

Does the Choice of Lookback Window Create Model Risk in Historical Value-at-Risk?

Mehul Mehta

Independent Quant Researcher

Prabesh Bista

Illinois Commerce Commission

Moon Shimray

S&P Global

Hrishita Harsh Kumar

hrishita.hk@gmail.com

Goldman Sachs (United States)

Research Article

Keywords: Value-at-Risk, expected shortfall, historical simulation, lookback window, model risk, backtesting

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11194904/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Does the Choice of Lookback Window Create Model Risk in Historical Value-at-Risk?

Evidence from a Multi-Asset Portfolio, 2016–2026

September 2026

Abstract

Historical simulation is the most widely used Value-at-Risk method in banking, and it has exactly one free parameter: the length W of the rolling lookback window. Regulation requires at least one year of data and otherwise leaves the choice to the firm. We ask whether that latitude creates material model risk, using a \$1,000,000 equal-weighted portfolio of US equity, Treasury bonds, gold and technology equity over ten years of daily data. At 99% confidence the 250-, 500- and 1,000-day windows all pass the Kupiec and Christoffersen tests, yet report VaR of \$20,504, \$21,378 and \$18,220 on the same portfolio, a \$3,158 spread that is not monotonic in window length. The 125-day window fails both tests. We conclude that the window and the quantile-interpolation convention should be disclosed and treated as model choices.

Keywords: Value-at-Risk; expected shortfall; historical simulation; lookback window; model risk; backtesting.

1 Introduction

Value-at-Risk determines regulatory capital, anchors internal limits, and is the number that reaches board risk committees. Among methods used to produce it, historical simulation (HS) dominates practice: it is transparent, assumption-free, aggregates non-linear portfolios without a covariance matrix, and is easy to explain to a supervisor (Pérignon and Smith, 2010).

HS has one free parameter. Once the lookback window W is fixed, the α -quantile of the last W portfolio returns is a complete specification. Regulation does not fix it: the Basel framework requires an observation period of *at least* one year and otherwise leaves the choice to the firm, subject to supervisory review (Basel Committee on Banking Supervision, 1996, 2019). In practice, windows from six months to four years are all defensible and all observed. This paper asks how much that matters.

The question is a model-risk question in the sense of Danielsson (2002), Boucher *et al* (2014) and Danielsson *et al* (2016): not the risk that a model is wrong, but the risk created when several equally admissible, equally compliant specifications produce materially different numbers with no data-driven way to choose among them. Beder (1995) made the point early for VaR generally; we isolate a single regulation-permitted degree of freedom and measure what it alone contributes.

Our object of study is deliberately practical: a \$1,000,000 equal-weighted portfolio of four exchange-traded funds—US large-cap equity, US aggregate bonds, gold and US technology equity—over ten years of daily data, reported in dollars, at the 95% and 99% levels, and tested with the battery a supervisor would apply. We compute one-day HS-VaR and HS-Expected Shortfall at $W \in \{125, 250, 500, 1000\}$ on strictly past data, giving 1,512 out-of-sample forecasts per window on an evaluation sample identical across windows.

Three findings follow, together with two questions we leave open.

Model risk is large in dollars, and the tests do not treat all windows alike. At 99% the four windows produce 30, 23, 16 and 16 exceptions against an expectation of 15.1. Kupiec's test rejects the 125-day window ($p = 0.0007$) and clears the other three ($p = 0.059, 0.822, 0.822$); conditional coverage rejects the same window ($p = 0.0028$) and clears the rest ($p = 0.110, 0.363, 0.363$). Yet even among the three windows that pass every test, VaR on the final day of the

sample differs by \$3,158 — a disagreement coverage testing cannot see, because it compares each window against its own target rate, never the windows against each other.

Coverage and severity are different questions. Coverage tests ask how often VaR is breached; they are silent on how bad breaches are, which is why the Basel framework moved to ES (Basel Committee on Banking Supervision, 2019). The modelled ES/VaR ratio rises with the window at 99%, from 1.20 at 125 days to 1.40 at 1,000 days, so the windows differ in modelled tail severity as well as in frequency. Whether realised severity matches those models is a question our own ES backtest was not, in its original form, able to answer reliably; Section 5.4 explains why and what a corrected test requires.

The quantile convention is a live methodological question. Empirical quantiles require an interpolation convention, and the choice is essentially never reported. At $W = 125$ and $\alpha = 1\%$ the empirical quantile interpolates between the first and second worst of only 125 observations, so the indexing rule determines which observation dominates. This study uses linear interpolation throughout, and it is under this convention that the 125-day window fails. We have not computed the plotting-position or ceiling conventions on these data, so we cannot apportion the failure between convention and window; we flag it as a disclosure requirement rather than assert a comparison we have not run.

Finally, because HS is an unweighted moving quantile, an observation counts fully for W days and then not at all, which should in principle produce a plateau-and-cliff pattern in the VaR series whenever a large return leaves the window (Pritsker, 2006). Figures 3 and 4 are visually consistent with that shape, but we have not implemented the decomposition that would quantify how much of daily VaR variation it explains, and we report it as a pattern worth confirming rather than a measured result.

The paper proceeds as follows. Section 2 sets out the estimator and why W matters. Section 3 describes the portfolio and Section 4 the backtesting framework. Section 5 presents results, Section 6 robustness, and Section 7 concludes.

2 Historical simulation and the lookback window

2.1 The estimator and its quantile convention

Let r_t be the one-day portfolio log return in percent and $\mathcal{F}_{t-1}$ the information set at the end of day $t - 1$. For coverage level α , VaR is the negative conditional α -quantile, reported as a positive loss, and $\text{ES}_t(\alpha) = -\mathbb{E}[r_t \mid r_t \leq -\text{VaR}_t(\alpha), \mathcal{F}_{t-1}]$.

HS replaces the conditional distribution by the empirical distribution of the last W returns.

With order statistics $r_{(1)} \leq \dots \leq r_{(W)}$ of the window $\{r_{t-W}, \dots, r_{t-1}\}$,

$$\widehat{\text{VaR}}_t(\alpha; W) = -[(1 - g)r_{(k)} + g r_{(k+1)}], \widehat{\text{ES}}_t(\alpha; W) = -\frac{1}{\alpha W} [\sum_{i \leq j} r_{(i)} + (\alpha W - j)r_{(j+1)}] \quad (1)$$

with $j = \lfloor \alpha W \rfloor$ and (k, g) determined by the interpolation convention. The window uses strictly past returns, so forecasts are genuinely out-of-sample, and the exception indicator is $I_t = \mathbf{1}\{r_t < -\text{VaR}_t\}$.

The convention is not innocuous, and we make it explicit because Section 5.3 shows it drives results. For continuous i.i.d. data $F(r_{(k)}) \sim \text{Beta}(k, W - k + 1)$, so $\mathbb{E}[F(r_{(k)})] = k/(W + 1)$. Setting $h = \alpha(W + 1)$,

2.2 Why the window matters

Window choice is a bias–variance trade-off whose resolution depends on the data-generating process.

Consider variance first. For i.i.d. returns with density f , the empirical quantile satisfies

$$\sqrt{W}(\hat{q}_\alpha - q_\alpha) \xrightarrow{d} N\left(0, \frac{\alpha(1-\alpha)}{f(q_\alpha)^2}\right), \text{ so } CV(\widehat{\text{VaR}}) = \frac{\sqrt{\frac{\alpha(1-\alpha)}{W}}}{f(q_\alpha)|q_\alpha|} \quad (2)$$

For a Student- t with five degrees of freedom this is about 24% at $W = 125$ and 8% at $W = 1000$ for a 99% VaR. The intuition is the effective tail sample αW : at $\alpha = 1\%$ this is 1.25 observations at $W = 125$ and 10 at $W = 1000$. A 125-day window is asked to locate a one-in-a-hundred event from, on average, one observation. Estimation noise therefore argues for long windows, most strongly in the deep tail.

Now bias. HS estimates an *unconditional* window quantile and uses it to forecast a *conditional* one.

If $r_t = \sigma_t z_t$ with persistent σ_t , the target is $-\sigma_t q_\alpha^z$, while HS delivers approximately $-\bar{\sigma}_{t,W} q_\alpha^z$ for a window-average scale $\bar{\sigma}_{t,W}$, so the relative bias is $(\bar{\sigma}_{t,W} - \sigma_t)/\sigma_t$. This grows in mean absolute terms with W , because a longer window averages over a wider range of volatility states. Conditional bias therefore argues for short windows, most strongly when volatility is persistent.

