.011	1.284 †
	2000	50.00	1.012	1.008	1.441 †

Table 1. *Cont.*

α	n	$k = n\alpha$	Student- t_5	Skewed- t_5	GARCH- t_5
5.0%	250	12.50	1.039	1.042	1.145
	500	25.00	1.022	1.016	1.339 [†]
	1000	50.00	1.003	1.005	1.494 [†]
	2000	100.00	1.006	1.008	1.587 [†]

3.4. Operational Sample-Size Rule

The Le Cam constant c_L in Appendix A depends on local density geometry and is not directly estimable. For operational sample-size calculations, we therefore instantiate the $(n\alpha)^{-1/2}$ rate with the CLT-motivated plug-in scale $C = \hat{\sigma}_{\text{tail}}$, estimated from tail residuals (see Chapter 6, [20]). Incorporating the random-count correction from Section 3.3, the required window length for tolerance ε solves

$$\frac{f(n, \alpha) C}{\sqrt{n\alpha}} \leq \varepsilon \quad \Longleftrightarrow \quad n \geq \frac{f(n, \alpha)^2}{\alpha} \left(\frac{C}{\varepsilon} \right)^2, \quad (7)$$

where

$$f(n, \alpha) = \sqrt{1 + \frac{1 - \alpha}{n\alpha}}. \quad (8)$$

Because $f(n, \alpha)$ depends on n , the rule is implicit; in practice, fixed-point iteration converges in a few steps. At $\alpha = 2.5\%$ and $n = 250$, the correction factor is $f(250, 0.025) = 1.075$, which raises the required window by approximately 15%. For $k = n\alpha \geq 25$, the correction is below 2% and can usually be ignored.

Table 2 reports the implied window lengths for four tail levels and four tolerances, calibrated from a Student- t_5 reference distribution. The table has two immediate implications. First, sub-FRTB tail levels are effectively infeasible on realistic calibration windows: at $\alpha = 0.5\%$ and $\varepsilon = 10\%$, the required window is about 44,700 trading days. Second, at the FRTB level $\alpha = 2.5\%$, a 250-day window implies a corrected tolerance of approximately 0.43 C, which is of the same order as many inter-model ES differences.

Table 2. Minimum calibration window n (trading days) for tolerance ε at tail level α . Cells are computed from Equation (7), with C calibrated from a Student- t_5 reference distribution and $f(n, \alpha)$ from Equation (8). The implicit equation is solved by fixed-point iteration. ε is expressed in standardised units.

α	C	$\varepsilon = 10\%$	20%	30%	50%
0.5%	1.491	44,660	11,311	5132	1958
1.0%	1.335	17,921	4553	2075	800
2.5%	1.152	5348	1365	627	246
5.0%	1.038	2174	558	257	102

Asset-specific calibrations are reported in Table A11 in Appendix I and summarised in Section 6. Empirical ratios should be interpreted as distances to the CLT-motivated plug-in benchmark, not as tests of the exact Le Cam constant.

4. Precision-Fragile Pairwise Comparison Screen

4.1. Screen Definition

We call a comparison precision-fragile if its absolute ES-recalibration difference is smaller than the plug-in tolerance. The screen is not a hypothesis test and does not attach

a nominal size. It is a conservative diagnostic: it asks whether the observed difference is larger than the precision floor implied by the available effective tail count.

The threshold is forecaster-agnostic, conditional on approximate first-stage calibration: it depends on the per-cell tail dispersion $\hat{C}$ and the effective tail count $n\alpha$, not on the model class itself. Adding or substituting forecasters changes the set of pairwise comparisons, but not the precision floor used to evaluate each pair. Severely miscalibrated forecasters can still inflate the precision-fragile share through the VaR-miscalibration channel, which the VaR-first diagnostic examines in Section 6.3.

Formally, for a given asset and tail level α , the comparison between forecasters 1 and 2 is classified as precision-fragile when

$$\left| \bar{r}_n^{(1)} - \bar{r}_n^{(2)} \right| < \frac{\sqrt{\hat{C}_1^2 + \hat{C}_2^2}}{\sqrt{n\alpha}}, \quad (9)$$

where $\bar{r}_n^{(j)}$ denotes the mean ES recalibration correction for forecaster j over the relevant evaluation windows. The threshold treats the two correction estimates as independently noisy. Because the same return realisation enters both forecasters' recalibration losses, pairwise correction estimates are likely positively correlated. Ignoring this covariance inflates the tolerance and therefore classifies marginal cases as precision-fragile; the screen is conservative by design.

Although the algebraic form of the threshold resembles a pooled standard error, its substantive content differs in three respects. First, the scale is anchored to the effective tail count $n\alpha$ —6.25 observations at FRTB parameters, not the 250 days that constitute the nominal window—so the tolerance reflects the information actually available in the tail rather than the total calibration length. Second, the precision floor is a fixed comparison standard: it is invariant to the forecaster set, whereas the realised precision-fragile share depends on which models are compared. Third, the screen is offered as a conservative diagnostic, not a hypothesis test with nominal size; it flags comparisons whose observed difference is smaller than the precision floor, without attaching a rejection probability.

4.2. Rate Tests and VaR-First Diagnostic

Three diagnostics assess whether the $(n\alpha)^{-1/2}$ rate governs empirical recalibration dispersion, and whether deviations from the benchmark are linked to first-stage quantile failure.

First, a cross-sectional scaling regression across all (asset, forecaster, α) cells:

$$\log\left(\widehat{\text{SD}}^{\text{det}}(\hat{r}_n)\right) = a + b \log(n\alpha) + \gamma \log(\hat{\sigma}_{\text{tail}}) + u. \quad (10)$$

The rate prediction is $b = -1/2$, while the CLT plug-in scale implies $\gamma = 1$. Standard errors are clustered by asset.

Second, a non-overlapping window-length scaling test at $\alpha = 1\%$, where variation in the effective tail count is most visible. The calibration length is varied over

$$n \in \{250, 500, 750, 1000\}, \quad (11)$$

with step size equal to n , so adjacent windows do not overlap. For each (asset, forecaster) pair, we estimate

$$\log\left(\widehat{\text{SD}}(\hat{r}_n)\right) = a + b \log(n) + u, \quad (12)$$

using HC1 robust standard errors. The pooled fixed-effects specification is

$$\log(\widehat{\text{SD}}(\hat{r}_{i,j,n})) = \alpha_i + \beta_j + b \log(n) + u_{i,j,n}, \quad (13)$$

where α_i and β_j denote asset and forecaster fixed effects.

Third, the VaR-first diagnostic distinguishes intrinsic ES estimation noise from first-stage quantile failure. For each (asset, forecaster) pair, VaR backtesting statistics are computed at $\alpha = 1\%$. The main diagnostic uses the Christoffersen conditional-coverage statistic [21]; the Kupiec unconditional-coverage statistic [22] is reported as a robustness check. Computational details of the conditional-coverage statistic and the associated Fisher exact test are given in Appendix E. Excess ES recalibration dispersion should concentrate in cells with poor first-stage VaR calibration.

5. Simulation Evidence

This section isolates the mechanism behind the VaR-first diagnostic. The empirical results in Section 6 show that excess ES recalibration dispersion is concentrated in poorly VaR-calibrated cells. The simulation below asks whether this pattern can arise when the first-stage VaR model fails to track time-varying volatility.

The simulation generates 30 independent paths of length 10,000 from a GARCH(1,1)- t_5 process with $\omega = 10^{-6}$, $\alpha_g = 0.10$, and $\beta_g = 0.85$. For each path, the first-stage VaR-ES model uses a convex blend of the true conditional volatility σ_t and the unconditional volatility $\bar{\sigma}$:

$$\sigma_{\text{model},t} = (1 - \delta) \sigma_t + \delta \bar{\sigma}, \quad \delta \in \{0, 0.70, 0.85, 1\}. \quad (14)$$

At $\delta = 0$, the model uses the true conditional volatility. At $\delta = 1$, it ignores conditional volatility dynamics and uses the unconditional volatility. Additive ES corrections are estimated on rolling 250-day windows, advanced in 21-day steps, at $\alpha = 2.5\%$. The ES dispersion ratio R is computed as the raw standard deviation of the ES correction divided by the oracle benchmark

$$\frac{\sigma_{\text{tail}}}{\sqrt{n\alpha}}. \quad (15)$$

Table 3 shows that R increases from 0.95 under the oracle volatility specification to 1.76 under the unconditional-volatility specification. The hit rate rises at the same time, indicating worsening VaR calibration. The transition through $R = 1$ occurs between the $\delta = 0.70$ and $\delta = 0.85$ designs, where the hit rate is already far above the nominal 2.5%. The simulation therefore supports the interpretation that regime-dependent VaR error can inflate ES recalibration dispersion, consistent with the VaR-first diagnostic.

Table 3. VaR-miscalibration simulation. The data-generating process is GARCH(1,1)- t_5 with $\alpha = 2.5\%$ and $n = 250$, using 30 paths of 10,000 days each. R is the ratio of raw ES-correction SD to the oracle benchmark $\sigma_{\text{tail}} / \sqrt{n\alpha}$.

VaR Model	Hit Rate	Median UC Stat	ES Dispersion Ratio (R)
Correct (true σ_t)	0.0250	0.5	0.95
Mild (30% cond. + 70% uncond.)	0.0961	23.9	0.97
Moderate (15% cond. + 85% uncond.)	0.1314	45.9	1.33
Severe (unconditional $\bar{\sigma}$)	0.1770	80.0	1.76

6. Financial-Risk Application

This section illustrates the precision-audit framework on 24 global financial assets and four forecasters under FRTB calibration parameters. The forecasters are not ranked as a

model-selection exercise; they provide heterogeneous first-stage forecast sequences against which the precision audit can be evaluated.

6.1. Data and Forecasting Setup

The empirical analysis uses daily log returns for 24 global assets: equity indices, bonds, commodities, cryptocurrencies, and currencies. The sample period is 2002–2026, with exact start and end dates varying by asset; details are reported in Appendix C. Forecasts are computed at $\alpha \in \{1\%, 2.5\%, 5\%\}$, with $\alpha = 2.5\%$ as the primary ES analysis level. Additive Fissler–Ziegel recalibration is performed on rolling $n = 250$ -day calibration windows, advanced in monthly steps of 21 trading days.

The parametric benchmark is a GJR-GARCH(1,1) model with Student- t innovations [15], estimated by maximum likelihood on a rolling 1000-day training window, with VaR and ES computed analytically from the fitted conditional distribution. The GARCH-family specification is motivated both by the conditional heteroskedasticity of the return series and by the literature. Engle’s ARCH-LM test [23] rejects the no-ARCH null at the 1% level for all 24 assets (Appendix K, Table A18), confirming that conditional-variance dynamics must be modelled. The GARCH model of Bollerslev [24] and its leverage-augmented GJR variant [15] are the standard parametric tools for this purpose, and the GARCH(1,1) lag order is supported by the evidence in Hansen and Lunde [25] that richer volatility models rarely outperform it out of sample. The GJR variant is retained because it additionally captures the well-documented asymmetric (leverage) response of equity and commodity volatility. The remaining forecasters are time-series foundation models run zero-shot. TimesFM 2.5 [16] uses nine quantile heads with a Student- t fit by quantile matching. Chronos-Small [17] generates 1000 Monte Carlo forecast samples, from which empirical quantiles and conditional tail means yield VaR and ES directly. Moirai 2.0 [18,19] produces 1000 forecast samples with a Student- t fit to extract VaR and ES. Implementation details of the extraction procedures are reported in Appendix C.

Given base forecasts $(\hat{V}_t, \hat{E}_t)$, the second stage estimates an additive correction pair $(\hat{q}_n, \hat{r}_n)$ by minimising the Fissler–Ziegel FZ_0 loss [4,5]. The corrected forecasts are

$$\tilde{V}_t = \hat{V}_t + \hat{q}_n, \quad \tilde{E}_t = \hat{E}_t + \hat{r}_n, \quad (16)$$

where $\hat{q}_n$ and $\hat{r}_n$ are held fixed within each calibration window. The loss function is

$$S(v, e, x) = \frac{\mathbb{1}\{x \leq v\}(x - v)}{\alpha e} + \frac{v}{e} + \log(-e) - 1, \quad e < 0. \quad (17)$$

The optimisation uses Nelder–Mead with three random restarts per window. Convergence failure is below 0.5% across all cells.

6.2. Precision-Fragile ES Comparisons

At the FRTB ES level $\alpha = 2.5\%$, 29 of 144 pairwise comparisons are precision-fragile under the plug-in screen in Equation (9), corresponding to a share of 20.1% (Table 4). The share increases as the tail level becomes more extreme, consistent with the $(n\alpha)^{-1/2}$ precision limit.

The 20.1% headline should be interpreted as a baseline diagnostic, not as a universal constant. Table 5 reports the precision-fragile share under alternative noise scales and detrending choices. At $\alpha = 2.5\%$, the share ranges from 16.0% to 29.2%, with the paired block bootstrap at 22.9%.

Table 4. Precision-fragile ES model comparisons. Each tail level has $\binom{4}{2} \times 24 = 144$ pairwise comparisons.

α	Comparisons	Fragile	% Fragile
1.0%	144	39	27.1%
2.5%	144	29	20.1%
5.0%	144	26	18.1%

Table 5. Precision-fragile share under alternative noise scales and detrending methods. The plug-in screen uses Equation (9). The paired bootstrap uses block length 21 days and 1000 replications.

Noise Scale	Fragile @ $\alpha = 1\%$	Fragile @ $\alpha = 2.5\%$	Fragile @ $\alpha = 5\%$
Plug-in $\hat{C}/\sqrt{n\alpha}$	27.1%	20.1%	18.1%
Detrended SD: MA-126	16.0%	16.0%	15.3%
Detrended SD: MA-252	17.4%	17.4%	15.3%
Detrended SD: MA-504	20.8%	17.4%	17.4%
Detrended SD: HP ($\lambda = 1600$)	17.4%	17.4%	16.0%
Detrended SD: Rolling median 252	18.1%	17.4%	15.3%
Raw SD (no detrending)	36.8%	29.2%	25.0%
Paired block bootstrap (block = 21)	—	22.9%	—

Because the precision-fragile share is itself an estimate subject to sampling variability, we attach a formal confidence interval to it (Table 6). A cluster bootstrap that resamples the 24 assets with replacement—thereby preserving the within-asset dependence across the six pairwise comparisons and the noise in $\hat{C}$ —yields, at $\alpha = 2.5\%$, a 95% interval of [16.0%, 25.0%] around the 20.1% point estimate (10^4 replications). A paired moving-block bootstrap over the rolling recalibration windows (block length 21, 10^4 replications), which instead resamples the time dimension that drives the floor sampling noise $CV(\hat{C}) = 0.76$, gives a consistent interval of [20.1%, 25.7%] around a mean of 22.9%. At $\alpha = 1\%$ and 5% the corresponding intervals are [20.8%, 34.0%] and [15.3%, 21.5%]. The intervals confirm that a non-negligible share of ES comparisons—of the order of one in five—is precision-fragile at FRTB parameters, while making explicit that the point estimate carries a width of roughly ± 5 percentage points; quantitative interpretation of any single share should respect this band.

Table 6. Precision-fragile share with bootstrap confidence intervals. Point shares use the plug-in screen, Equation (9). The cluster bootstrap resamples the 24 assets (10^4 replications); percentile 95% intervals are reported.

Tail Level α	Fragile/Total	Point Share	95% Cluster-Bootstrap CI
1.0%	39/144	27.1%	[20.8%, 34.0%]
2.5%	29/144	20.1%	[16.0%, 25.0%]
5.0%	26/144	18.1%	[15.3%, 21.5%]

Paired moving-block bootstrap at $\alpha = 2.5\%$ (block 21, 10^4 reps): mean 22.9%, 95% CI [20.1%, 25.7%].

Table 7 summarises the three main choices for estimating the precision scale at $\alpha = 2.5\%$. The qualitative conclusion is unchanged: a material fraction of ES comparisons is smaller than the precision floor supported by the available tail data.

The sample-size rule reveals substantial cross-asset heterogeneity. At $\alpha = 2.5\%$, the estimated tail-dispersion scale spans more than an order of magnitude: USD/JPY requires roughly 84 trading days for 50 bp precision, whereas BTC requires roughly 4900 days. For the median asset, the 250-day calibration window delivers approximately 37 bp precision.

Seven of 24 assets require more than 250 days to reach a 50 bp tolerance, so the sample-size rule is as much an asset-selection diagnostic as a window-length diagnostic.

Table 7. Precision-fragile share under alternative $\hat{C}$ estimation methods at $\alpha = 2.5\%$.

$\hat{C}$ Estimation	Precision-Fragile Share
Full-sample $\hat{C}$	20.1%
Rolling-window $\hat{C}_t$	16.0%
Paired block bootstrap	22.9%

6.3. Empirical Diagnostics

Figure 1 compares the detrended standard deviation of the ES correction $\hat{r}_n$ with the plug-in precision benchmark $\hat{C}/\sqrt{n\alpha}$ across all (asset, forecaster, α) cells. The 45-degree line corresponds to $R = 1$. Points below the line indicate that the plug-in benchmark is conservative for that cell; points above the line indicate excess dispersion, which is examined through the VaR-first diagnostic below.

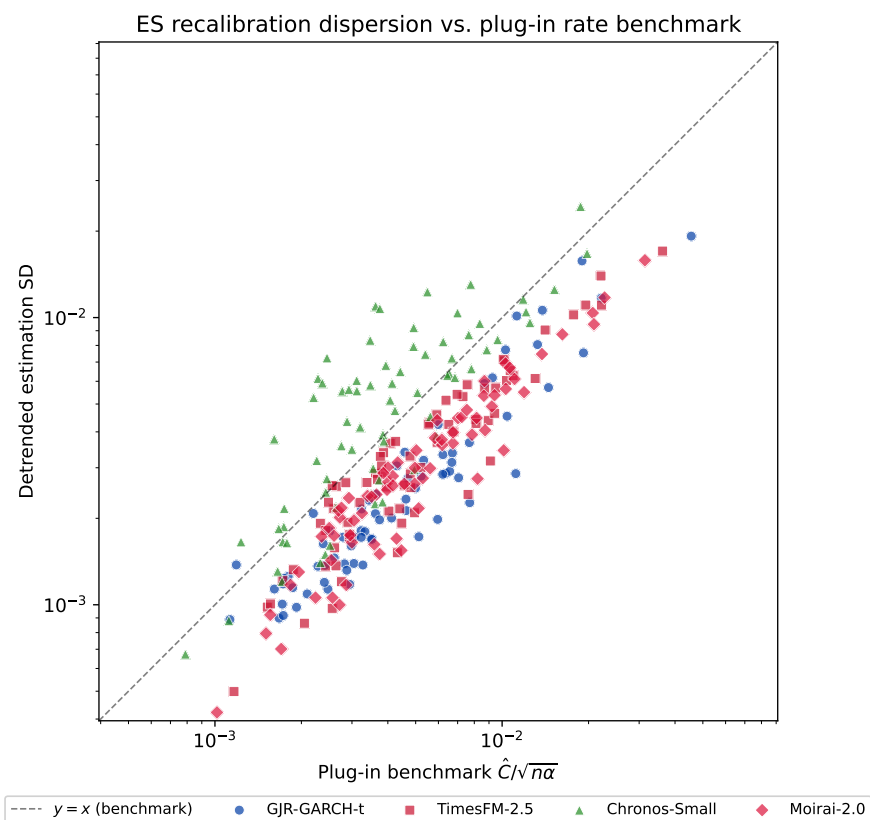


Figure 1. Detrended standard deviation of $\hat{r}_n$ versus the plug-in precision benchmark $\hat{C}/\sqrt{n\alpha}$. Each point is one (asset, forecaster, α) cell. The dashed 45-degree line marks $R = 1$.

Figure A3 in Appendix I disaggregates the same comparison by forecaster. GJR-GARCH- t and Moirai 2.0 mostly lie below the 45-degree line, while Chronos-Small straddles it.

The cross-sectional scaling regression in Equation (10) tests whether empirical ES recalibration dispersion follows the predicted effective-tail-count rate. Table 8 reports the results across 288 cells. The estimated slope is $\hat{b} = -0.436$ with clustered standard error 0.039, close to the theoretical value $-1/2$. The tail-dispersion coefficient is $\hat{\gamma} = 0.870$ with standard error 0.036, below the plug-in prediction $\gamma = 1$.

The estimate $\hat{\gamma} < 1$ does not contradict the $(n\alpha)^{-1/2}$ rate, which concerns scaling in the effective tail count. Two mechanisms can reduce the estimated tail-dispersion coefficient: full-sample estimation of $\hat{\sigma}_{\text{tail}}$ may average across volatility regimes, and rolling-window overlap may reduce the measured dispersion of $\hat{r}_n$.

Table 9 reports the benchmark ratio R at $\alpha = 2.5\%$. GJR-GARCH- t , TimesFM 2.5, and Moirai 2.0 have median ratios below one, whereas Chronos-Small has a median $R = 1.13$ and reaches $R = 2.69$ for Natural Gas. This pattern is consistent with excess ES recalibration dispersion being linked to first-stage VaR failure rather than to the precision benchmark itself.

Table 8. Cross-sectional scaling regression. The sample contains 288 (asset, forecaster, α) cells. Standard errors are clustered by asset. Theoretical predictions are $b = -0.50$ and $\gamma = 1.00$.

Model	$\hat{b}$ (SE)	$\hat{\gamma}$ (SE)	R^2
$\log(n\alpha) + \log(\hat{\sigma}_{\text{tail}})$	−0.436 (0.039)	0.870 (0.036)	0.676
+ forecaster FE	−0.429 (0.028)	0.943 (0.027)	0.842
$\log(n) + \log(\hat{\sigma}_{\text{tail}})$	—	0.926 (0.047)	0.529
$\log(\hat{\sigma}_{\text{tail}})$ only	—	0.926 (0.047)	0.529

Table 9. Benchmark ratio $R_{i,m,\alpha}$ at $\alpha = 2.5\%$, across 24 assets. The numerator is the detrended standard deviation of $\hat{r}_n$ using a 252-day moving-average filter; the denominator is the plug-in precision benchmark.

Forecaster	Median	Q1	Q3	Min	Max
Chronos-Small	1.13	0.83	1.53	0.60	2.69
GJR-GARCH- t	0.53	0.48	0.63	0.34	0.94
Moirai 2.0	0.59	0.52	0.67	0.35	0.75
TimesFM 2.5	0.63	0.55	0.77	0.35	0.90

The controlled simulation in Section 5 establishes the mechanism: as the first-stage volatility model moves from the oracle specification ($\delta = 0$) to the unconditional-volatility specification ($\delta = 1$), VaR calibration degrades, and the ES dispersion ratio rises from $R = 0.95$ to $R = 1.76$ (Table 3). This regime-dependent VaR error channel is independent of any single forecaster and provides the causal logic for the VaR-first diagnostic.

The empirical data corroborate this mechanism. Figure 2 plots the Christoffersen conditional-coverage statistic against R at $\alpha = 1\%$. The pooled Spearman correlation is $\rho = 0.776$ ($p < 0.001$). Excluding Chronos-Small, the correlation remains positive and significant ($\rho = 0.513$, $p < 0.001$), whereas within Chronos-Small alone, it is small and statistically insignificant. All 12 cells with $R > 1$ belong to Chronos-Small and exhibit severe VaR miscalibration (Table A16 in Appendix I). The pooled association is strengthened by the separation of Chronos-Small from the remaining forecasters, and the ex-Chronos correlation becomes insignificant at $\alpha = 2.5\%$ and 5% (Table 10), consistent with the reduced power of the diagnostic at less extreme tail levels.

Table 10. VaR-first diagnostic robustness across tail levels. Entries report Spearman's rank correlations between the Christoffersen conditional-coverage statistic and benchmark ratio R at the diagnostic α indicated. *** denotes $p < 0.001$.

Diagnostic α	n Pairs	Spearman ρ (All)	Spearman ρ (Excl. Chronos)	Comment
1%	76	0.776 ***	0.513 ***	FRTB VaR gatekeeper
2.5%	87	0.590 ***	0.192	Matched to FRTB ES level
5.0%	95	0.559 ***	0.146	Scaling validation

Because the headline association is driven by a single severely miscalibrated forecaster, the operationally relevant question is whether the diagnostic remains informative among reasonably well-calibrated models. We test this directly on the subset {GJR-GARCH- t , TimesFM 2.5, Moirai 2.0} using two inferential designs (Table 11). A marginal permutation test that shuffles R across all cells reproduces the weak cross-sectional signal at $\alpha = 1\%$ ($\rho = 0.513$, $p < 0.001$) but not at $\alpha = 2.5\%$ or 5% ($p = 0.13$ and 0.23). Crucially, a within-asset permutation test—which shuffles R only among forecasters sharing the same asset, and therefore conditions on asset fixed effects, the comparison most relevant to model monitoring—fails to reject at every tail level, including $\alpha = 1\%$ ($p = 0.46$, 0.51 , 0.41). A cluster-robust panel regression $R_{ij} = \alpha_i + \beta \log(\text{CC}_{ij}) + u_{ij}$ with asset fixed effects confirms this: $\hat{\beta}$ is insignificant at all three tail levels (Table 12). The cross-sectional $\alpha = 1\%$ signal is thus a between-asset pattern—assets with intrinsically harder tails exhibit both poorer VaR coverage and higher R —and not a within-asset ordering that would let the diagnostic rank well-calibrated competitors. We therefore scope the VaR-first diagnostic accordingly: it is established as a detector of gross first-stage miscalibration (the Chronos-Small stress case) and of the cross-asset tail-difficulty gradient, but it is not established as a tool for discriminating among well-calibrated ES forecasts at the FRTB level $\alpha = 2.5\%$. This narrower reading is the honest one and is consistent with the between-group nature of the result noted in Section 8.

Table 11. VaR-first diagnostic within the well-calibrated forecaster subset. Spearman ρ between the Christoffersen CC statistic and the benchmark ratio R , with marginal (free) and within-asset (asset-fixed-effects) permutation p -values (2×10^4 permutations). *** $p < 0.001$ on the within-asset test.

α	Forecaster Subset	n	Spearman ρ	p (Free Perm.)	p (Within-Asset Perm.)
1.0%	Pooled (4 models)	76	+0.776 ***	<0.001	<0.001
1.0%	Well-calibrated subset	52	+0.513	<0.001	0.458
2.5%	Pooled (4 models)	87	+0.590 ***	<0.001	<0.001
2.5%	Well-calibrated subset	63	+0.192	0.131	0.508
5.0%	Pooled (4 models)	95	+0.559 ***	<0.001	<0.001
5.0%	Well-calibrated subset	71	+0.146	0.227	0.411

Table 12. Panel regression $R_{ij} = \alpha_i + \beta \log(\text{CC}_{ij}) + u_{ij}$ on the well-calibrated subset, with asset fixed effects and standard errors clustered by asset.

α	$\hat{\beta}_{\log \text{CC}}$	Cluster SE	p
1.0%	−0.004	0.030	0.900
2.5%	+0.011	0.023	0.623
5.0%	+0.011	0.017	0.530

Chronos-Small is therefore best interpreted as a stress case for the diagnostic, not as a competitive benchmark. Its excess dispersion is consistent with time-varying VaR misalignment; after separating these severely miscalibrated cells, ES recalibration dispersion is broadly comparable across the remaining forecasters.

Table 13 reports the headline results under forecaster subsets. Excluding Chronos-Small increases the precision-fragile share because the remaining forecasters produce ES corrections that are harder to distinguish. The precision-fragile share is not a model-free empirical constant; it depends on the dispersion of the forecasters included in the comparison set. The forecaster-agnostic object is the precision threshold, not the realised share.

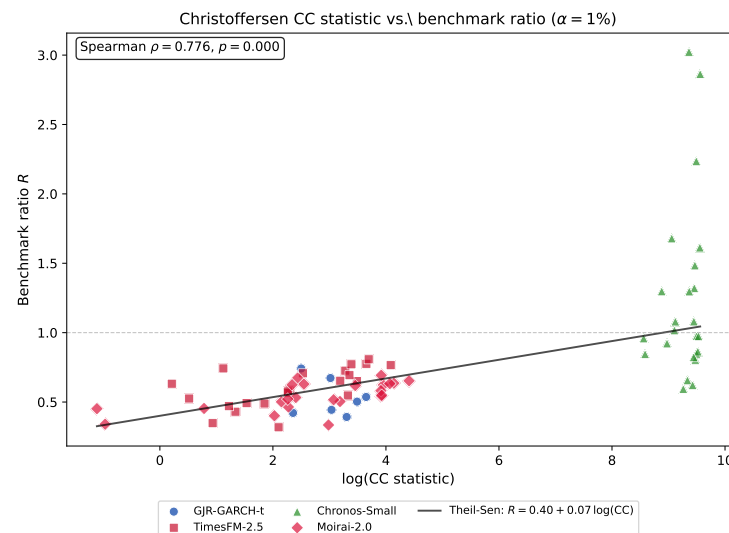


Figure 2. VaR-first diagnostic at $\alpha = 1\%$. The horizontal axis is the Christoffersen conditional-coverage statistic [21], a measure of first-stage VaR miscalibration (larger values indicate worse calibration); the vertical axis is the benchmark ratio $R = \widehat{SD}^{\det}(\hat{r}_n) / \hat{B}$, the detrended ES-recalibration dispersion expressed in units of the plug-in precision floor, so that $R > 1$ marks excess dispersion relative to the effective-tail-count benchmark. Each point is one (asset, forecaster) pair; the four forecasters are pooled. The positive association shows that excess ES dispersion concentrates in poorly VaR-calibrated cells: the pooled Spearman rank correlation is $\rho = 0.776$ ($p < 0.001$), and $\rho = 0.513$ once the severely miscalibrated Chronos-Small cells are excluded.