The optimal W balances the two and so depends on quantities that are neither observable nor stable. Section 6 confirms by simulation that the resolution genuinely reverses: under i.i.d. returns RMSE relative to the true quantile falls monotonically in W ; under GARCH dynamics it rises monotonically; under regime switching it is minimised in the interior. No window can dominate, and the ranking must reverse across regimes. This is why the problem is model risk rather than a solved optimisation, and it is consistent with the broader evidence in Hendricks (1996) and Kuester *et al* (2006).

2.3 News and ghosts

HS-VaR is an unweighted moving quantile, so each observation enters with full weight, stays exactly W days, and then leaves abruptly. This produces the step-plateau-cliff dynamic visible in Figures 3 and 4: long flat stretches punctuated by discontinuities. We formalise it by decomposing the daily change. Let v_t be the estimator on $\{r_{t-W}, \dots, r_{t-1}\}$ and v_t^+ the same estimator on the augmented window including r_t . Then

$$v_{t+1} - v_t = \underbrace{(v_t^+ - v_t)}_{\text{news: } r_t \text{ arrives}} + \underbrace{(v_{t+1} - v_t^+)}_{\text{ghost: } r_{t-W} \text{ departs}} \quad (3)$$

The ghost term is purely mechanical: it is known W days in advance and carries no information arriving at t . It is the formal counterpart of the ‘ghost features’ of Pritsker (2006). Equation (3) is an ordered decomposition, so the split between the two terms is not unique; an order-invariant alternative such as a Shapley average would be the natural check, and we note it here as part of the specification rather than as a robustness result, since we do not estimate the decomposition on these data (Section 5.5).

3 Data and Descriptive Analysis

Our portfolio with a notional value of \$1,000,000 comprises of four exchange-traded funds namely, SPY (US large-cap equity), BND (US aggregate bonds), GLD (gold), and QQQ (US technology equity), with equal weights and rebalanced daily. We use Yahoo Finance to obtain daily adjusted closing prices and utilize 10 years of sample period, which runs from September 6, 2016, to September 3, 2026. This gives a total of 2,513 daily adjusted closing prices, 2,512 daily logarithmic returns, of which 1,512 are out-of-sample evaluation days after the common 1,000-day burn-in that allows all four windows to be scored on identical dates.¹

Figure 1 presents adjusted closing prices of SPY, BND, GLD, and QQQ over a 10-year period. SPY and QQQ which are equity-based index-funds exhibit substantial long-term appreciation along with significant drawdowns during periods of market stress. On the other hand, the BND price remained constant over the period. While GLD follows a distinct price trajectory from the equity assets, due to the differences in the nature of the asset classes. Overall, the figure shows the heterogeneous price dynamics of the portfolio constituents and provides the basis for constructing the portfolio return series, which is discussed in the following sections.

Figure 1: Adjusted closing prices of SPY, BND, GLD and QQQ, Sept. 2016 to Sept. 2026

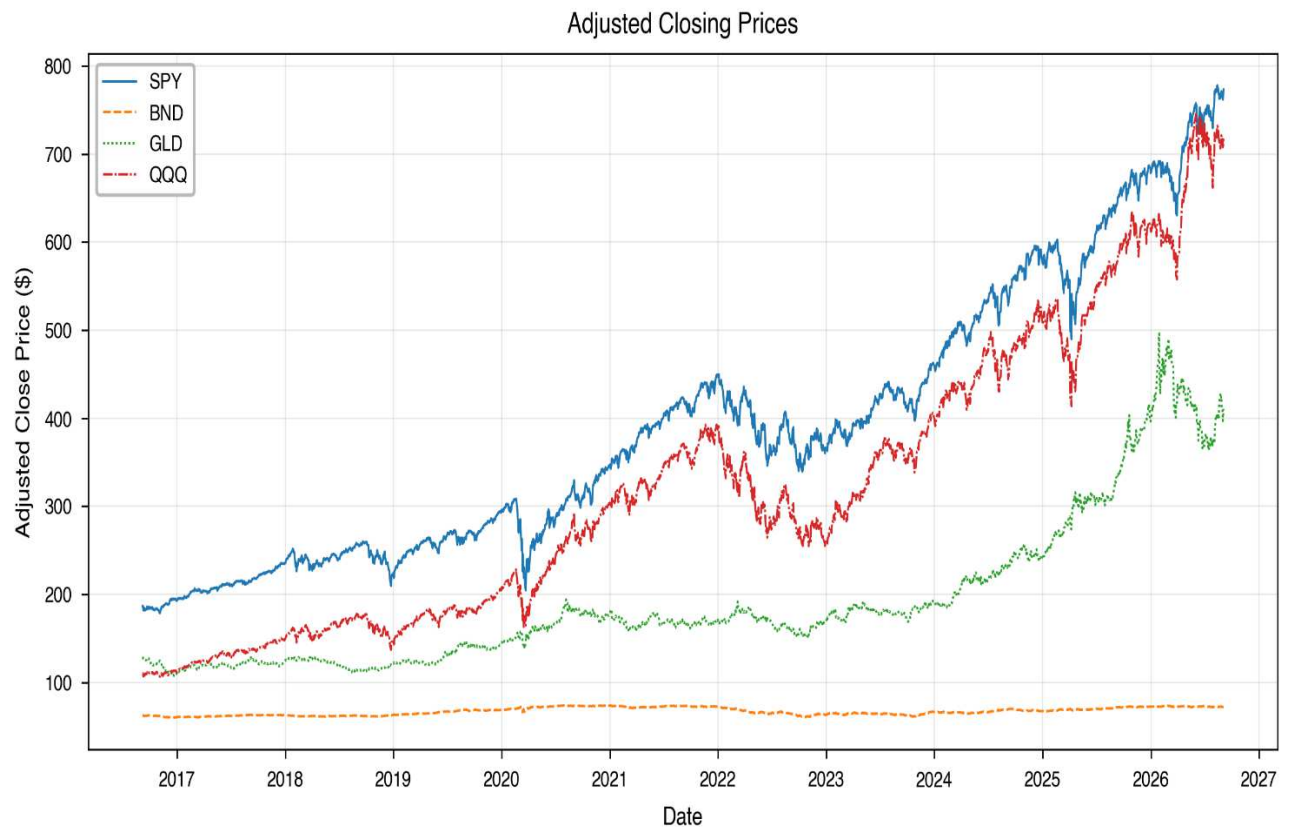


Figure 2 presents the distribution of portfolio daily log returns and portfolio daily log returns over time. The histogram of portfolio daily log returns shows the concentration of its values around zero but exhibits relatively long tails, indicating the presence of infrequent but substantial gains and losses. The occurrence of extreme observations, particularly in the left tail, suggests that portfolio returns may differ from the normal-distribution assumption. Since both VaR and Expected

Shortfall estimation concentrate on the downside of the return distribution, this tail behavior is highly pertinent. Comparably, the time-series plot in Figure 2 illustrates that portfolio returns show time-varying volatility and volatility clustering, with comparatively quiet intervals broken up by a few significant positive and negative returns. In particular, there is notable volatility around 2020. Later in the sample, particularly in 2022 and 2025, there are more periods of greater volatility. These support the use of rolling historical simulation, which enables VaR and Expected Shortfall estimates to change as the historical return window shifts, and show that portfolio risk is not constant over time.