Table 13. Headline results under forecaster subsets. Precision-fragile share is computed at $\alpha = 2.5\%$. Spearman ρ is the correlation between the Christoffersen conditional-coverage statistic and benchmark ratio R at $\alpha = 1\%$. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Sample	Precision-Fragile Share	Median R	VaR-Diagnostic ρ
All forecasters	20.1%	0.64	0.776 ***
Excl. Chronos-Small	40.3%	0.58	0.513 ***
Only calibrated (GJR, Moirai)	8.3%	0.56	0.509 **

6.4. Implications for Tail-Risk Practice

At $\alpha = 2.5\%$ with $n = 250$ and a typical $\hat{C} \approx 1.3\%$, the implied recalibration tolerance is

$$\varepsilon_n = \frac{\hat{C}}{\sqrt{n\alpha}} \approx 52 \text{ bp.} \quad (18)$$

This tolerance is of the same order as many inter-model ES differences, so it provides a simple way to report the sampling uncertainty attached to ES recalibration and model comparison.

The precision audit can be implemented as a four-step workflow:

1. report the effective tail count $n\alpha$, not only the window length n ;
2. estimate $\hat{C} = \hat{\sigma}_{\text{tail}}$ from tail residuals;
3. compute the plug-in precision floor $\hat{C} / \sqrt{n\alpha}$;
4. flag ES comparisons whose absolute recalibration difference falls below the corresponding pairwise tolerance as precision-fragile.

A precision-fragile comparison does not imply that the two models are equivalent. It means that the observed ES-recalibration difference is smaller than the precision budget supported by the available tail data. Such comparisons should not be used as decisive evidence for model replacement without additional data, stronger structural assumptions,

or complementary diagnostics. Appendix J summarises the benchmark, sample-size rule, and diagnostic screen as a workflow.

7. Robustness and Sensitivity

This section collects robustness checks on the main empirical findings.

Cutoff multiplier sensitivity. Replacing the baseline tolerance in Equation (9) by $\kappa \sqrt{\hat{C}_1^2 + \hat{C}_2^2} / \sqrt{n\alpha}$ gives the shares in Table 14. Even at $\kappa = 0.5$, 16.0% of comparisons remain precision-fragile.

Table 14. Precision-fragile share at $\alpha = 2.5\%$ under alternative cutoff multipliers κ . Each specification has $\binom{4}{2} \times 24 = 144$ pairwise comparisons.

κ	Precision-Fragile	Share (%)
0.5	23	16.0%
1.0	29	20.1%
1.5	45	31.2%
2.0	65	45.1%

VaR-first diagnostic across tail levels. Table 10 reports the diagnostic at all three tail levels. The association between the conditional-coverage statistic and R is strongest at $\alpha = 1\%$, supporting the choice of this level as the diagnostic anchor. At $\alpha = 2.5\%$ and 5% , the full-sample correlation remains significant, but the correlation excluding Chronos-Small becomes insignificant.

Window-length scaling. A non-overlapping window-length scaling test at $\alpha = 1\%$ provides a complementary check; details are reported in Appendix D. Three of four forecasters are consistent with the $(n\alpha)^{-1/2}$ rate. The pooled test rejects $H_0 : b = -0.50$ at $p = 0.033$, driven by TimesFM 2.5; excluding TimesFM 2.5, all subsamples are consistent with the theoretical slope within their 95% confidence intervals.

HS-250 naive benchmark. As a robustness check, Appendix H reports results for a naive Historical Simulation benchmark (HS-250) that uses the empirical quantile and tail mean from the same 250-day window. HS-250 lies close to the plug-in precision floor ($R = 0.96$ at $\alpha = 2.5\%$), and all of its comparisons with TimesFM 2.5 and Moirai 2.0 are precision-fragile, reinforcing the interpretation that the binding constraint is often the effective tail count rather than model complexity.

Comparison with Diebold–Mariano and the Model Confidence Set. A natural question is whether the precision-fragile screen identifies anything that established forecast-comparison procedures do not already capture. To answer it, we reconstruct a genuinely out-of-sample, per-day Fissler–Ziegel (FZ₀) loss series for each forecaster—the rolling (q, r) correction is estimated on each 250-day window and then applied to the next 21-day block - and feed the loss differentials to a Diebold–Mariano test [26] with a Newey–West HAC variance and the small-sample correction of Harvey et al. [27], and to a Model Confidence Set (MCS, [28]) per asset. At $\alpha = 2.5\%$ the DM test cannot distinguish 125 of the 144 forecaster pairs at the 5% level, and the 75% MCS retains on average 3.0 of the 4 forecasters (all four for 10 of 24 assets, at least three for 17). These established procedures therefore independently confirm that ES forecast comparisons are frequently statistically unresolvable at FRTB parameters. The screen does not contradict them: 90% of the 29 precision-fragile pairs are also DM-indistinguishable (Table 15), and the three exceptions are all TimesFM 2.5 versus Moirai 2.0 pairs in which DM detects a difference in the VaR component of the joint loss that the ES-correction screen, by construction, does not target. The incremental value of the screen is not a sharper rejection rule but a different object: it attributes indistinguishability to the effective tail count $n\alpha$ through a closed-form floor

that requires neither a loss-differential time series nor resampling, and pinpoints why comparisons are unresolvable. Full details are in Appendix L.

Table 15. Precision-fragile screen versus the Diebold–Mariano test at $\alpha = 2.5\%$ (144 forecaster pairs). DM uses out-of-sample FZ_0 loss differentials with Newey–West HAC variance and the Harvey et al. [27] correction.

	DM Cannot Distinguish ($p \geq 0.05$)	DM Distinguishes ($p < 0.05$)
Precision-fragile	26	3
Not precision-fragile	99	16

$P(\text{DM cannot distinguish} \mid \text{fragile}) = 90\%$; DM cannot distinguish 125/144 pairs overall; mean MCS size 3.0/4 forecasters; $P(\text{both retained in MCS} \mid \text{fragile}) = 69\%$.

8. Limitations

Six limitations delimit the interpretation.

First, the empirical benchmark uses the CLT-motivated plug-in scale $\hat{C}$, not the Le Cam constant c_L . The latter depends on local density geometry and is not estimated from data. Hence, values $R < 1$ indicate that the plug-in benchmark is conservative for that cell; they do not contradict the lower-bound argument.

Second, the non-overlapping window-scaling test is data-intensive. At $n = 1000$, it requires roughly 5000 trading days, leaving too few usable assets for Chronos-Small to estimate a meaningful model-specific slope.

Third, the theoretical benchmark is oracle-style: it is conditional on a fixed first-stage forecast sequence. The empirical rolling design approximates this condition because forecasts are predictable from past information, but they are not independent of the full calibration path. A strictly held-out design would isolate the oracle assumption more cleanly.

Fourth, absolute benchmark ratios depend on how low-frequency volatility-regime shifts are removed. The forecaster ranking is stable across the 126-, 252-, and 504-day filters reported in Appendix C, but the levels of R and the precision-fragile share vary with the detrending choice.

Fifth, the pooled window-scaling test rejects $H_0 : b = -0.50$ at $p = 0.033$, driven by TimesFM 2.5 (Appendix D). The rejection disappears in restricted samples.

Sixth, the VaR-first diagnostic is partly a between-group result. The pooled association between the Christoffersen conditional-coverage statistic and R is strengthened by the separation of Chronos-Small from the remaining forecasters. Formal permutation and panel tests (Tables 11 and 12) show that, once asset fixed effects are controlled, the association does not survive within the well-calibrated forecaster subset at any tail level. The diagnostic should therefore be read as a detector of gross first-stage miscalibration and of cross-asset tail difficulty, not as an established tool for ranking well-calibrated ES forecasts at $\alpha = 2.5\%$.

These limitations define the scope of the audit. The precision floor is a necessary condition for meaningful ES comparison, not a sufficient condition for model superiority. The proposed quantities should be read as precision diagnostics for ES model comparison under tail-data scarcity.

9. Conclusions

This paper studies how the effective tail sample size constrains pairwise comparisons of expected shortfall forecasts. The relevant precision budget is $n\alpha$, not the nominal calibration window length n . At the FRTB tail level $\alpha = 2.5\%$ and $n = 250$, the effective tail count is only $n\alpha = 6.25$.

The paper converts the known $(n\alpha)^{-1/2}$ information limit into a precision-audit framework with four components: a plug-in precision benchmark, a finite-sample sample-size rule, a precision-fragile pairwise comparison screen, and a VaR-first diagnostic. The audit is diagnostic, not a replacement for formal model validation.

Across 24 global assets and four forecasters, roughly one in five pairwise ES comparisons is precision-fragile: the observed difference in recalibrated ES corrections is smaller than the precision floor supported by the available tail data. This share ranges from 16% to 29% across estimation variants and remains economically material across alternative noise scales, detrending methods, and bootstrap specifications.

The VaR-first diagnostic shows that excess ES recalibration dispersion is concentrated in cells with poor first-stage VaR calibration. VaR miscalibration is therefore a major channel for inflated ES dispersion, consistent with the simulation evidence in Section 5. A naive Historical Simulation benchmark lies close to the plug-in precision floor, reinforcing the interpretation that, at standard tail levels, the effective number of tail observations can be more binding than model complexity.

The precision-fragile share is not a universal constant. It depends on the asset universe, forecaster set, tail-dispersion estimate, and detrending convention. What is invariant is the information constraint: ES recalibration precision scales with $n\alpha$, not with the nominal window length alone.

The analysis is an ex-post oracle-style precision audit, not a deployable real-time rule. Future work should develop fully held-out, real-time, dependence-aware versions of the audit that relax the fixed-first-stage assumption. Extensions to multivariate tail-risk settings and to formal decision-theoretic frameworks for the comparison screen are also left for future investigation.

Author Contributions: Conceptualization, D.T.P. and M.M.-M.-P.; methodology, D.T.P. and M.M.-M.-P.; software, D.T.P.; validation, D.T.P. and M.M.-M.-P.; formal analysis, D.T.P. and M.M.-M.-P.; investigation, D.T.P. and M.M.-M.-P.; resources, D.T.P.; data curation, D.T.P.; writing—original draft preparation, D.T.P. and M.M.-M.-P.; writing—review and editing, D.T.P. and M.M.-M.-P.; visualization, D.T.P. and M.M.-M.-P.; supervision, D.T.P. and M.M.-M.-P.; project administration, D.T.P. and M.M.-M.-P.; funding acquisition, D.T.P. and M.M.-M.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the Marie Skłodowska-Curie Actions under the European Union’s Horizon Europe research and innovation programme for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635; and the project “IDA Institute of Digital Assets”, CF166/15.11.2022, contract number CN760046/23.05.2023; the project “AI for Energy Finance (AI4EFin)”, CF162/15.11.2022, contract number CN760048/23.05.2023; the project “Accountable Governance and Responsible Innovation in Artificial Intelligence”, CF158/15.11.2022, contract number CN760047/23.05.2023, financed under Romania’s National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8. We acknowledge the support of the project “MA’AT — Autonomous Model for Textual Assistance”, SMIS Code 2021+: 330941, funding contract no. 390090/11.11.2025, project co-financed by the European Regional Development Fund through the Smart Growth, Digitalisation and Financial Instruments Programme 2021–2027 (POCIDIF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Daily data are sourced from Yahoo Finance for the assets listed in Appendix C. Replication code and intermediate data files are available on the https://github.com/Quantlet/HMD_ES/ (accessed on 28 May 2026) platform; accompanying slides are available on the <https://quantinar.com/course/100058/finite-sample-precision-limits-for-expected-shortfallforecast-comparisons> (accessed on 28 May 2026) platform.

Acknowledgments: During the preparation of this manuscript, the authors used generative AI tools for language editing, L^AT_EX assistance, and code-checking support. The authors reviewed and edited all generated content and take full responsibility for the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ES	Expected Shortfall
VaR	Value-at-Risk
FRTB	Fundamental Review of the Trading Book
FZ	Fissler–Ziegel
CLT	Central limit theorem
CC	Christoffersen conditional coverage
HS	Historical Simulation

Appendix A. Minimax Lower Bound and Proof

The distribution class used in the Le Cam construction is

$$\mathcal{P}(\alpha, w, m_0) := \left\{ P : P \ll \lambda, f_P = \frac{dP}{d\lambda}, \mathbb{E}_P[X^-] < \infty, f_P(x) \geq m_0, \right. \quad (\text{A1})$$

$\forall x \in [\text{VaR}_\alpha(P) - w, \text{VaR}_\alpha(P)]$, where $w > 0$ is the tail window width and $m_0 > 0$ is a density lower bound on the tail window.

Theorem A1 (Minimax lower bound). *There exists a constant $c_L > 0$, depending only on the tail-window width w and density lower bound m_0 , such that for any estimator T_n of $\text{ES}_\alpha(P)$ based on n i.i.d. observations from $P \in \mathcal{P}(\alpha, w, m_0)$,*

$$\inf_{T_n} \sup_{P \in \mathcal{P}(\alpha, w, m_0)} \mathbb{E}_P |T_n - \text{ES}_\alpha(P)| \geq \frac{c_L}{\sqrt{n\alpha}}. \quad (\text{A2})$$

By Remark 1, the same bound applies to any additive recalibration estimator $\hat{r}_n$ under a fixed first-stage forecast.

To illustrate the magnitude of c_L , consider a Student- t_5 return distribution with $w = \sigma_{\text{tail}}$ and $m_0 = f_{t_5}(\text{VaR}_\alpha - w)$. Then c_L evaluates to 0.0017 at $\alpha = 1\%$, 0.0024 at $\alpha = 2.5\%$, and 0.0031 at $\alpha = 5\%$. The plug-in constant $C = \sigma_{\text{tail}}$ at the same calibration is 1.39, 1.16, and 1.03 respectively, giving $c_L/C \approx 0.1\text{--}0.3\%$. The Le Cam constant is conservative by construction: the two-point perturbation uses the smallest density on the tail window, while the CLT variance integrates the full tail distribution. The plug-in benchmark $\hat{C}/\sqrt{n\alpha}$ is therefore the operational object.