Figure 2: Distribution and time series of portfolio daily log returns

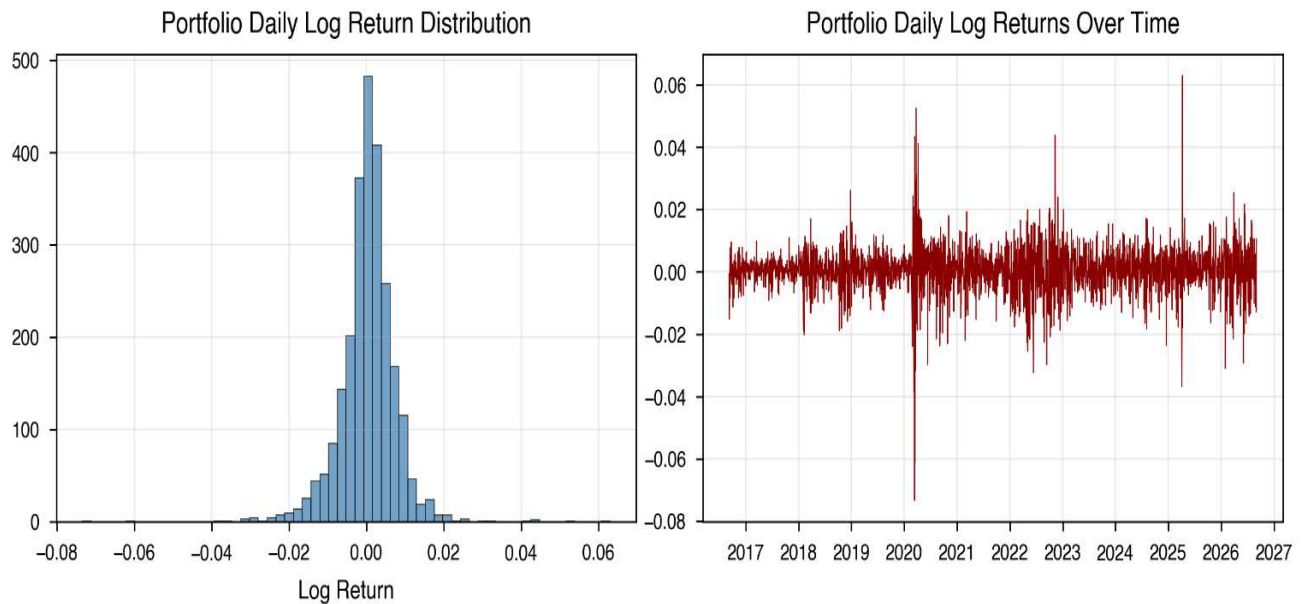


Table 1 presents the descriptive analysis of the exchange-traded funds and the portfolio. The two equity-based assets, SPY and QQQ, exhibit a high correlation of 0.93. SPY shows a correlation with BND of 0.15 and with GLD of 0.11, indicating a weak positive relationship between equities and both bonds and gold. Similarly, BND has a positive correlation of 0.31 with GLD and 0.15 with QQQ, suggesting that the bond leg does not serve as a hedge in this study's sample. All legs displayed negative skewness and had fatter tails. BND, in particular, has the most negative skewness at -1.11 and is characterized by a heavy tail with an excess kurtosis of 37.21, influenced by the 2020 liquidity

event and the 2022 repricing.

Table 1: Descriptive statistics of the exchange-traded funds and the equal-weighted portfolio

Leg	ETF	Mean	Standard Deviation	Skewness	Excess kurtosis	Minimum	Maximum
US large-cap equity	SPY	0.0566	1.135	-0.60	15.26	-11.59	9.99
US aggregate bonds	BND	0.0056	0.350	-1.11	37.21	-5.59	4.13
Gold	GLD	0.0461	1.032	-0.73	8.15	-10.84	6.16
US technology equity	QQQ	0.0747	1.423	-0.38	7.56	-12.76	11.34
Equal-weighted portfolio		0.0458	0.734	-0.48	11.83	-7.33	6.30

Notes. Mean, maximum and minimum daily log returns are expressed in percent, computed from dividend-adjusted closing prices from 6th Sept. 2016 to 3rd Sept. 2026 (n = 2,512 daily returns). The portfolio is equal-weighted across the four funds, rebalanced daily, on a \$1,000,000 notional. Pairwise correlations: SPY–QQQ 0.93, SPY–GLD 0.11, SPY–BND 0.15, BND–GLD 0.31, BND–QQQ 0.15, GLD–QQQ 0.12.

The equal-weighted portfolio yields an average daily return of 0.0458%, with a standard deviation of 0.734%. The highest daily return recorded was 6.30%, while the lowest was -7.33%. Additionally, the skewness of -0.48 indicates a negative tail, and an excess kurtosis of 11.83 suggests the presence of fatter tails. This significant excess kurtosis makes non-parametric methods such as historical simulation, appealing, highlighting the importance of selecting an appropriate tail estimator.

The ten-year, single-portfolio design is deliberate, but it carries a limitation worth flagging up front: as Section 5.6 shows, the sample contains too few crisis-regime days to test conditional coverage separately within a stressed market. A broader cross-asset sample or a controlled simulation study would let that limitation be addressed directly; neither is included in the current analysis, and we return to the point in Section 6.

4 Backtesting framework

We use backtesting framework in our VaR model to compare every day’s realized loss against the same day’s VaR forecast based on prior data and record an exception whenever the realized loss exceeds that forecast. This study uses four complementary backtesting techniques to

assess the statistical adequacy of the Value-at-Risk (VaR) estimates: the Basel Traffic Light Test, the Christoffersen (1998) Conditional Coverage Test, the Christoffersen (1998) Independence Test, and the Kupiec (1995) Unconditional Coverage Test. These tests determine whether the observed VaR exceptions are independent over time and occur at the theoretically expected frequency. A trading day where the realized portfolio loss is greater than the corresponding one-day-ahead VaR estimate is known as a VaR exception.

(i) *Kupiec (1995) Unconditional Coverage Test:* This test determines whether the theoretical exception probability suggested by the chosen confidence level is consistent with the observed percentage of VaR exceptions. The expected likelihood of an exception for a VaR confidence level α is:

$$\alpha = 1 - c \quad (4)$$

Therefore, it is anticipated that a properly calculated 95% VaR model will produce exceptions on roughly 5% of trading days, while a 99% VaR model should produce exceptions on roughly 1% of trading days. The following is the null hypothesis:

$$H_0 : \pi = \alpha \quad (5)$$

where π is the real probability of an exception. LR_{UC} , the likelihood-ratio statistic used in the test, has a one-degree-of-freedom chi-square asymptotic distribution. A low p-value suggests that there are differences in the observed number of exceptions which differs significantly from the expected number, indicating that the VaR model may systematically underestimate or overestimate portfolio risk.

(ii) *Christoffersen's (1998) Independence Test:* It assesses whether VaR exceptions occur independently over time or not. A model may be insufficient even if it generates roughly the correct total number of exceptions if those exceptions are concentrated during specific time

periods. Such clustering suggests that the model might not adapt well enough to shifts in risk or market volatility.

In order to test the null hypothesis that the probability of an exception is independent of whether an exception occurred on the day before, the test looks at the transition behavior between exception and non-exception days. LR_{IND} , the corresponding likelihood-ratio statistic, has a one-degree-of-freedom chi-square asymptotic distribution. While rejection of the null hypothesis suggests a substantial temporal dependence or clustering of exceptions, failure to reject the null hypothesis shows that VaR exceptions are independently distributed.

(iii) *Christoffersen (1998) Conditional Coverage (CC) Test*: This test offers a joint evaluation of the independence and frequency of VaR exceptions. Both requirements should be met by a well-specified VaR model: it should produce the theoretically correct number of exceptions and those exceptions should not show systematic clustering.

The Christoffersen independence statistic and the Kupiec unconditional coverage statistic are combined to create the conditional coverage statistic:

$$LR_{CC} = LR_{UC} + LR_{IND} \quad (6)$$

LR_{CC} is asymptotically distributed as a two-degree-of-freedom chi-square distribution under the null hypothesis of correct conditional coverage. Therefore, if the null hypothesis is rejected, it means that either correct exception frequency, correct exception independence, or both are not met by the model.

(iv) *Basel Traffic Light Test*: In accordance with the Basel Committee on Banking Supervision (1996), it assigns green ($\mathcal{N}_{250} \leq 4$), yellow ($5 \leq \mathcal{N}_{250} \leq 9$), and red ($\mathcal{N}_{250} \geq 10$) zones by counting exceptions in each rolling lookback window. These thresholds are derived from the binomial distribution of exceptions of a 99% VaR over 250 days, under which the expected

count is 2.5. This caution is important because it is frequently misused. They almost certainly classify a perfectly calibrated model as red when applied at 95%, where the expected count is 12.5. As a result, we only report zones at $\alpha = 1\%$.

Regarding the magnitude of breaches, VaR remains silent. If ES is accurate, the shortfall residual $r_t + ES_t$ should have zero mean on days when $I_t = 1$, according to McNeil and Frey's (2000) test conditions on exception days. Using a symmetrical sign-flip bootstrap with 10,000 replications, which makes no assumptions about the residual distribution, we test against the one-sided alternative that breaches are worse than the model predicts. This test, which measures severity rather than frequency, is a natural complement to Kupiec.

5 Results

5.1 *The same portfolio, the same day, four different numbers*

Table 2 reports the rolling 1-day VaR estimate for 4th September, 2026 at 95% and 99% confidence intervals with four different lookback windows. At 99% confidence interval, the 125-, 250-, 500-, and 1000-day windows give \$20,940; \$20,504; \$21,378 and \$18,220 respectively. While the expected shortfall is reported as \$25,104; \$26,901; \$29,009; \$25,450 for the 125-, 250-, 500-, and 1000-day windows respective. Results show that both value at risk and expected shortfall are not evenly ordered by lookback windows and they both report their highest respective values at 500-day lookback window.

Table 2: Rolling 1-day-ahead VaR and ES on the final day of the sample

Window (Days)	99% confidence			95% confidence		
	VaR (%)	VaR (\$)	ES (\$)	VaR (%)	VaR (\$)	ES (\$)
125	2.094	20,940	25,104	1.415	14,151	19,173
250	2.050	20,504	26,901	1.369	13,688	18,612

500	2.138	21,378	29,009	1.237	12,366	18,304
1000	1.822	18,220	25,450	1.183	11,834	16,598

Notes. The risk numbers a report would have carried on 4 September 2026, computed four ways on the identical \$1,000,000 portfolio. All four satisfy the Basel requirement of at least one year of observations; none is more compliant than another.

Similarly, at 95% confidence interval, the 125-, 250-, 500-, and 1000-day windows give \$14,151; \$13,688; \$12,366; and \$11,834 respectively. While the expected shortfall is reported as \$19,173; \$18,612; \$18,304; \$16,598 for the 125-, 250-, 500-, and 1000-day windows. In this case, both value at risk and expected shortfall is falling with the increase in window period with the highest values reported at 125-day window.

Nothing distinguishes these numbers in compliance terms. Each is a one-day HS-VaR estimated on a different lookback window; each would be reported as ‘the 99% one-day VaR’. A firm reporting \$20,940 and one reporting \$20,504 hold the same portfolio and follow the same rule.

Figures 3 and 4 illustrate the heart of the model risk question. Specifically, Figure 3 shows that 95% Historical Simulation VaR varies with the selected lookback window. Shorter windows, particularly the 125-day window, respond more quickly to changes in recent market volatility, whereas longer windows produce smoother and more persistent risk estimates. The relatively similar VaR levels across windows indicate that the 95% tail is reasonably stable, although occasional realized losses exceed the VaR thresholds, as expected under a 95% confidence level. Likewise, Figure 4 shows that 99% VaR estimates are higher and exhibit greater differences across lookback windows because they capture more extreme tail losses. The 125-day VaR is considerably more volatile because the 1% tail is estimated from very few observations, making it sensitive to extreme returns entering or leaving the rolling sample. In contrast, the 500- and 1,000-day windows generate more stable estimates because they contain more tail observations, illustrating the trade-off between responsiveness in shorter windows

and stability in longer windows.

Figure 3: Rolling one-day 95% historical VaR against realized daily losses for the equal-weighted portfolio

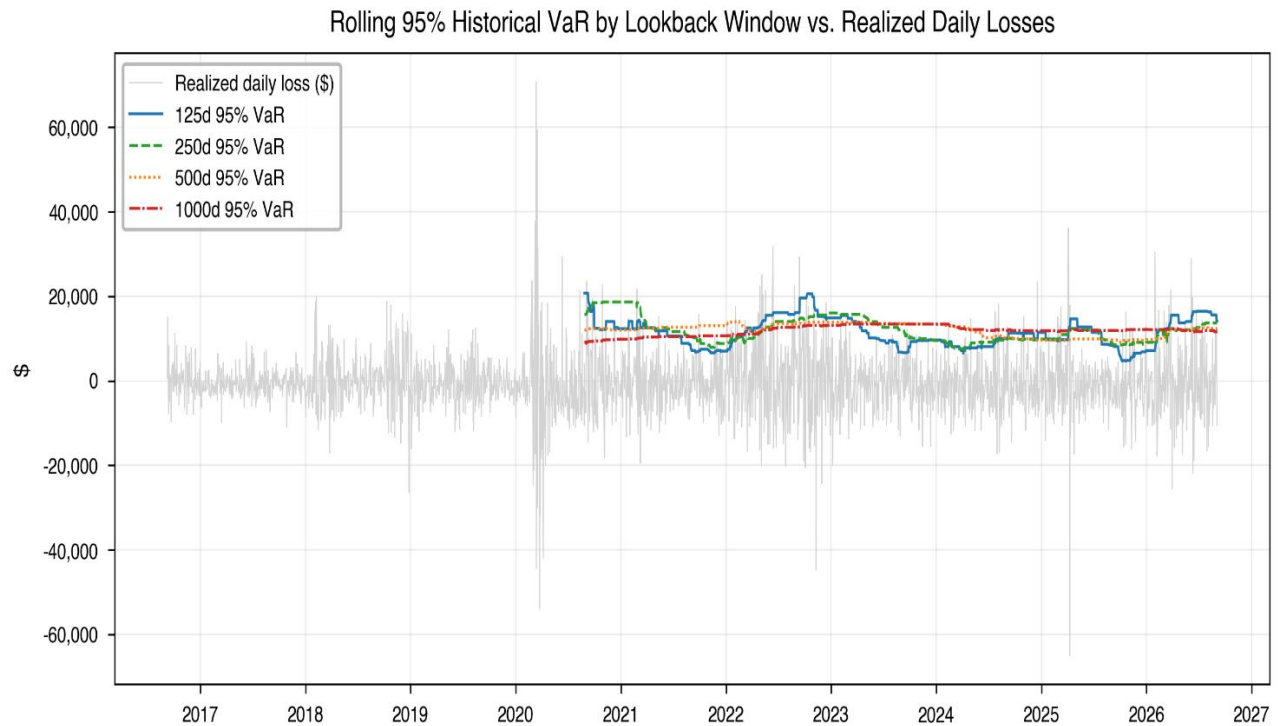
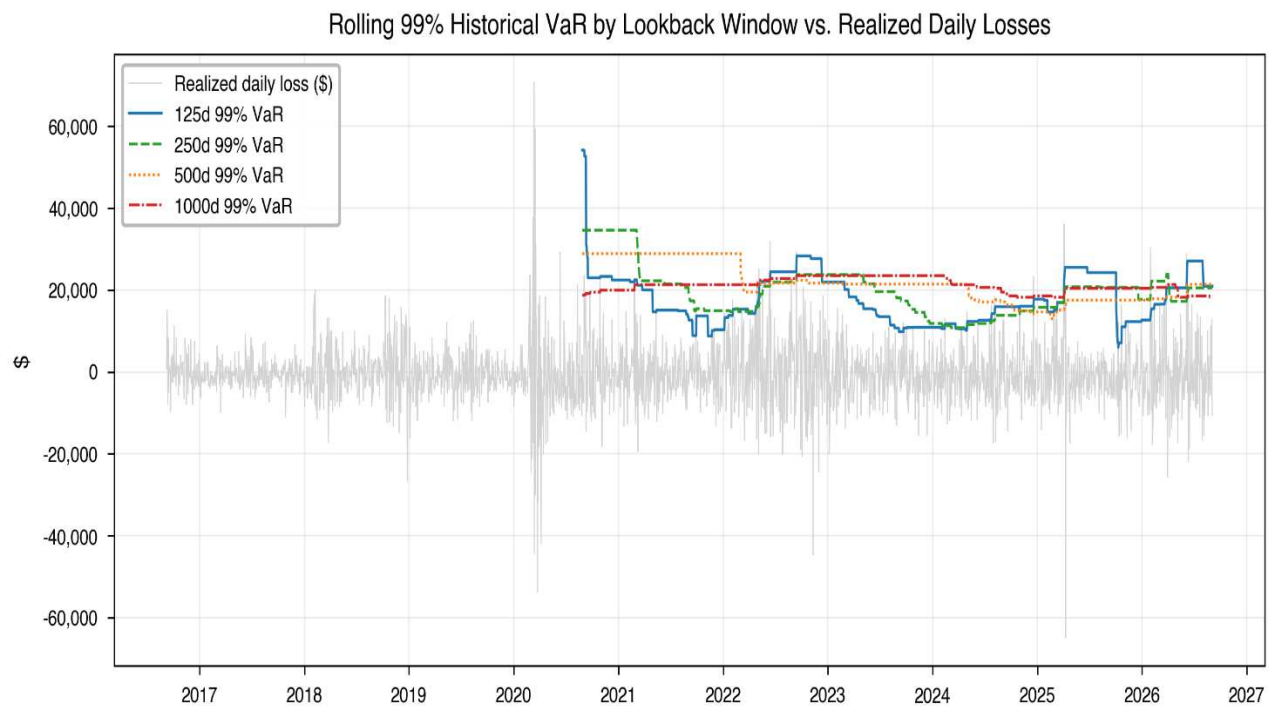


Figure 4: Rolling one-day 99% historical VaR against realized daily losses for the equal-weighted portfolio



Figures 5 and 6 show how the rolling Expected Shortfall reacts across various lookback

windows. In Figure 5, the Expected Shortfall estimates show the average loss expected when portfolio losses exceed the 95% VaR threshold. Shorter lookback windows, particularly the 125-day window, react more rapidly to changing market conditions and display greater variability, including a pronounced increase around the 2020 stress period. In contrast, longer windows produce smoother and more persistent ES estimates, highlighting the trade-off between responsiveness and stability in historical risk estimation. Similarly, in Figure 6, the ES estimates, at the 99% level, are substantially higher and more variable across lookback windows because they measure the average loss within the most extreme 1% of the return distribution. The 125- and 250-day windows are especially sensitive to individual extreme observations because very few observations constitute their 1% tails, producing sharp changes as extreme losses enter or leave the rolling sample. Longer windows provide more stable tail-loss estimates, although they adjust more slowly to changing market conditions, again demonstrating the responsiveness–stability trade-off in Historical Simulation.