Lemma A1 (Antisymmetric two-point perturbation). *Fix $\alpha \in (0, 1/2)$ and let P_0 have density g_0 satisfying $g_0(x) \geq 2m_0$ on $[v_0 - w, v_0]$, where $v_0 := \text{VaR}_\alpha(P_0)$. For $\delta > 0$, define*

$$f_\pm(x) := g_0(x) \pm \frac{\delta}{w} h\left(\frac{x - v_0 + w}{w}\right), \quad (\text{A3})$$

where $h(u) = \sin(2\pi u)$ on $[0, 1]$ and $h = 0$ elsewhere. Let $\kappa := \inf_{u \in [0, 1]} 4\pi(1 - u)/[1 - \cos(2\pi u)] \approx 2.76$ (attained near $u \approx 0.63$). Then, provided that $\delta \leq \min(1, \kappa) m_0 w$:

(i) $f_\pm$ are valid densities with $f_\pm(x) \geq m_0$ on $[v_0 - w, v_0]$, so $P_\pm \in \mathcal{P}(\alpha, w, m_0)$;

- (ii) $\text{VaR}_\alpha(P_+) = \text{VaR}_\alpha(P_-) = v_0$;
 (iii) $|\text{ES}_\alpha(P_+) - \text{ES}_\alpha(P_-)| = w\delta/(\pi\alpha)$.

Proof. (i) Since $\int_0^1 h(u) du = 0$ and the perturbation is supported on $[v_0 - w, v_0]$, both $f_\pm$ integrate to 1. On the support, $f_\pm(x) \geq 2m_0 - \delta/w \geq m_0$ when $\delta \leq m_0 w$, so $P_\pm \in \mathcal{P}(\alpha, w, m_0)$.

(ii) The antisymmetry of $\sin(2\pi u)$ about $u = 1/2$ ensures that the perturbation adds zero net mass to $(-\infty, v_0]$: $\int_{v_0-w}^{v_0} (\delta/w) h((x - v_0 + w)/w) dx = \delta \int_0^1 \sin(2\pi u) du = 0$. Hence $F_\pm(v_0) = \alpha$. It remains to show $F_\pm(x) < \alpha$ for all $x < v_0$. For $x \leq v_0 - w$ the perturbation vanishes, so $F_\pm(x) = F_0(x) < \alpha$. For $x \in (v_0 - w, v_0)$, set $u := (x - v_0 + w)/w \in [0, 1)$ and compute the cumulative perturbation exactly:

$$\int_{v_0-w}^x \frac{\delta}{w} \sin\left(\frac{2\pi(t - v_0 + w)}{w}\right) dt = \frac{\delta}{2\pi} [1 - \cos(2\pi u)]. \quad (\text{A4})$$

The baseline CDF margin satisfies $F_0(v_0) - F_0(x) = \int_x^{v_0} g_0(t) dt \geq 2m_0(v_0 - x) = 2m_0 w(1 - u)$. Hence $F_+(x) < \alpha$ requires $\delta [1 - \cos(2\pi u)] / (2\pi) < 2m_0 w(1 - u)$, i.e.,

$$\delta < \frac{4\pi m_0 w (1 - u)}{1 - \cos(2\pi u)}. \quad (\text{A5})$$

The ratio $(1 - u) / [1 - \cos(2\pi u)]$ is bounded below on $[0, 1)$: it equals $1/(2\pi^2 u)$ near $u = 0$ (bounded) and tends to $+\infty$ as $u \uparrow 1$ (the cosine perturbation vanishes faster than the linear baseline margin). Let $\kappa := \inf_{u \in [0, 1)} 4\pi(1 - u) / [1 - \cos(2\pi u)] > 0$. Then $\delta \leq \kappa m_0 w$ suffices. Since κ is a universal positive constant, the constraint $\delta \leq m_0 w$ (from part (i)) can be tightened to $\delta \leq \min(1, \kappa) m_0 w$ if needed. For F_- , the cumulative perturbation has the opposite sign, so $F_-(x) = F_0(x) - \delta [1 - \cos(2\pi u)] / (2\pi) < F_0(x) < \alpha$ on $(v_0 - w, v_0)$. Therefore $F_\pm(x) < \alpha$ for all $x < v_0$, and $\text{VaR}_\alpha(P_\pm) = v_0$.

(iii) The ES difference is

$$\begin{aligned} \text{ES}_\alpha(P_+) - \text{ES}_\alpha(P_-) &= \frac{1}{\alpha} \int_{v_0-w}^{v_0} x [f_+(x) - f_-(x)] dx \\ &= \frac{2\delta}{\alpha w} \int_{v_0-w}^{v_0} x \sin(2\pi(x - v_0 + w)/w) dx. \end{aligned} \quad (\text{A6})$$

Substituting $u = (x - v_0 + w)/w$, the integral becomes

$$w \int_0^1 (v_0 - w + wu) \sin(2\pi u) du.$$

The constant term vanishes because $\int_0^1 \sin(2\pi u) du = 0$. Integration by parts gives $\int_0^1 u \sin(2\pi u) du = -1/(2\pi)$, so $|\text{ES}_\alpha(P_+) - \text{ES}_\alpha(P_-)| = w\delta/(\pi\alpha)$. $\square$

Proof of Theorem A1. Apply Lemma A1 with perturbation amplitude

$$\delta = c \sqrt{\alpha/n} \quad (\text{A7})$$

for a constant $c > 0$ to be chosen (the constraint $\delta \leq m_0 w$ is satisfied for n large enough).

Step 1: ES separation. By Lemma A1(iii),

$$\Delta := |\text{ES}_\alpha(P_+) - \text{ES}_\alpha(P_-)| = \frac{w\delta}{\pi\alpha} = \frac{wc\sqrt{\alpha/n}}{\pi\alpha} = \frac{wc}{\pi\sqrt{n\alpha}}. \quad (\text{A8})$$

Step 2: χ^2 divergence. Since f_+ and f_- differ only on $[v_0 - w, v_0]$,

$$\chi^2(P_+ \| P_-) = \int_{v_0-w}^{v_0} \frac{(f_+ - f_-)^2}{f_-} dx = \int_{v_0-w}^{v_0} \frac{4\delta^2 \sin^2(2\pi u)}{w^2 f_-(x)} dx. \quad (\text{A9})$$

Using $f_- \geq m_0$ on the support and $\int_0^1 \sin^2(2\pi u) du = 1/2$, this gives

$$\chi^2(P_+ \| P_-) \leq \frac{4\delta^2}{w^2} \cdot \frac{1}{m_0} \cdot w \cdot \frac{1}{2} = \frac{2\delta^2}{m_0 w} = \frac{2c^2 \alpha}{m_0 w n}. \quad (\text{A10})$$

Step 3: TV control. For product measures,

$$1 + \chi^2(P_+^n \| P_-^n) = (1 + \chi^2(P_+ \| P_-))^n, \quad (\text{A11})$$

so $\chi^2(P_+^n \| P_-^n) \leq \exp(n \chi^2(P_+ \| P_-)) - 1$, using $1 + x \leq e^x$. By the total-variation/ χ^2 inequality,

$$\text{TV}(P_+^n, P_-^n)^2 \leq \chi^2(P_+^n \| P_-^n) \leq \exp(2c^2 \alpha / (m_0 w)) - 1. \quad (\text{A12})$$

For small c , this is approximately $2c^2 \alpha / (m_0 w)$. To ensure $\text{TV}(P_+^n, P_-^n) \leq 1 - 1/\sqrt{2}$, it suffices to take c small enough that $\exp(2c^2 \alpha / (m_0 w)) - 1 \leq (1 - 1/\sqrt{2})^2$; since $\alpha < 1/2$, this is satisfied when $c^2 \leq (m_0 w / 2) \log(1 + (1 - 1/\sqrt{2})^2)$.

Step 4: Le Cam bound. By Le Cam's two-point lemma,

$$\inf_{T_n} \max_{j \in \{+, -\}} \mathbb{E}_{P_j} |T_n - \text{ES}_\alpha(P_j)| \geq \frac{1}{2} \Delta (1 - \text{TV}(P_+^n, P_-^n)) \geq \frac{\Delta}{2\sqrt{2}}. \quad (\text{A13})$$

Substituting the expression for Δ from Step 1 and taking c at its maximum admissible value from Step 3,

$$\inf_{T_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P |T_n - \text{ES}_\alpha(P)| \geq \frac{w}{2\sqrt{2} \pi} \cdot \frac{\sqrt{(m_0 w / 2) \log(1 + (1 - 1/\sqrt{2})^2)}}{\sqrt{n\alpha}} =: \frac{c_L}{\sqrt{n\alpha}}, \quad (\text{A14})$$

where $c_L = w \sqrt{(m_0 w / 2) \log(1 + (1 - 1/\sqrt{2})^2)} / (2\sqrt{2} \pi) > 0$ depends only on w and m_0 . $\square$

Appendix B. Finite-Sample Calibration Details

The random-count correction in Section 3.3 arises because the tail count $K_n := \sum_{i=1}^n \mathbb{1}\{X_i \leq \text{VaR}_\alpha\} \sim \text{Bin}(n, \alpha)$ is random. Conditional on $K_n = k$, the tail-average variance is $\sigma_{\text{tail}}^2 / k$; a delta-method expansion of $\mathbb{E}[1/K_n]$ using $\text{Var}(K_n) = n\alpha(1 - \alpha)$ gives the unconditional variance $\sigma_{\text{tail}}^2 f(n, \alpha)^2 / (n\alpha)$, where $f(n, \alpha) = \sqrt{1 + (1 - \alpha)/(n\alpha)}$.

An Edgeworth correction is not used. At tail levels relevant in practice, the conditional tail distribution of heavy-tailed returns is highly skewed (conditional $\gamma_1 \approx -4$ for Student- t_5 at $\alpha \leq 5\%$). At $k \leq 10$, the one-term Edgeworth approximation shifts the quantile inward and produces intervals narrower than the Gaussian approximation, contrary to Monte Carlo evidence. The random-count factor and direct Monte Carlo calibration are therefore preferred.

Under GARCH dynamics, the pattern in Table 1 separates finite-sample tail-count effects from volatility-regime effects. At short windows ($n = 250$), the unconditional σ_{tail} can exceed the within-regime tail dispersion, producing $\hat{f} < 1$. At longer windows, paths traverse multiple volatility regimes, and the empirical SD exceeds the unconditional

asymptotic prediction ($\hat{f} > 1.2$, flagged with †). This is the serial-dependence inflation that the detrending analysis in Section 3.2 is designed to remove.

An important consequence is that at short windows under realistic GARCH dynamics, the plug-in benchmark overstates the precision floor ($\hat{f} < 1$), so the precision-fragile screen uses a wider tolerance than the within-regime noise alone would justify. Its conclusions remain valid as an upper bound on the fraction of unreliable model rankings.

Table A1 quantifies the sampling variability of the precision floor itself. Because $B = \hat{C}/\sqrt{n\alpha}$ and $\hat{C}$ is a sample standard deviation of a small tail subsample, B inherits non-negligible sampling variance. Under Student- t_5 at FRTB parameters ($n = 250$, $n\alpha = 6.25$), the coefficient of variation of $\hat{C}$ across 50,000 replications is 0.76, and the 90% simulation range of $B/\bar{B}$ spans $[0.25, 2.32]$. Doubling the window to $n = 500$ ($n\alpha = 12.5$) reduces the CV to 0.56, and at $n = 1000$ it falls to 0.43. GARCH(1,1)- t_5 dynamics produce similar or slightly larger CVs at each window length, consistent with serial-dependence effects. These results reinforce the conservative reading: at standard FRTB windows, the precision floor is itself a noisy estimate, and marginal precision-fragile classifications should be treated as tentative.

Table A1. Sampling variability of the precision floor $B = \hat{C}/\sqrt{n\alpha}$ at $\alpha = 2.5\%$. $CV(\hat{C})$ is the coefficient of variation across 50,000 Monte Carlo replications. $B/\bar{B}$: 90% range gives the 5th–95th percentile ratio of the realised floor to its mean.

n	$n\alpha$	Student- t_5		GARCH(1,1)- t_5	
		$CV(\hat{C})$	$B/\bar{B}$: 90% Range	$CV(\hat{C})$	$B/\bar{B}$: 90% Range
250	6.25	0.76	[0.25, 2.32]	0.78	[0.17, 2.43]
500	12.50	0.56	[0.40, 1.99]	0.64	[0.32, 2.13]
750	18.75	0.49	[0.48, 1.84]	0.57	[0.39, 2.00]
1000	25.00	0.43	[0.52, 1.76]	0.52	[0.44, 1.94]

Appendix C. Data and Computational Details

Table A2 lists the 24 assets with tickers, asset class, sample period, and number of observations. All returns are daily log returns. Data source: Yahoo Finance.

All returns are daily log returns $r_t = \log(P_t/P_{t-1})$, expressed in decimal form (not percentages). Each asset uses its own trading calendar; no cross-asset calendar synchronisation is imposed. Missing observations (holidays, halts) are dropped, and the rolling-window count n reflects actual trading days.

The forecasting setup, base forecasters, VaR–ES extraction, FZ recalibration specification, detrended SD measure, and plug-in benchmark are described in Sections 2–6. Below are additional implementation details.

TimesFM 2.5 quantile levels are $\tau \in \{0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$; the Student- t fit minimises the sum of squared quantile deviations with unconstrained degrees of freedom. For the non-overlapping window-length scaling test (Appendix D), each block of size n is treated as an independent calibration sample; all four forecasters are included.

For each asset, forecaster, and tail level, base forecasts $(\hat{V}_t, \hat{E}_t)$ are extracted from the corresponding predictive distribution. GJR-GARCH- t yields VaR and ES analytically from the fitted conditional Student- t distribution. Chronos-Small computes both quantities nonparametrically from 1000 simulated forecast paths. TimesFM 2.5 and Moirai 2.0 require a tail-completion step: a Student- t distribution is fitted to the model output by quantile matching for TimesFM 2.5 and by sample fitting for Moirai 2.0, and VaR and ES are then extracted from the fitted distribution. The Student- t tail-completion step is a forecaster-specific extraction device, not part of the proposed audit.

Table A2. Asset universe.

Ticker	Name	Class	Sample
SP500	S&P 500	Equity	2000-01–2026-03
STOXX	Euro Stoxx 50	Equity	2004-04–2026-03
GDAXI	DAX	Equity	2000-01–2026-03
FCHI	CAC 40	Equity	2000-01–2026-03
FTSE100	FTSE 100	Equity	2000-01–2026-03
NIKKEI	Nikkei 225	Equity	2000-01–2026-03
HSI	Hang Seng	Equity	2000-01–2026-03
BOVESPA	Bovespa	Equity	2000-01–2026-03
NIFTY	Nifty 50	Equity	2007-09–2026-03
ASX200	ASX 200	Equity	2000-01–2026-03
ICLN	iShares Clean Energy	Equity	2008-06–2026-03
TLT	US 20Y+ Treasury	Bond	2002-07–2026-03
IBGL	Euro Gov Bond	Bond	2008-01–2026-03
DJCI	DJ Commodity	Commodity	2009-10–2021-01
GOLD	Gold	Commodity	2000-08–2026-03
WTI	WTI Crude	Commodity	2000-08–2026-03
NATGAS	Natural Gas	Commodity	2000-08–2026-03
CBU0	Copper	Commodity	2011-03–2026-03
BTC	Bitcoin	Crypto	2014-09–2026-03
ETH	Ethereum	Crypto	2017-11–2026-03
EURUSD	EUR/USD	FX	2003-12–2026-03
GBPUSD	GBP/USD	FX	2003-12–2026-03
USDJPY	USD/JPY	FX	2000-01–2026-03
AUDUSD	AUD/USD	FX	2006-05–2026-03

As a robustness check, the headline ratios are recomputed using moving-average windows of 126 and 504 days (Table A3). The qualitative ranking is preserved in every specification.