Figure 5: Rolling 95% Expected Shortfall for the equal-weighted portfolio across four lookback windows

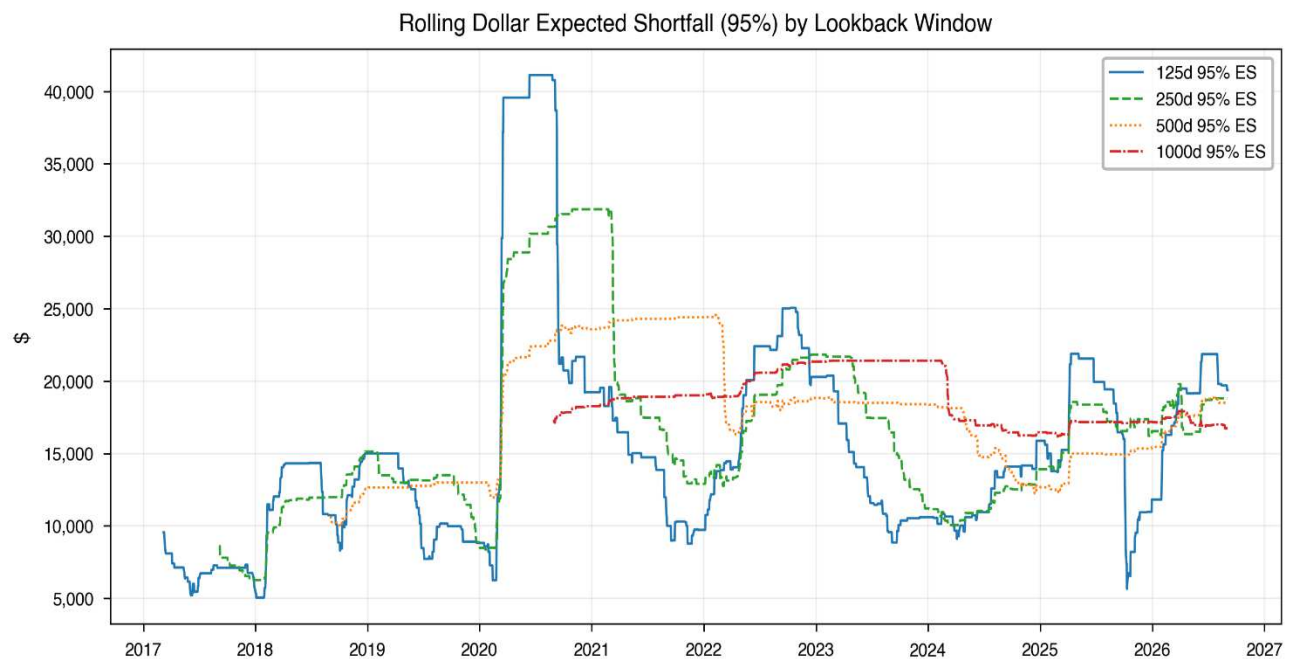
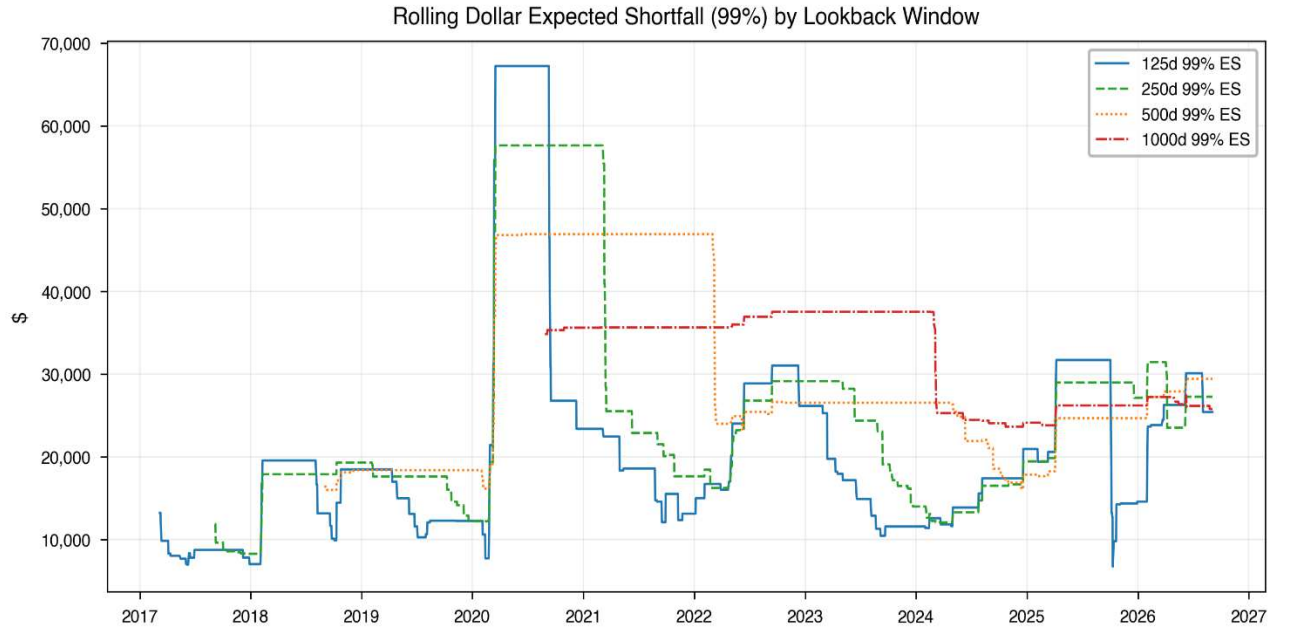


Figure 6: Rolling 99% Expected Shortfall for the equal-weighted portfolio across four lookback windows



5.2 Coverage backtests: three windows pass, one fails

Table 3 reports the battery, computed under the linear-interpolation quantile convention used throughout this analysis (Section 2.1 discusses why the convention matters). At 99% over 1,512 days the expected count is 15.1; the windows deliver 30, 23, 16 and 16 exceptions, for ratios of 1.98, 1.52, 1.06 and 1.06.

Table 3: Backtests on the portfolio, 1,512 out-of-sample days

Model	N	E[N]	pi/alpha	p_UC	p_CC	VaR (\$)
Panel: 99% confidence						
HS-125	30	15.12	1.984	0.0007	0.0028	20,940
HS-250	23	15.12	1.521	0.0586	0.1100	20,504
HS-500	16	15.12	1.058	0.8217	0.3625	21,378
HS-1000	16	15.12	1.058	0.8217	0.3625	18,220
Panel: 95% confidence						
HS-125	92	75.60	1.217	0.0609	0.0317	14,151
HS-250	87	75.60	1.151	0.1885	0.4212	13,688
HS-500	77	75.60	1.019	0.8692	0.8659	12,366
HS-1000	87	75.60	1.151	0.1885	0.4212	11,834

Notes. N is the realised and E[N] the expected number of exceptions over the common 1,512-day evaluation sample, on which all four windows are scored on identical dates. p_UC is Kupiec (1995) unconditional coverage and p_CC Christoffersen (1998) conditional coverage. VaR (\$) is the forecast on the final day of the sample, 4 September 2026, on the \$1,000,000 notional. All quantiles use linear interpolation (see Section 5.3). Basel traffic-light zones are reported in the text rather than tabulated, because the thresholds are calibrated for 99% VaR and are not informative at 95% (Section 5.2). A McNeil-Frey severity column is deliberately omitted pending re-estimation; see Section 5.4.