Table A3. Sensitivity of median benchmark ratio R at $\alpha = 2.5\%$ to the detrending moving-average window. The forecaster ranking is invariant across specifications; the absolute level shifts materially.

Forecaster	MA-126	MA-252	MA-504
GJR-GARCH- t	0.40	0.53	0.76
TimesFM 2.5	0.42	0.63	1.07
Chronos-Small	0.81	1.13	1.70
Moirai 2.0	0.41	0.59	0.98

Table A4 reports the effect of estimating $\hat{C}$ on each rolling calibration window rather than once over the full evaluation sample.

Table A4. Median benchmark ratio R under full-sample vs. rolling-window estimation of $\hat{C} = \hat{\sigma}_{\text{tail}}$.

Forecaster	$\alpha = 1\%$		$\alpha = 2.5\%$	
	R_{full}	R_{roll}	R_{full}	R_{roll}
GJR-GARCH- t	0.63	0.83	0.57	1.06
TimesFM 2.5	0.57	1.34	0.62	1.35
Chronos-Small	1.00	1.41	1.13	1.62
Moirai 2.0	0.57	1.49	0.58	1.26

Table A5. Sensitivity of the tail-dispersion coefficient $\hat{\gamma}$ to the σ_{tail} estimation method. Cross-sectional regression (10) with SEs clustered by asset (24 clusters).

σ_{tail} Measure	$\hat{\gamma}$	SE	$t(\gamma = 1)$	$\hat{b}$
Full-sample	0.866	0.055	−2.43	−0.440
Rolling (250-day)	0.758	0.055	−4.40	−0.520
Rolling + forecaster FE	0.727	0.064	−4.29	−0.516

Table A6. TimesFM first-stage extraction noise diagnostic. Dependent variable: $\log \text{SD}(\hat{r}_n)$ across 24 assets at $\alpha = 1\%$. Column (A) regresses on $\log \sigma_{\text{tail}}$ only; column (B) adds $\log \bar{\nu}^{-1}$ (mean inverse fitted Student- t degrees of freedom) as a proxy for quantile-fit noise. HC1 standard errors in parentheses.

	(A)	(B)
$\log \sigma_{\text{tail}}$	0.915 (0.058)	0.854 (0.094)
$\log \bar{\nu}^{-1}$	—	0.162 (0.145)
R^2	0.820	0.828
N	24	24

Appendix D. Window-Length Scaling Test

A clarification on the high coefficients of determination in the per-asset scaling regressions (e.g., R^2 up to 0.997 in Tables A14 and A17) is warranted. These regressions are descriptive log–log fits of a recalibration-dispersion summary on the window length, with only four design points $n \in \{250, 500, 750, 1000\}$; the near-unit R^2 reflects the fact that dispersion declines as an almost deterministic power law of n , exactly as the $(n\alpha)^{-1/2}$ theory predicts. They are not predictive return models fitted using many regressors, so the high R^2 does not indicate overfitting, and there is no in-sample versus out-of-sample gap that would give rise to optimism bias: the object of inference is the slope b (with the 95% confidence intervals reported alongside each fit), not the fit quality. The cross-sectional scaling regression (10) that underlies the headline rate estimate is, conversely, deliberately conservative—it pools 288 cells, reports standard errors clustered by asset, and yields a modest R^2 —so neither the per-asset nor the pooled rate evidence is an artefact of overfitting. The genuine small-sample limitation here is the opposite one, namely the scarcity of non-overlapping windows at long n , which we state explicitly as a limitation in Section 8.

A non-overlapping window-length scaling test at $\alpha = 1\%$ varies the calibration window length across $n \in \{250, 500, 750, 1000\}$ for all 24 assets and four forecasters, with a step size equal to n so that adjacent windows share zero observations. Per-forecaster median slopes are −0.55 (GJR-GARCH- t , 21 assets), −0.72 (TimesFM 2.5, 20 assets), and −0.56 (Moirai 2.0, 20 assets); Chronos-Small retains only 2 assets and is excluded from per-model inference. Across 63 retained (asset, forecaster) pairs, 63% of 95% CIs contain −0.50 (Table A9).

The pooled fixed-effects estimate is $\hat{b} = -0.585$ (SE = 0.040), with 95% CI $[-0.66, -0.51]$ (Table A7). Restricting the panel to GJR-GARCH- t and Moirai 2.0 gives $\hat{b} = -0.542$ ($p = 0.298$), which does not reject −0.50. Excluding TimesFM alone yields $\hat{b} = -0.532$ ($p = 0.530$). Table A8 reports all four specifications.

Table A7. Pooled fixed-effects rate test. $\log(\text{SD}(\hat{r}_n)) = \alpha_i + \beta_j + b \log(n)$ across 4 forecasters, 24 assets and window lengths $n \in \{250, 500, 750, 1000\}$, non-overlapping windows only.

Statistic	Value
Pooled slope $\hat{b}$	−0.585
Standard error	0.040
95% CI	[−0.663, −0.507]
t -stat ($H_0: b = -0.50$)	−2.13
p -value (two-sided)	0.033
R^2	0.811
Forecasters	4
Assets	24
Observations	280

Table A8. Restricted pooled FE rate tests. Same specification as Table A7, restricted to subsamples that isolate different sources of contamination.

Sample	Cells	$\hat{b}$	SE	95% CI	$p(b = -0.5)$
All forecasters	280	−0.585	0.040	[−0.66, −0.51]	0.033
Excl. TimesFM	198	−0.532	0.052	[−0.63, −0.43]	0.530
VaR-pass only	35	−0.780	0.131	[−1.04, −0.52]	0.032
GJR + Moirai only	167	−0.542	0.041	[−0.62, −0.46]	0.298

Table A9. Non-overlapping window-length scaling: per-forecaster summary. Median OLS slope from $\log(\text{SD}) = a + b \log(n)$ with HC1 robust standard errors. Chronos-Small retains only 2 assets due to short forecast histories.

Forecaster	Assets	Median $\hat{b}$	IQR	−0.5 ∈ CI	R^2 (Med)
GJR-GARCH- t	21	−0.55	[−0.64, −0.45]	67%	0.911
TimesFM 2.5	20	−0.72	[−0.85, −0.56]	50%	0.919
Chronos-Small	2	−0.38	[−0.39, −0.37]	50%	0.610
Moirai 2.0	20	−0.56	[−0.68, −0.44]	75%	0.913
Pooled	63	−0.57	[−0.73, −0.45]	63%	0.918

Appendix E. CC Statistic Details and Fisher Exact Test

The Christoffersen conditional-coverage statistic [21] decomposes as $\text{CC} = \text{UC} + \text{IND}$, where $\text{UC} = -2 \log(\alpha^{n_1}(1 - \alpha)^{n_0} / \hat{\pi}^{n_1}(1 - \hat{\pi})^{n_0})$ is the Kupiec unconditional-coverage likelihood ratio and IND is a first-order Markov independence test; under the null, $\text{CC} \sim \chi^2_2$. For Chronos-Small, hit rates of $\approx 37\%$ (vs. the nominal 1%) produce $\text{UC} > 7000$, so the large CC values in Figure 2 are arithmetically expected, not anomalous.

The Fisher exact 2×2 test on the binary Kupiec-pass vs. $R > 1$ contingency table gives $p = 0.35$, underpowered because only 12 of 96 (asset, forecaster) pairs pass Kupiec at $\alpha = 1\%$. The Christoffersen CC statistic, being continuous, provides a finer-grained diagnostic than the binary pass/reject classification (Table A15).

Appendix F. Conceptual Overview

Figure A1 summarises the paper’s logical chain from the known ES rate to the operational diagnostics.

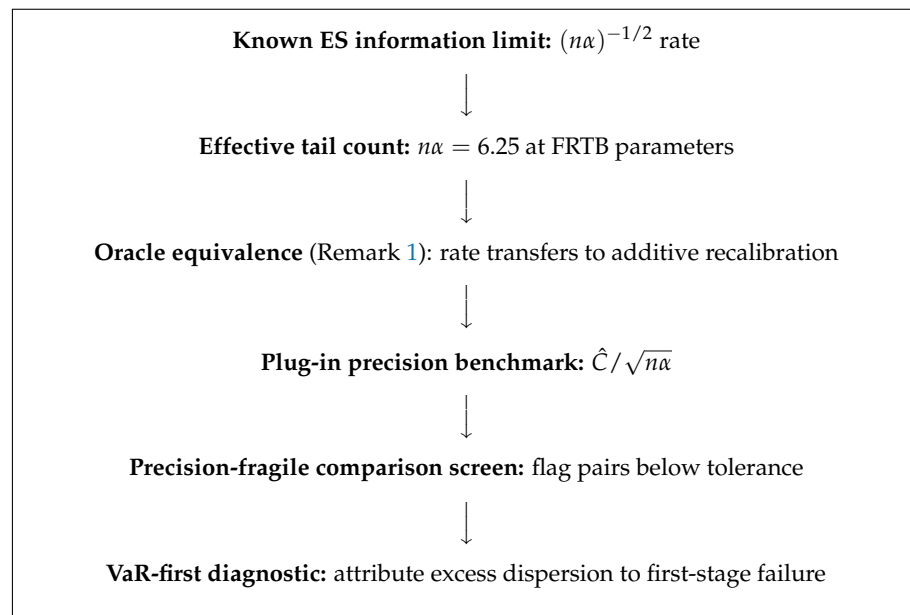


Figure A1. Logical chain of the paper: from known ES rate to operational diagnostics [1,2].

Appendix G. Detrending Illustration

Figure A2 shows the raw correction path $\hat{r}_n$, its 252-day moving average, and the detrended residual for one representative asset (S&P 500, GJR-GARCH- t , $\alpha = 2.5\%$).

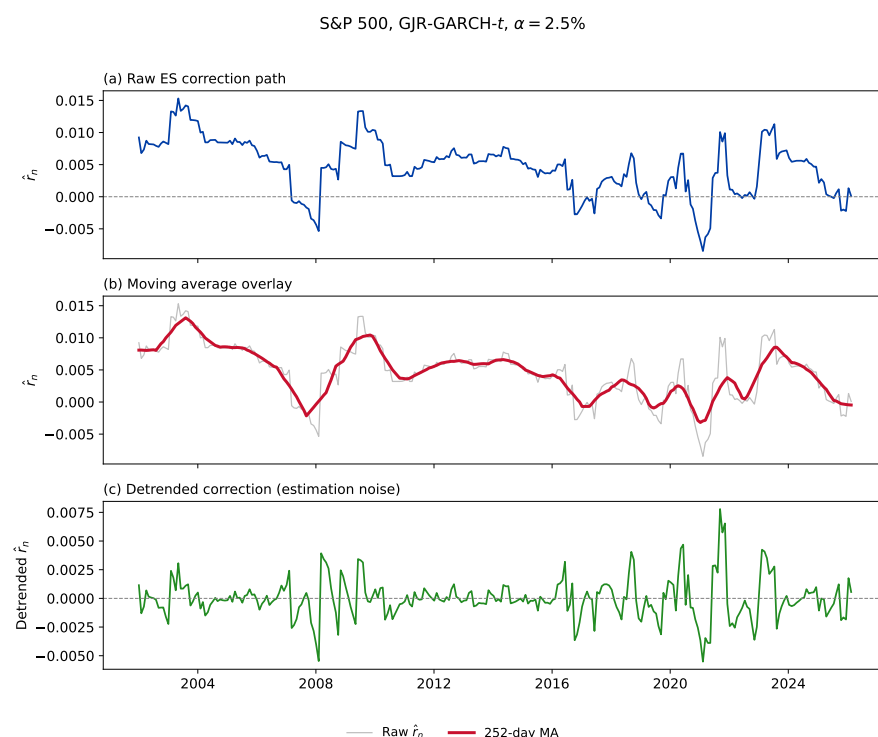


Figure A2. Raw ES correction path, 252-day moving average, and detrended correction for S&P 500 under GJR-GARCH- t at $\alpha = 2.5\%$.

Appendix H. HS-250 Naive Benchmark

Table A10 adds a Historical Simulation benchmark (HS-250) as a diagnostic reference for the effective tail-count constraint, not as a production recommendation. For each forecast date, the HS-250 base VaR and ES are the empirical quantile and conditional tail mean from the trailing 250-day window of observed returns. The FZ recalibration

correction is then estimated on the same rolling audit window, as for the other forecasters, so the benchmark ratio R is computed on an identical footing.

At $\alpha = 2.5\%$, HS-250 achieves a median benchmark ratio $R = 0.96$, close to the plug-in precision floor. Pairwise comparisons between HS-250 and the four model-based forecasters at $\alpha = 2.5\%$ yield 50 of 96 additional precision-fragile comparisons (52.1%).

Table A10. Forecaster comparison at $\alpha = 2.5\%$ including the HS-250 naive benchmark. R is the median benchmark ratio; $\bar{r}_n$ is the mean recalibration shift; $\hat{\sigma}_{\text{tail}}$ is the median empirical tail dispersion. HS-250 is included as a diagnostic reference, not as a model-selection recommendation.

Forecaster	Median R	Mean $\bar{r}_n$	Median $\hat{\sigma}_{\text{tail}}$
GJR-GARCH- t	0.53	0.0081	0.85%
TimesFM 2.5	0.63	−0.0015	1.09%
Moirai 2.0	0.59	−0.0045	1.10%
Chronos-Small	1.13	−0.0319	0.91%
HS-250 (naive)	0.96	−0.0040	1.25%

Appendix I. Additional Tables and Figures

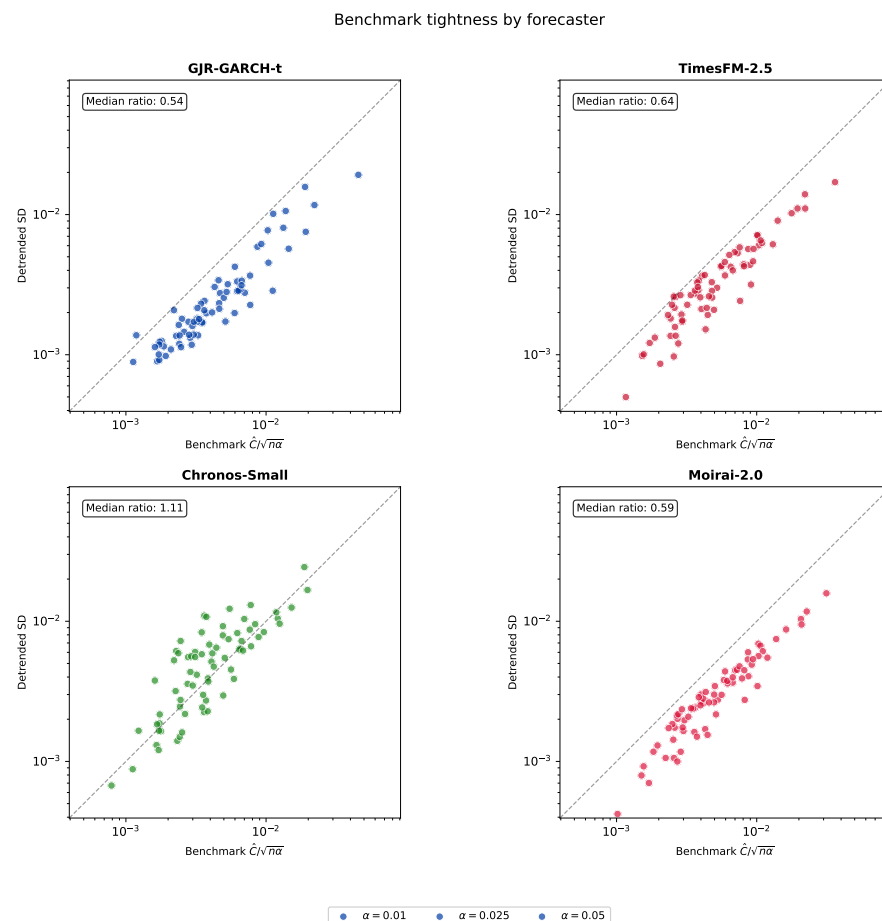


Figure A3. Detrended ES-recalibration dispersion $\widehat{SD}^{\text{det}}(\hat{r}_n)$ (vertical axis) versus the plug-in precision benchmark $\hat{B} = \hat{C}/\sqrt{n\alpha}$ (horizontal axis), shown in a 2×2 grid with one forecaster per panel: GJR-GARCH- t , TimesFM-2.5, Chronos-Small, and Moirai-2.0. Each point is one (asset, tail-level) cell, pooling all 24 assets and the three tail levels $\alpha \in \{1\%, 2.5\%, 5\%\}$; the 45° line marks equality $R = 1$, so points above it have excess dispersion. The inset in each panel reports that forecaster's median benchmark ratio R over all assets and tail levels.

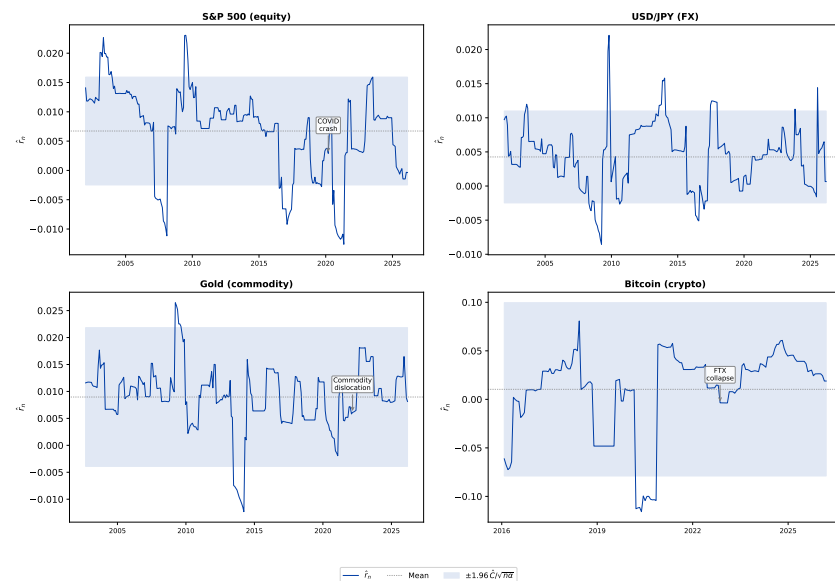


Figure A4. Rolling-window ES correction $\hat{r}_n$ across asset classes at $\alpha = 1\%$ under GJR-GARCH- t ($n = 250$, step = 21 days), shown as a scaling illustration. Shaded bands are plug-in 95% reference intervals $\bar{r} \pm 1.96 \frac{\hat{C}}{\sqrt{n\alpha}}$.

Table A11. Asset-specific minimum calibration window n at $\alpha = 2.5\%$. $\hat{\sigma}_{\text{tail}}$ is the empirical tail dispersion from GJR-GARCH- t residuals; ε_{250} is the plug-in benchmark at the 250-day window (with finite-sample correction); right-hand columns report the required n for the stated tolerance ε in basis points of return. Assets sorted by $\hat{\sigma}_{\text{tail}}$.

Asset	$\hat{\sigma}_{\text{tail}}$ (%)	ε_{250} (bp)	$\varepsilon = 25$ bp	$\varepsilon = 50$ bp	$\varepsilon = 100$ bp	$\varepsilon = 200$ bp
TLT	0.45	19	161	56	23	9
CBU0	0.55	24	228	74	29	12
USDJPY	0.60	26	262	84	32	13
ASX200	0.60	26	266	85	33	13
DJCI	0.63	27	287	91	34	14
FTSE100	0.65	28	306	96	36	15
SP500	0.70	30	348	107	39	16
AUDUSD	0.72	31	367	112	41	17
STOXX	0.75	32	392	119	43	18
GDAXI	0.81	35	453	135	48	20
EURUSD	0.81	35	455	135	48	20
GBPUSD	0.82	35	465	138	49	20
NIFTY	0.88	38	527	154	54	22
FCHI	0.91	39	561	163	56	23
GOLD	1.03	44	717	203	68	27
NIKKEI	1.16	50	893	248	80	31
BOVESPA	1.16	50	898	249	80	31
IBGL	1.28	55	1091	298	94	35
ICLN	1.33	57	1175	319	100	37
HSI	1.59	68	1657	441	132	47
NATGAS	2.31	99	3463	894	248	80
WTI	2.60	112	4371	1121	306	96
ETH	3.45	148	7665	1945	513	151
BTC	5.54	238	19,656	4943	1264	342
Median	0.85	36	496	146	51	21

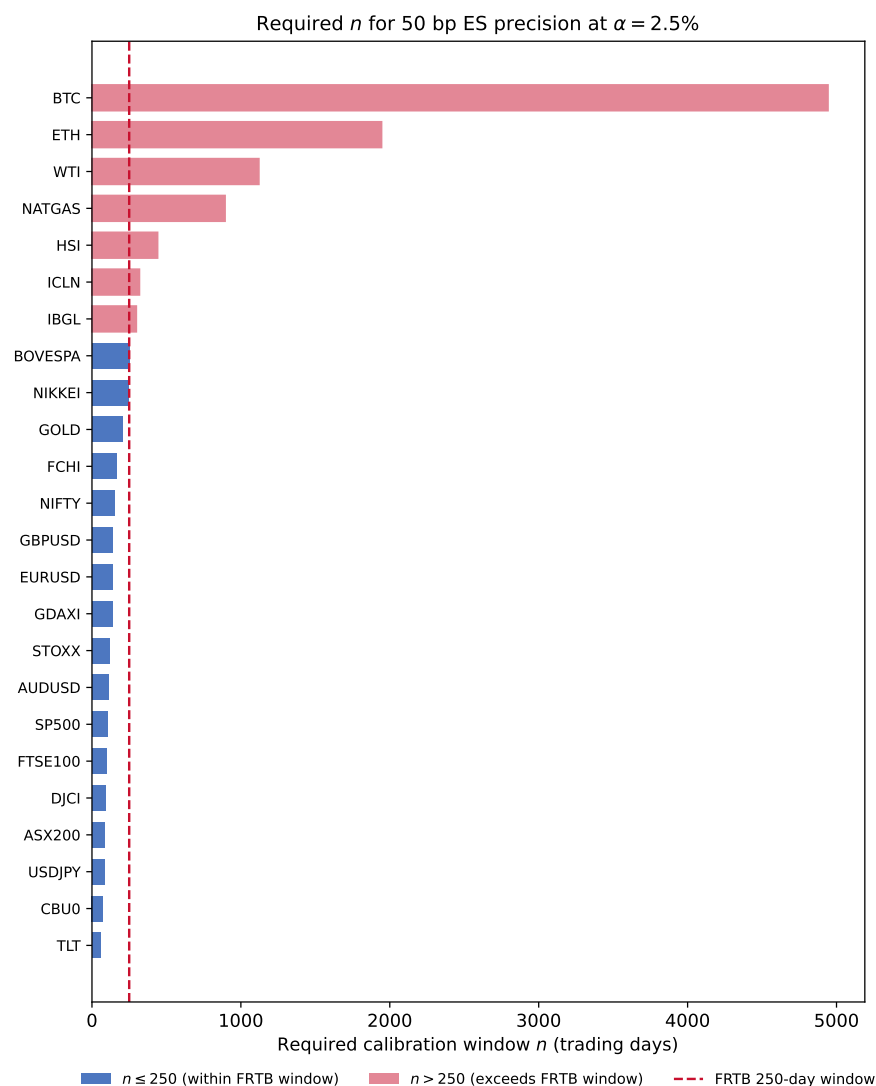


Figure A5. Required calibration window for 50 bp ES precision by asset at $\alpha = 2.5\%$. The vertical line marks the 250-day default; red color means longer calibration windows, blue color means lower calibration window.

Table A12. Asset-specific minimum calibration window n at $\alpha = 1\%$. $\hat{\sigma}_{\text{tail}}$ is the empirical tail dispersion from GJR-GARCH- t residuals; ε_{250} is the plug-in benchmark at the 250-day window; right-hand columns report the required n for the stated tolerance ε in basis points of return. Assets sorted by $\hat{\sigma}_{\text{tail}}$. For standardised-unit requirements, see Table 2.

Asset	$\hat{\sigma}_{\text{tail}}$ (%)	ε_{250} (bp)	$\varepsilon = 25 \text{ bp}$	$\varepsilon = 50 \text{ bp}$	$\varepsilon = 100 \text{ bp}$	$\varepsilon = 200 \text{ bp}$
USDJPY	0.54	34	473	119	30	8
ASX200	0.56	35	495	124	31	8
FTSE100	0.58	36	530	133	34	9
TLT	0.59	37	561	141	36	9
DJCI	0.68	43	739	185	47	12
CBU0	0.72	46	840	210	53	14
SP500	0.74	47	882	221	56	14
STOXX	0.79	50	998	250	63	16
GDAXI	0.83	52	1094	274	69	18

Table A12. *Cont.*

Asset	$\hat{\sigma}_{\text{tail}}$ (%)	ε_{250} (bp)	$\varepsilon = 25 \text{ bp}$	$\varepsilon = 50 \text{ bp}$	$\varepsilon = 100 \text{ bp}$	$\varepsilon = 200 \text{ bp}$
AUDUSD	0.94	60	1423	356	89	23
EURUSD	0.98	62	1544	386	97	25
FCHI	0.98	62	1549	388	97	25
GOLD	1.04	66	1726	432	108	27
NIKKEI	1.06	67	1802	451	113	29
NIFTY	1.12	71	1992	498	125	32
BOVESPA	1.22	77	2366	592	148	37
GBPUSD	1.22	77	2379	595	149	38
ICLN	1.37	87	3014	754	189	48
IBGL	1.76	112	4979	1245	312	78
NATGAS	1.78	112	5061	1266	317	80
HSI	2.29	145	8414	2104	526	132
ETH	3.00	190	14,404	3601	901	226
WTI	3.04	192	14,746	3687	922	231
BTC	7.20	456	83,038	20,760	5190	1298
Median	1.01	64	1636	409	103	26

Table A13. Asset-level benchmark ratio R at $\alpha = 1\%$ under GJR-GARCH- t . Assets sorted by R .

Asset	Detr. SD	Bound	Ratio	Kupiec p
IBGL	0.0029	0.0112	0.26	0.000
GBPUSD	0.0023	0.0077	0.29	0.000
AUDUSD	0.0020	0.0060	0.33	0.000
NIFTY	0.0028	0.0071	0.39	0.000
WTI	0.0075	0.0192	0.39	0.000
HSI	0.0057	0.0145	0.39	0.000
BTC	0.0192	0.0456	0.42	0.003
GOLD	0.0029	0.0066	0.44	0.000
EURUSD	0.0028	0.0062	0.46	0.000
BOVESPA	0.0037	0.0077	0.48	0.000
ASX200	0.0017	0.0035	0.48	0.000
NIKKEI	0.0034	0.0067	0.50	0.000
STOXX	0.0025	0.0050	0.51	0.000
TLT	0.0020	0.0037	0.53	0.000
FCHI	0.0033	0.0062	0.54	0.000
GDAXI	0.0028	0.0052	0.54	0.000
SP500	0.0028	0.0047	0.59	0.000
FTSE100	0.0024	0.0036	0.67	0.000
USDJPY	0.0023	0.0034	0.67	0.000
ICLN	0.0059	0.0087	0.68	0.000
DJCI	0.0030	0.0043	0.71	0.000
CBU0	0.0034	0.0046	0.74	0.143
ETH	0.0158	0.0190	0.83	0.006
NATGAS	0.0101	0.0112	0.90	0.000

Table A14. Non-overlapping window-length scaling for 21 assets (GJR-GARCH- t , $\alpha = 1\%$, step = n). OLS slope $\hat{b}$ from $\log(\text{SD}) = a + b \log(n)$ with HC1 robust standard errors. Assets sorted by slope. †: dropped for <3 valid window lengths.