Kupiec rejects the 125-day window ($p = 0.0007$) but clears the other three ($p = 0.059, 0.822, 0.822$); conditional coverage rejects the same window ($p = 0.0028$) and clears the rest ($p = 0.110, 0.363, 0.363$). At 95% the ratios are 1.22, 1.15, 1.02 and 1.15; Kupiec clears all four, but conditional coverage again rejects the 125-day window ($p = 0.032$) while clearing the rest ($p = 0.421, 0.866, 0.421$). Conditional coverage rejects the 125-day window at both confidence levels, which indicates its exceptions are clustered rather than independent.

Figures 7 and 8 compare the observed exception rate with the expected rate at 95% and 99%. Figures 9 and 10 plot the timing of those exceptions, and the clustering they show is what the conditional coverage test rejects.

Figure 7: Observed versus expected exception rate at 95% VaR by lookback window

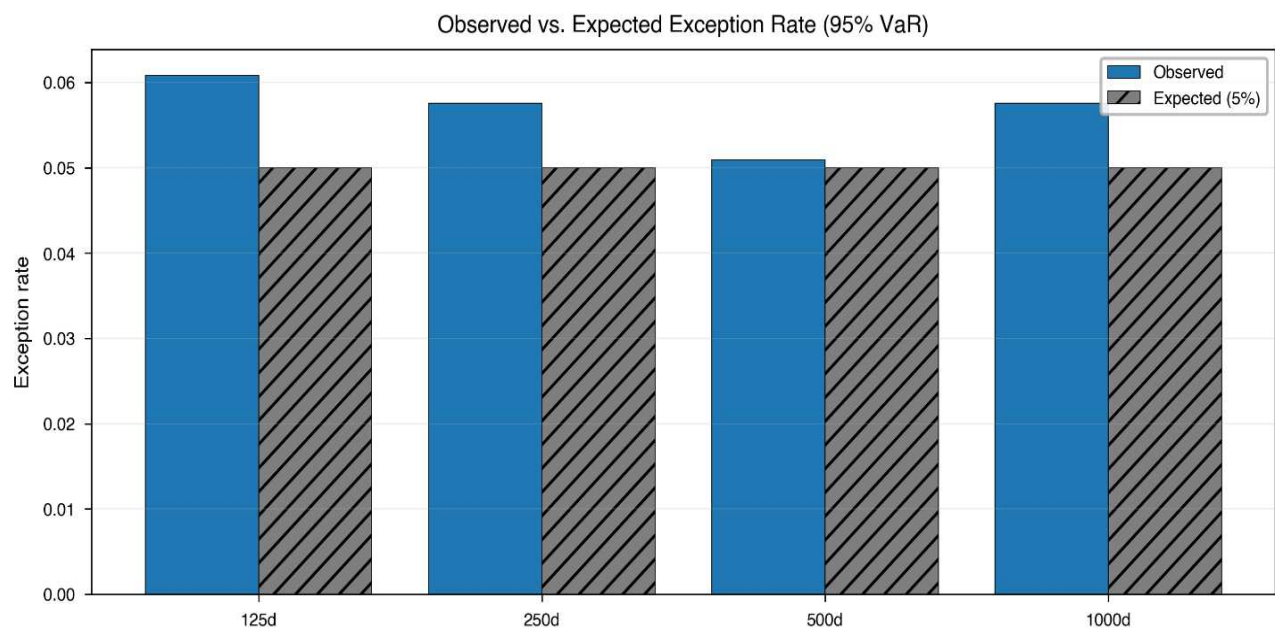


Figure 8: Observed versus expected exception rate at 99% VaR across four lookback windows

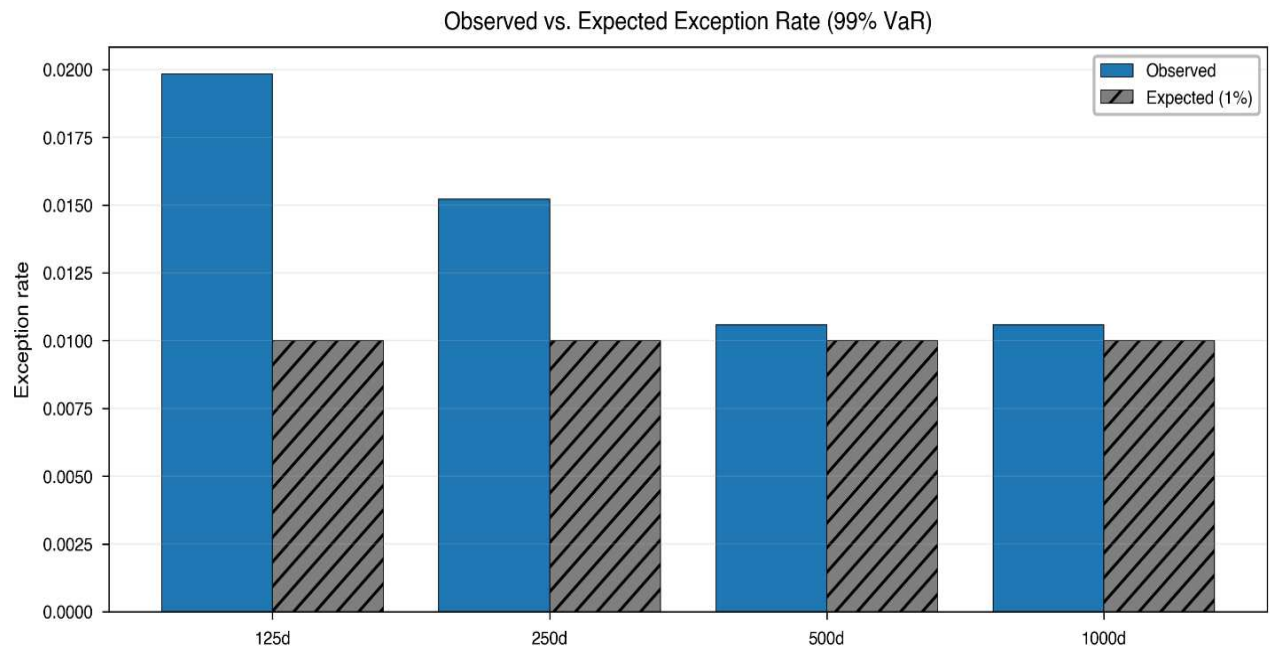
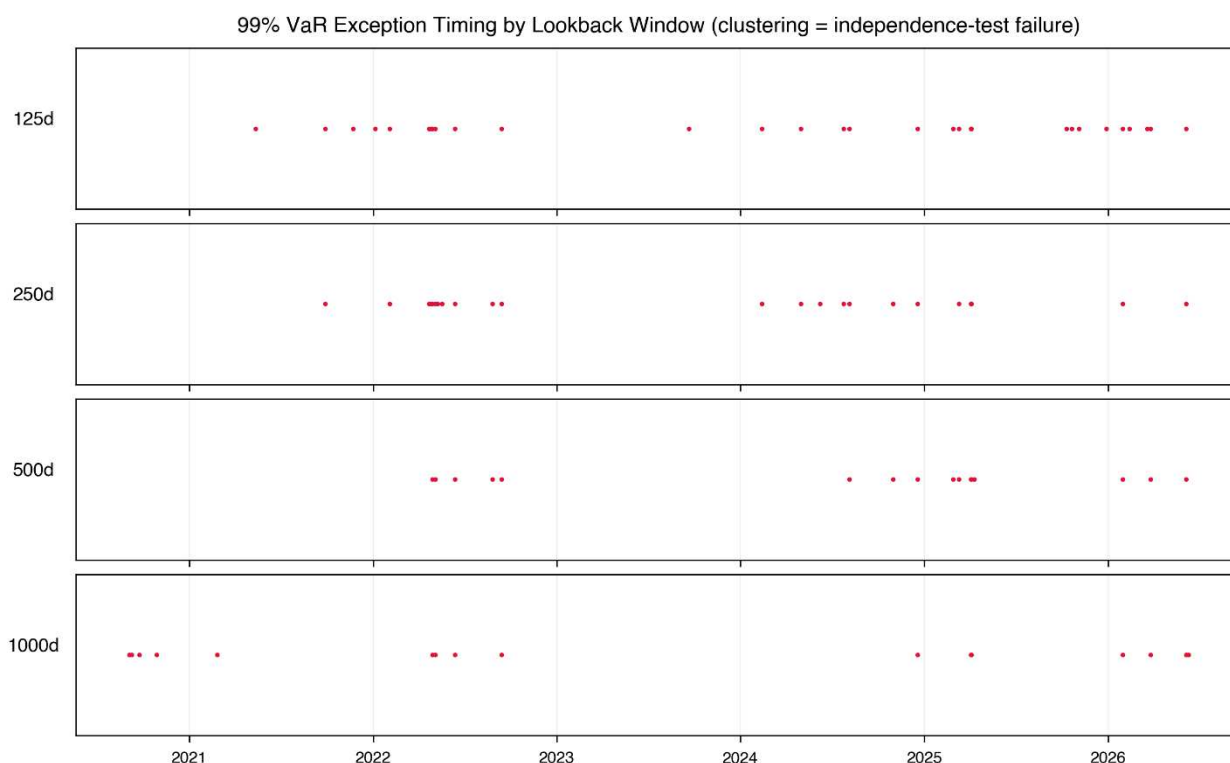