Asset	Windows	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$\hat{b}$	95% CI	R^2	$-0.5 \in \text{CI}$
NIFTY	30	0.0067	0.0043	0.0028	–	−0.78	[−0.92, −0.63]	0.982	N
AUDUSD	34	0.0049	0.0032	0.0021	–	−0.77	[−0.92, −0.61]	0.978	N
HSI	50	0.0144	0.0092	0.0068	0.0051	−0.73	[−0.81, −0.65]	0.993	N
FCHI	51	0.0089	0.0073	0.0043	0.0037	−0.66	[−0.84, −0.49]	0.911	Y
GOLD	50	0.0061	0.0045	0.0022	0.0030	−0.64	[−1.02, −0.27]	0.741	Y
EURUSD	45	0.0074	0.0045	0.0048	0.0026	−0.64	[−0.95, −0.32]	0.824	Y
GBPUSD	45	0.0068	0.0044	0.0029	0.0032	−0.62	[−0.84, −0.39]	0.906	Y
STOXX	43	0.0050	0.0039	0.0039	0.0019	−0.58	[−1.08, −0.08]	0.679	Y
USDJPY	53	0.0040	0.0034	0.0025	0.0017	−0.56	[−0.83, −0.29]	0.875	Y
IBGL	30	0.0054	0.0042	0.0029	–	−0.56	[−0.78, −0.34]	0.926	Y
FTSE100	51	0.0058	0.0044	0.0030	0.0028	−0.55	[−0.65, −0.46]	0.952	Y
NATGAS	50	0.0226	0.0135	0.0112	0.0110	−0.54	[−0.72, −0.37]	0.940	Y
TLT	45	0.0046	0.0023	0.0029	0.0020	−0.53	[−0.78, −0.27]	0.712	Y
WTI	50	0.0166	0.0104	0.0098	0.0079	−0.51	[−0.59, −0.43]	0.960	Y
ASX200	51	0.0051	0.0042	0.0030	0.0027	−0.46	[−0.55, −0.38]	0.954	Y
ICLN	29	0.0122	0.0105	0.0072	–	−0.45	[−0.70, −0.20]	0.861	Y
BOVESPA	50	0.0088	0.0079	0.0069	0.0050	−0.36	[−0.60, −0.13]	0.804	Y
BTC	27	0.0429	0.0353	0.0287	–	−0.36	[−0.44, −0.27]	0.973	N
NIKKEI	50	0.0089	0.0077	0.0054	0.0063	−0.32	[−0.50, −0.14]	0.753	N
SP500	51	0.0073	0.0064	0.0060	0.0048	−0.27	[−0.41, −0.14]	0.879	N
GDAXI	51	0.0077	0.0065	0.0064	0.0056	−0.21	[−0.26, −0.16]	0.943	N
CBU0	–	–	–	–	–	–	–	–	†
DJCI	–	–	–	–	–	–	–	–	†
ETH	–	–	–	–	–	–	–	–	†
Median						−0.55	IQR [−0.64, −0.45]		67%

Table A15. Directional VaR-first contingency check at $\alpha = 1\%$. A cell is classified as “excess” if $R > 1$. The Spearman rank correlation between the CC statistic and R supersedes the binary Fisher test; see Figure 2.

	Ratio ≤ 1	Ratio > 1
VaR pass (Kupiec $p > 0.05$)	12	0
VaR reject (Kupiec $p \leq 0.05$)	72	12
Spearman rank correlation	$\rho = 0.776$, $p = 0.000$	

Table A16. Excess-dispersion case audit: all (asset, forecaster) cells with $R > 1$ at $\alpha = 1\%$. CC stat is the Christoffersen conditional-coverage statistic; $\hat{\sigma}_{\text{tail}}$ is the tail dispersion scale. All 12 excess cells belong to Chronos-Small.

Asset	Forecaster	R	Kupiec p	CC Stat	$\hat{\sigma}_{\text{tail}}$	Cause
EURUSD	Chronos-Small	3.02	<0.001	11,636.6	0.0057	Severe VaR miscalibration
USDJPY	Chronos-Small	2.86	<0.001	14,189.2	0.0059	Severe VaR miscalibration
FTSE100	Chronos-Small	2.24	<0.001	13,251.3	0.0087	Severe VaR miscalibration
ICLN	Chronos-Small	1.68	<0.001	8581.9	0.0123	Severe VaR miscalibration
ASX200	Chronos-Small	1.61	<0.001	14,112.9	0.0078	Severe VaR miscalibration
NIKKEI	Chronos-Small	1.48	<0.001	12,923.6	0.0111	Severe VaR miscalibration
SP500	Chronos-Small	1.32	<0.001	12,789.4	0.0099	Severe VaR miscalibration
BTC	Chronos-Small	1.30	<0.001	7186.8	0.0297	Severe VaR miscalibration
GBPUSD	Chronos-Small	1.30	<0.001	11,699.6	0.0044	Severe VaR miscalibration
HSI	Chronos-Small	1.08	<0.001	12,690.9	0.0106	Severe VaR miscalibration
NIFTY	Chronos-Small	1.08	<0.001	9146.5	0.0080	Severe VaR miscalibration
IBGL	Chronos-Small	1.02	<0.001	8993.9	0.0061	Severe VaR miscalibration

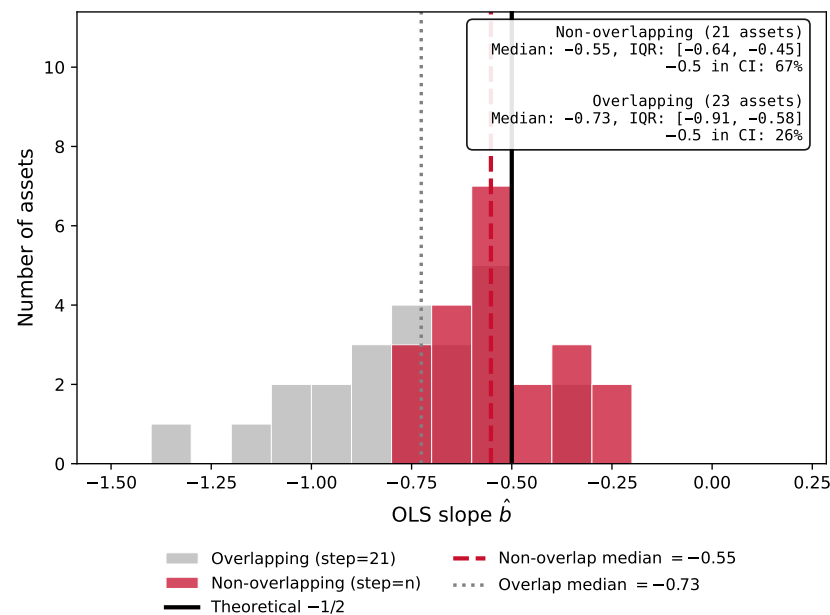


Figure A6. Distribution of per-asset OLS window-scaling slopes for two sampling designs: overlapping windows (grey, step = 21 days) and non-overlapping windows (red, step = n). The grey histogram is the same overlapping-window distribution plotted on its own in Figure A11; this figure simply overlays the non-overlapping design on top of it for direct comparison. Both panels therefore report the identical overlapping median of -0.73 (26% of asset confidence intervals contain $-1/2$); the present figure adds the non-overlapping series, whose median is -0.55 with 67% coverage. Eliminating window overlap removes the serial-correlation bias that steepens the overlapping slopes and moves the median from -0.73 toward the theoretical $-1/2$ (solid vertical line).

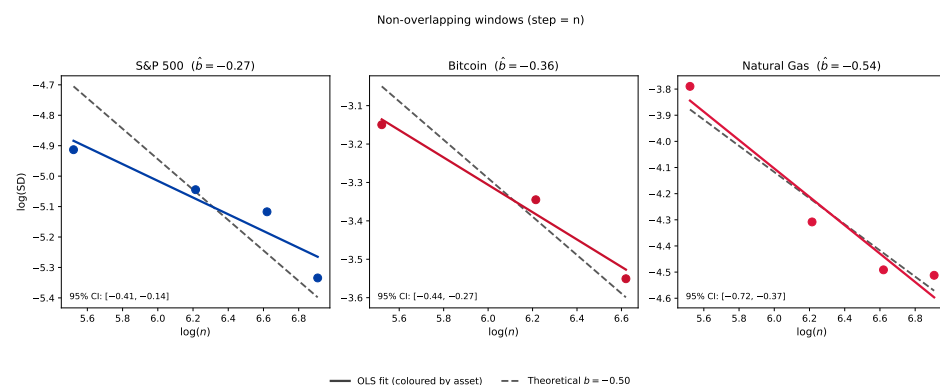


Figure A7. Log-log scaling of recalibration SD versus window length using non-overlapping windows. Each panel is a single asset and the points and OLS fit are coloured accordingly: S&P 500 (blue), Bitcoin (red) and Natural Gas (crimson). Dashed grey: theoretical $b = -0.50$; solid colour: OLS fit.

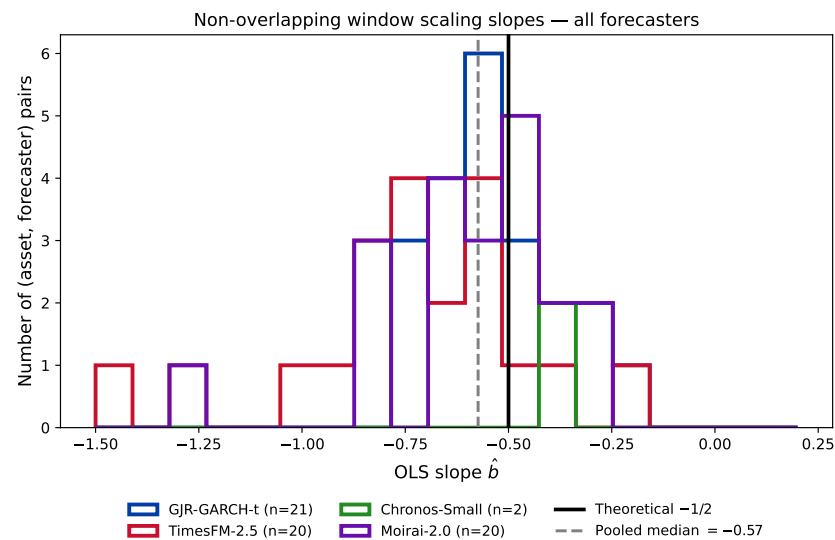


Figure A8. Distribution of non-overlapping OLS slopes across all four forecasters, coloured by forecaster: GJR-GARCH-t (blue), TimesFM-2.5 (red), Chronos-Small (green) and Moirai-2.0 (purple); the number of asset pairs per forecaster is given in the legend. Pooled median = -0.57 (63 asset-forecaster pairs); vertical line: theoretical $-1/2$.

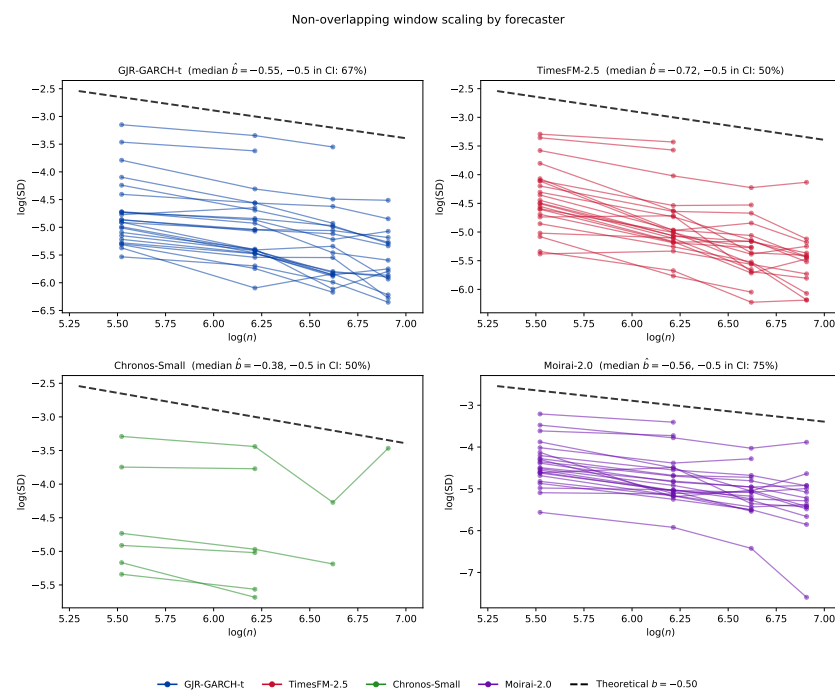


Figure A9. Non-overlapping window scaling by forecaster. Each panel is one forecaster (named in the panel title) and each solid line is a single asset, coloured by forecaster: GJR-GARCH-t (blue), TimesFM-2.5 (red), Chronos-Small (green) and Moirai-2.0 (purple). Dashed black: theoretical $b = -0.50$.

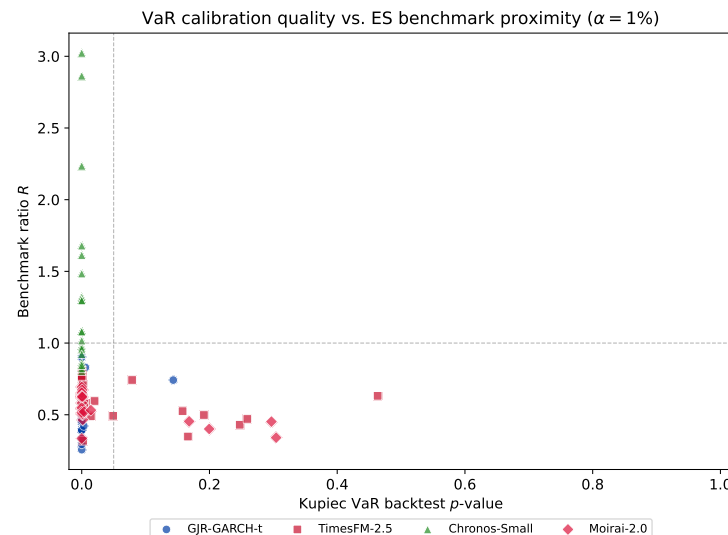


Figure A10. Kupiec p -value vs. benchmark ratio R at $\alpha = 1\%$ (binary variant of Figure 2). Each point is one (asset, forecaster) pair, with markers and colours denoting the forecaster (see legend). The two dashed grey reference lines mark the Kupiec rejection threshold ($p = 0.05$, vertical) and the benchmark-parity level ($R = 1$, horizontal); points in the upper-left quadrant are VaR-rejected cells with excess dispersion. Fisher's exact test gives $p = 0.35$ due to the small number of VaR-pass cells.

Table A17. Overlapping window-length scaling (step = 21). Same design as Table A14 but with overlapping windows. The steeper median slope (-0.73 vs. -0.55) is attributable to rolling-window overlap.