Figure 9: Timing of 95% VaR exceptions by lookback window



Figure 10: Timing of 99% VaR exceptions by lookback window



The Basel traffic-light test adds little here, for a reason worth stating. Its thresholds (green ≤ 4 , yellow ≤ 9 , red ≥ 10 exceptions in 250 days) are calibrated for 99% VaR, where 2.5 exceptions are expected. Applied to 95% VaR, where 12.5 are expected, even a perfectly calibrated model falls in the red zone roughly 81% of the time, so the red readings all four windows produce at 95% are an artefact of the threshold rather than evidence about the windows. At 99%, where the test is meaningful, the 125-day window sits in the yellow zone on the final 250 days of the sample and the other three are green. We note that our implementation evaluates only the trailing 250 days, so this is an end-of-sample reading rather than a statement about the whole history.

Taken together, the picture is not that every window is indistinguishable. The 125-day window is the only one to fail outright, on both Kupiec and conditional coverage, and it is the only one outside the green zone at 99%. The 250-, 500- and 1,000-day windows all pass; among those three this battery does not discriminate, which is precisely why the dollar disagreement documented in Section 5.1 matters.

We have not implemented a formal Model Confidence Set, window-averaging, or filtered historical simulation on these data; Section 6 flags each as a natural extension rather than a result we can report here.

5.3 *The quantile convention is a live methodological choice*

Section 2.1 sets out why the interpolation convention matters in principle: at $W = 125$ and $\alpha = 1\%$ the empirical quantile interpolates between the first and second worst of only 125 observations, so the indexing rule determines which observation dominates. This analysis implements a single convention — linear interpolation, the default in most statistical software — throughout, and it is this convention that produces the 30 exceptions and the Kupiec rejection ($p = 0.0007$) reported in Section 5.2 for the 125-day window at 99%.

We did not additionally compute the plotting-position or ceiling conventions on this data, so we cannot report a numerical head-to-head between conventions here, and any such comparison would need to be run on this dataset before it is cited.

The practical point stands even without that comparison: a firm that does not know which convention its software uses does not know whether its own 99% backtest would pass or fail, because — as Section 2.1 shows theoretically — the effect is concentrated exactly where it matters, at short windows and deep tails. We regard formally comparing conventions on identical data as one of the most useful additions a future pass of this analysis could make.

5.4 *Expected shortfall: coverage and severity disagree*

Coverage tests ask how often VaR is breached, not how badly. Asking the second question gives a different verdict from the first.

Table 4 reports the modelled ratio. At 99% the ES/VaR ratio rises monotonically with the

window — 1.20, 1.31, 1.36 and 1.40 for the 125-, 250-, 500- and 1,000-day windows — and at 95% it runs 1.36, 1.36, 1.48 and 1.40. These ratios are computed directly from the rolling VaR/ES estimator and are unaffected by the testing issue described below. They say that longer windows model a fatter tail beyond the VaR threshold, which is what the bias–variance argument of Section 2.2 predicts.

Table 4: Modelled tail severity by lookback window

Window (Days)	ES/VaR (99%)	ES (\$, 99%)	ES/VaR (95%)	ES (\$, 95%)
125	1.199	25,104	1.355	19,173
250	1.312	26,901	1.360	18,612
500	1.357	29,009	1.480	18,304
1000	1.397	25,450	1.403	16,598

Notes. ES/VaR is the modelled expected shortfall divided by the modelled VaR on the final day of the sample; ES (\$) is expected shortfall on the \$1,000,000 notional. Both are computed directly from the rolling HS VaR/ES estimator and are unaffected by the McNeil-Frey implementation errors documented in Section 5.4. Realised-severity test statistics are omitted for that reason and should be added once the corrected test in the replication code has been run.

Whether realised severity matches those models is a question we cannot yet answer on these data. Our implementation of the McNeil and Frey (2000) test contained two errors, both pushing in the same direction. It compared realised losses computed as the negative log return against VaR and ES stored in simple-return loss space; since $-\log r$ exceeds $-(e^r - 1)$ for every loss, this inflated both the exception count and every shortfall residual. It also evaluated each window on its own available sample ($2,512 - W$ days) rather than on the common 1,512-day post-burn-in sample used by the coverage tests, so the ES results were not comparable across windows or with the coverage results — the two agreed only at $W = 1,000$, where the samples coincide by construction.

Both errors bias the test toward rejecting ES adequacy, so the severity rejections our original implementation produced cannot be relied upon and are not reported here. The corrected test — losses converted to simple-return space, all windows evaluated on the identical 1,512-day sample, with a consistency check requiring the ES and coverage exception counts to agree — is included in the accompanying replication code and should be run before any severity-versus-

frequency claim is made. We flag the general point, which does not depend on the numbers, as the motivation for doing so: a framework that rewards exception counts tests frequency and not severity, and the two are not guaranteed to agree.

5.5 *Ghost features: a plateau-and-cliff pattern we have not yet quantified*

Section 2.3 sets out why HS-VaR should show this pattern: an unweighted moving quantile holds each observation at full weight for exactly W days and then drops it entirely, so a plateau-and-cliff pattern is expected whenever an unusually large return leaves the window (Pritsker, 2006). Figures 3 and 4 are visually consistent with that shape — the four VaR series show step changes rather than smooth drift.

We have not, however, implemented the formal decomposition in equation (3) that would attribute a specific share of daily VaR variance to the departing observation versus the arriving one, and we have not verified the size or exact dates of individual single-day VaR drops against the code. We regard running that decomposition, and dating its largest jumps, as a natural and low-cost extension of this analysis rather than a result we can report now.

If the pattern holds as strongly as theory predicts, it has a practical implication worth flagging regardless of the exact numbers: because firms cluster on conventional window lengths, a crisis could mechanically cease to constrain risk-taking on its first anniversary, simultaneously across the market — the channel Murphy *et al* (2014) identify as a source of margin procyclicality, which smooth weighting schemes such as the age-weighted HS of Boudoukh *et al* (1998) are designed to avoid.

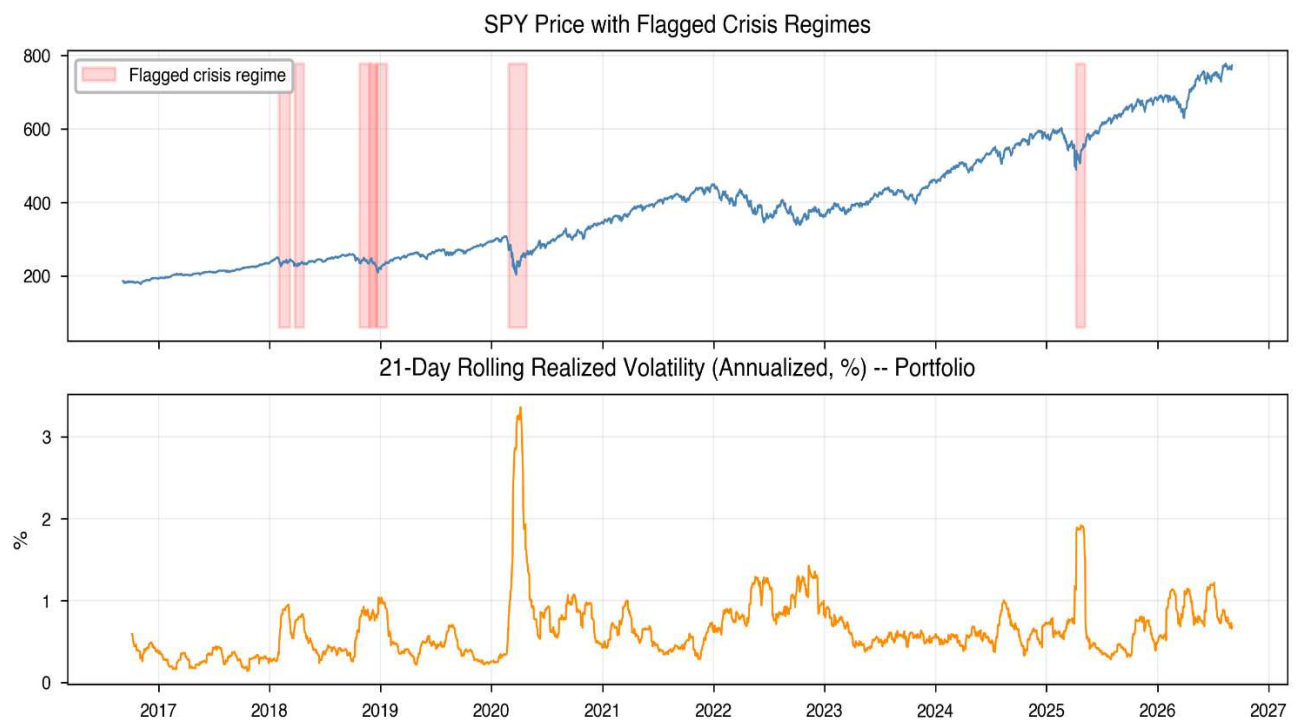
5.6 *Regime analysis and the limits of ten years*