Asset	$n = 250$	$n = 500$	$n = 1000$	$n = 2000$	$\hat{b}$	95% CI	R^2	$-0.5 \in \text{CI}$
DJCI	0.0065	0.0054	0.0041	0.0002	-1.31	$[-2.38, -0.23]$	0.586	Y
NIFTY	0.0055	0.0032	0.0016	0.0005	-1.18	$[-1.40, -0.96]$	0.964	N
AUDUSD	0.0045	0.0028	0.0017	0.0005	-1.02	$[-1.25, -0.78]$	0.946	N
ETH	0.0266	0.0215	0.0200	0.0016	-1.01	$[-1.95, -0.06]$	0.523	Y
HSI	0.0133	0.0076	0.0039	0.0019	-0.97	$[-1.07, -0.87]$	0.989	N
STOXX	0.0057	0.0039	0.0024	0.0008	-0.93	$[-1.19, -0.66]$	0.921	N
GDAXI	0.0075	0.0054	0.0032	0.0012	-0.89	$[-1.11, -0.67]$	0.941	N
ICLN	0.0127	0.0081	0.0041	0.0021	-0.86	$[-0.96, -0.76]$	0.986	N
NIKKEI	0.0079	0.0058	0.0038	0.0011	-0.85	$[-1.20, -0.51]$	0.853	N
BOVESPA	0.0091	0.0063	0.0037	0.0019	-0.77	$[-0.87, -0.66]$	0.981	N
USDJPY	0.0046	0.0031	0.0021	0.0009	-0.74	$[-0.91, -0.57]$	0.948	N
GOLD	0.0058	0.0029	0.0018	0.0012	-0.73	$[-0.84, -0.61]$	0.975	N
NATGAS	0.0201	0.0128	0.0087	0.0046	-0.71	$[-0.79, -0.63]$	0.987	N
FCHI	0.0083	0.0065	0.0040	0.0024	-0.64	$[-0.76, -0.52]$	0.966	N
EURUSD	0.0060	0.0044	0.0027	0.0017	-0.63	$[-0.70, -0.57]$	0.990	N
GBPUSD	0.0057	0.0040	0.0025	0.0017	-0.61	$[-0.64, -0.57]$	0.997	N
TLT	0.0046	0.0033	0.0023	0.0013	-0.60	$[-0.70, -0.50]$	0.971	Y
IBGL	0.0056	0.0036	0.0023	0.0017	-0.57	$[-0.62, -0.52]$	0.991	N
ASX200	0.0051	0.0043	0.0030	0.0016	-0.55	$[-0.69, -0.41]$	0.938	Y
SP500	0.0072	0.0056	0.0042	0.0023	-0.54	$[-0.67, -0.42]$	0.947	Y
FTSE100	0.0057	0.0038	0.0027	0.0020	-0.50	$[-0.54, -0.46]$	0.994	Y
BTC	0.0428	0.0325	0.0277	0.0186	-0.37	$[-0.44, -0.29]$	0.959	N
WTI	0.0143	0.0101	0.0086	0.0066	-0.34	$[-0.40, -0.28]$	0.969	N
Median					-0.73	IQR $[-0.91, -0.58]$		26%

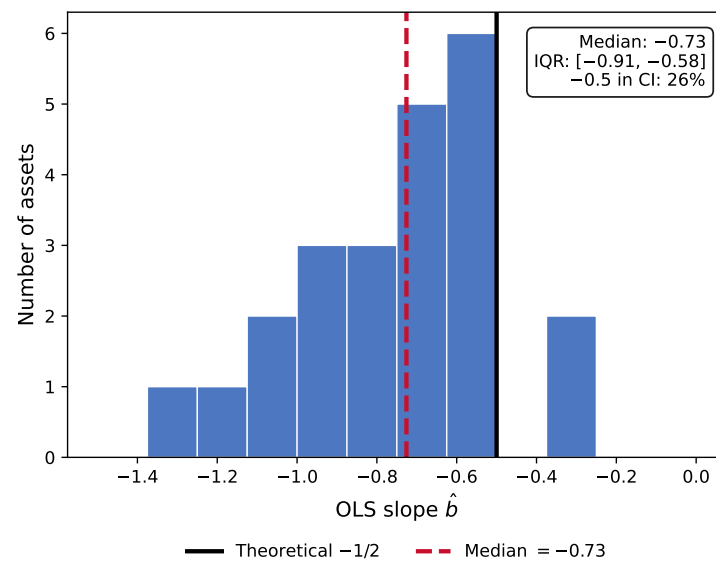


Figure A11. Overlapping-window slope distribution only (step = 21 days): median -0.73 , with only 26% of asset confidence intervals containing -0.50 . This is the overlapping-design histogram on its own; the same distribution appears as the grey series in Figure A6, where it is overlaid against the non-overlapping design (median -0.55). The two figures plot the identical overlapping data and differ only in that Figure A6 adds the non-overlapping comparison series, which shifts the median toward the theoretical $-1/2$.

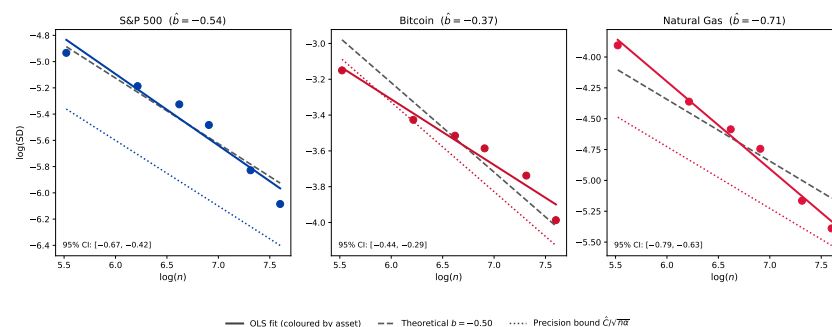


Figure A12. Overlapping-window log-log scaling for S&P 500, BTC, and Natural Gas (step = 21). Each panel is a single asset; the points, the solid OLS fit and the dotted precision bound $\hat{C}/\sqrt{n\alpha}$ are coloured by asset (S&P 500, blue; Bitcoin, red; Natural Gas, crimson), while the dashed grey line is the theoretical slope $b = -0.50$. All three slopes are steeper than in the non-overlapping design (Figure A7).

Appendix J. ES Recalibration Precision Audit

Procedure: ES Recalibration Precision Audit

Input: Returns $X_1, \dots, X_n$; base VaR–ES forecasts $(\hat{V}_t, \hat{E}_t)$; tail level α ; window length n .

1. Estimate $\hat{C} = \hat{\sigma}_{\text{tail}}$ from tail residuals $\{\hat{V}_t - X_t : X_t \leq \hat{V}_t\}$.
2. Compute the plug-in tolerance $\varepsilon_n = \frac{\hat{C}}{\sqrt{n\alpha}}$.
3. Estimate the additive ES correction $\hat{r}_n$ by FZ loss minimisation on the calibration window.
4. Compute the rolling dispersion of $\hat{r}_n$.
5. **If** dispersion $\leq \varepsilon_n$ **and** VaR is well aligned (Christoffersen CC $p > 0.05$, or Kupiec $p > 0.05$ as a coarser screen): classify residual uncertainty as intrinsic tail-sample noise.
6. **If** dispersion $> \varepsilon_n$: investigate VaR misalignment via the CC statistic (Section 6.3), time-varying first-stage bias, or non-stationarity, before interpreting ES results.

Output: Precision flag; sample-size requirement $n \geq (1/\alpha)(\hat{C}/\varepsilon)^2$ for desired tolerance ε ; diagnostic classification.

Appendix K. Conditional Heteroskedasticity Tests

Table A18 reports Engle’s ARCH-LM test [23] applied to the demeaned daily log returns of each asset, using ten lags. The Lagrange multiplier statistic is asymptotically χ^2_{10} under the null of no autoregressive conditional heteroskedasticity. The null is rejected at the 1% level for all 24 assets, with statistics ranging from 145 (ETH) to 2461 (USD/JPY). This pervasive conditional heteroskedasticity motivates the GARCH-family first-stage volatility model used throughout (Section 6.1).

Table A18. Engle ARCH-LM test (10 lags) on demeaned daily log returns. LM is the Lagrange multiplier statistic (χ^2_{10} under the no-ARCH null); n is the number of return observations.

Asset	n	LM (10)	p	Asset	n	LM (10)	p
SP500	6590	1764.4	<0.001	IBGL	4598	1091.8	<0.001
STOXX	5505	999.6	<0.001	DJCI	2822	192.5	<0.001
GDAXI	6654	1167.4	<0.001	GOLD	6409	311.5	<0.001
FCHI	6698	1133.4	<0.001	WTI	6416	1248.5	<0.001
FTSE100	6620	1374.3	<0.001	NATGAS	6414	467.2	<0.001
NIKKEI	6418	1205.3	<0.001	CBU0	3765	379.7	<0.001
HSI	6456	1063.9	<0.001	BTC	4200	144.9	<0.001
BOVESPA	6492	1774.3	<0.001	ETH	3050	146.0	<0.001
NIFTY	4538	637.5	<0.001	EURUSD	5783	2079.4	<0.001
ASX200	6621	1844.2	<0.001	GBPUSD	5795	262.3	<0.001
ICLN	4459	1537.6	<0.001	USDJPY	6803	2461.1	<0.001
TLT	5946	1285.7	<0.001	AUDUSD	5159	1519.5	<0.001

Appendix L. Diebold–Mariano and Model Confidence Set Benchmarks

This appendix details the comparison of the precision-fragile screen with the Diebold–Mariano (DM) test and the model confidence set (MCS), summarised in Section 7.

Loss construction. For each (asset, forecaster) pair at $\alpha = 2.5\%$ we form a genuinely out-of-sample per-day FZ₀ loss series. On each rolling 250-day window, the additive correction $(\hat{q}_n, \hat{r}_n)$ is estimated by FZ minimisation, then held fixed and used to score the immediately following 21-day block; concatenating the non-overlapping blocks gives one loss observation per day for each forecaster, defined on the common date support of all four forecasters so that loss differentials are aligned.

Diebold–Mariano. For each pair, we compute the mean loss differential, a Newey–West HAC long-run variance (Bartlett kernel, 21 lags) and the Diebold–Mariano statistic [26], applying the small-sample correction of Harvey et al. [27]; p -values use the t_{T-1} reference distribution. At $\alpha = 2.5\%$ the test fails to reject equal predictive accuracy for 125 of the 144 pairs at the 5% level. Of the 29 pairs flagged as precision-fragile by the screen, 26 (90%) are also DM-indistinguishable (Table 15); the three exceptions are TimesFM 2.5 versus Moirai 2.0 on BOVESPA, TLT and IBGL, where the DM statistic is driven by a difference in the VaR limb of the joint loss rather than by the ES correction targeted by the screen.

Model Confidence Set. For each asset, we apply the MCS procedure of Hansen et al. [28] to the four forecasters' loss matrix (range statistic, stationary bootstrap, block length 21, 10^3 replications, 10% size). The 90% MCS retains on average 3.0 of the 4 forecasters; it retains all four for 10 of the 24 assets, and at least three for 17. Conditional on a pair being precision-fragile, both members are jointly retained in the MCS for 69% of cases. Established multiple-comparison machinery therefore reaches the same broad conclusion as the screen—ES forecasts are frequently statistically indistinguishable at FRTB parameters—while the screen adds a closed-form, resampling-free attribution of that indistinguishability to the effective tail count $n\alpha$.

References

1. Zwingmann, T.; Holzmann, H. Asymptotics for the expected shortfall. *arXiv* **2016**, arXiv:1611.07222. <https://doi.org/10.48550/arXiv.1611.07222>.
2. Bartl, D.; Eckstein, S. Optimal nonparametric estimation of the expected shortfall risk. *arXiv* **2024**, arXiv:2405.00357. <https://doi.org/10.48550/arXiv.2405.00357>.
3. Chen, S.X. Nonparametric estimation of expected shortfall. *J. Financ. Econom.* **2008**, *6*, 87–107. <https://doi.org/10.1093/jfinec/nbm019>.
4. Fissler, T.; Ziegel, J.F. Higher order elicibility and Osband's principle. *Ann. Stat.* **2016**, *44*, 1680–1707. <https://doi.org/10.1214/16-AOS1439>.
5. Dimitriadis, T.; Bayer, S. A joint quantile and expected shortfall regression framework. *Electron. J. Stat.* **2019**, *13*, 1823–1871. <https://doi.org/10.1214/19-EJS1560>.
6. Patton, A.J.; Ziegel, J.F.; Chen, R. Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *J. Econom.* **2019**, *211*, 388–413. <https://doi.org/10.1016/j.jeconom.2018.10.008>.
7. Pele, D.T.; Bolovaneanu, V.; Ginavar, A.T.; Lessmann, S.; Härdle, W.K. Recalibrating Tail Risk Forecasts Under Temporal Dependence. 2026. Available online: <https://ssrn.com/abstract=6757685> (accessed on 28 May 2026)
8. Acerbi, C.; Székely, B. Back-testing expected shortfall. *Risk* **2014**, *27*, 76–81.
9. Nolde, N.; Ziegel, J.F. Elicibility and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.* **2017**, *11*, 1833–1874. <https://doi.org/10.1214/17-AOAS1041>.
10. Acerbi, C.; Tasche, D. On the coherence of expected shortfall. *J. Bank. Financ.* **2002**, *26*, 1487–1503. [https://doi.org/10.1016/S0378-4266\(02\)00283-2](https://doi.org/10.1016/S0378-4266(02)00283-2).
11. Du, Z.; Escanciano, J.C. Backtesting expected shortfall: Accounting for tail risk. *Manag. Sci.* **2017**, *63*, 940–958. <https://doi.org/10.1287/mnsc.2015.2342>.
12. Bayer, S.; Dimitriadis, T. Regression-based expected shortfall backtesting. *J. Financ. Econom.* **2022**, *20*, 437–471. <https://doi.org/10.1093/jfinec/nbaa013>.
13. Taylor, J.W. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *J. Bus. Econ. Stat.* **2019**, *37*, 121–133. <https://doi.org/10.1080/07350015.2017.1281815>.
14. Basel Committee on Banking Supervision. *Minimum Capital Requirements for Market Risk*; Technical Report d457; BCBS d457; Bank for International Settlements: Basel, Switzerland, 2019.
15. Glosten, L.R.; Jagannathan, R.; Runkle, D.E. On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Financ.* **1993**, *48*, 1779–1801. <https://doi.org/10.1111/j.1540-6261.1993.tb05128.x>.
16. Das, A.; Kong, W.; Sen, R.; Zhou, Y. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 10148–10167.
17. Ansari, A.F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Pineda Arango, S.; Kapoor, S.; et al. Chronos: Learning the language of time series. *arXiv* **2024**, arXiv:2403.07815.

18. Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; Sahoo, D. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 53140–53164.
19. Liu, C.; Aksu, T.; Liu, J.; Liu, X.; Yan, H.; Pham, Q.; Savarese, S.; Sahoo, D.; Xiong, C.; Li, J. Moirai 2.0: When less is more for time series forecasting. *arXiv* **2025**, arXiv:2511.11698. <https://doi.org/10.48550/arXiv.2511.11698>.
20. McNeil, A.J.; Frey, R.; Embrechts, P. *Quantitative Risk Management: Concepts, Techniques and Tools*, revised ed.; Princeton Series in Finance; Princeton University Press: Princeton, NJ, USA, 2015.
21. Christoffersen, P.F. Evaluating interval forecasts. *Int. Econ. Rev.* **1998**, *39*, 841–862.
22. Kupiec, P.H. Techniques for verifying the accuracy of risk measurement models. *J. Deriv.* **1995**, *3*, 73–84. <https://doi.org/10.3905/jod.1995.407942>.
23. Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **1982**, *50*, 987–1007. <https://doi.org/10.2307/1912773>.
24. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
25. Hansen, P.R.; Lunde, A. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *J. Appl. Econom.* **2005**, *20*, 873–889. <https://doi.org/10.1002/jae.800>.
26. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
27. Harvey, D.; Leybourne, S.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4).
28. Hansen, P.R.; Lunde, A.; Nason, J.M. The model confidence set. *Econometrica* **2011**, *79*, 453–497. <https://doi.org/10.3982/ECTA5771>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.