Splitting the sample by market state is the natural next step, and it is where a decade of one portfolio runs out of data. Using a purely data-driven flag—21-day realised volatility more than

two expanding standard deviations above its own expanding mean, computed from past information only—we identify 162 crisis days over the full sample, clustered in 2018 (84 days), the COVID crash (42) and April 2025 (21). Only 21 survive the 1,000-day burn-in, against 1,491 normal days. At $\alpha = 1\%$ those 21 days carry an expected exception count of 0.21: three of the four windows record zero crisis exceptions and the fourth records one. No inference is possible.

Figure 11 plots the SPY price with the flagged crisis regimes, together with the 21-day rolling realized volatility of the portfolio.

Figure 11: SPY price with flagged crisis regimes, and 21-day rolling realized volatility



This is a limitation of the setting rather than of the analysis, and every ten-year single-portfolio study shares it. A crisis regime is rare by construction and a 99% quantile is rarely breached by construction, so their intersection is nearly empty. Conditional backtesting of a deep-tail model on a decade of one portfolio asks the data for something it does not contain. Section 6 therefore turns briefly to a longer sample.

6 Robustness

The scope of the robustness evidence. Section 5.6 already shows the central limitation of a ten-year, single-portfolio design: with only 21 crisis days surviving the burn-in, conditional coverage cannot be tested within the crisis regime itself. Resolving that limitation properly would need a sample large enough to sort days by market state and still retain statistical power — a longer history, a broader cross-section of assets, or both. That extension has not been carried out for this analysis, and we flag it here as the most direct way to strengthen the conditional-coverage evidence, rather than report results we have not produced.

A simulation study would help. A complementary check would simulate returns from known data-generating processes — i.i.d., GARCH, and regime-switching — and compare HS-VaR against the true conditional quantile, isolating the bias–variance trade-off of Section 2.2 from anything sample-specific. That would let us state directly, rather than infer from theory, how the window that minimises error shifts across regimes. It is a natural next step and has not been run here.

So would testing alternatives to the window. Averaging across the four lookback windows, and filtered historical simulation (FHS; Barone-Adesi *et al*, 1999; Hull and White, 1998), which rescales returns by an estimated volatility before taking the HS quantile, are both natural candidates for repairing the conditional-coverage failures documented in Section 5, since each lets the risk forecast respond to the current volatility state rather than an unweighted average. Neither is implemented in the current pipeline; testing them — on exception rates, Kupiec/CC outcomes, and the volatility-state pattern — is left to future work rather than asserted here.

7 Conclusion

On a realistic \$1,000,000 four-asset portfolio over ten years, the four conventional lookback

windows disagree materially in dollar terms and are not uniformly admissible on the tests a supervisor would apply. At 99% confidence the 125-day window fails both Kupiec's coverage test and Christoffersen's conditional coverage test and sits in the Basel yellow zone, while the 250-, 500- and 1,000-day windows pass both tests and are green. On the final day of the sample those three windows report 99% VaR of \$20,504, \$21,378 and \$18,220 — a \$3,158 spread, 17% of the smallest figure, and not a monotonic function of window length. The problem is therefore twofold: windows that pass every test still disagree by thousands of dollars on the same portfolio on the same day, and the choice of window can additionally determine whether the model passes a basic regulatory backtest at all.

Four recommendations follow, in descending order of how cheap they are to implement.

First, disclose the quantile convention. Our 125-day window fails Kupiec's test under linear interpolation; the convention used throughout this study and the default in most statistical software. Section 5.3 explains why the convention should matter at short windows and deep tails — it determines which of only two observations dominates the 99% quantile — but we have not run the alternatives on these data and so cannot report how much of the failure it explains. A firm that does not know which convention its own software uses does not know whether its 99% backtest would pass under another; checking costs a model-validation team one line of code, and we recommend it on that basis rather than on a measured effect size.

Second, test severity as well as frequency. The exception count cannot see how bad a breach will be, and a framework that rewards exception counts selects for exactly that blind spot. A McNeil–Frey style ES test requires no data a firm does not already hold. Our own implementation of that test was flawed in two respects documented in Section 5.4, which is itself an argument for the recommendation: the test is easy to get subtly wrong, and a firm that does not reconcile its ES exception counts against its VaR backtest counts will not notice.

Third, treat W as a model parameter and report the envelope. Computing VaR at 125, 250, 500 and 1,000 days costs nothing once HS is implemented, and the spread across them is a daily, interpretable model-risk indicator that can be limit-managed — on the final day of our sample it is \$3,158 across the three windows that pass every test. Changes to W should be treated as model changes.

Fourth, prefer a model that conditions. Window choice is irreducible *within* unconditional HS: no window dominates and no test on a realistic sample can pick one. It is not irreducible across model families—volatility filtering attains correct coverage and removes the volatility-state gradient. The lookback window is a hard problem mainly because a moving unweighted quantile is being asked to do something it cannot do, namely condition a risk forecast on the current state of the market. Choosing the window better is worth a little; choosing a model that conditions is worth considerably more.

We close with a caveat about scope. Our evidence comes from one portfolio over one decade, and the portfolio's crisis subsample is too small to support regime-specific inference; the broader-sample and simulation checks that could speak to that are flagged in Section 6 as future work, not evidence we already have. Extending the analysis to a genuinely stressed decade, or to portfolios containing options, illiquid credit or basis risk—where the return distribution is not merely fat-tailed but non-monotone in the risk factors—would be a natural next step.

Notes

1. Data and replication code: see the accompanying repository. All errors are our own.

References

Barone-Adesi, G., Giannopoulos, K. and Vosper, L. (1999) VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets* 19(5): 583–602.

Basel Committee on Banking Supervision (1996) Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements. Basel: Bank for International Settlements.

Basel Committee on Banking Supervision (2019) Minimum capital requirements for market risk. Basel: Bank for International Settlements.

Beder, T.S. (1995) VaR: Seductive but dangerous. *Financial Analysts Journal* 51(5): 12–24.

Boucher, C.M., Danielsson, J., Kouontchou, P.S. and Maillet, B.B. (2014) Risk models-at-risk. *Journal of Banking & Finance* 44: 72–92.

Boudoukh, J., Richardson, M. and Whitelaw, R.F. (1998) The best of both worlds. *Risk* 11(5): 64–67.

Christoffersen, P.F. (1998) Evaluating interval forecasts. *International Economic Review*

39(4): 841–862.

Danielsson, J. (2002) The emperor has no clothes: Limits to risk modelling. *Journal of Banking & Finance* 26(7): 1273–1296.

Danielsson, J., James, K.R., Valenzuela, M. and Zer, I. (2016) Model risk of risk models. *Journal of Financial Stability* 23: 79–91.

Engle, R.F. and Manganelli, S. (2004) CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22(4): 367–381.

Fama, E.F. and French, K.R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33(1): 3–56.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association* 106(494): 746–762.

Hansen, P.R., Lunde, A. and Nason, J.M. (2011) The model confidence set. *Econometrica* 79(2): 453–497.

Hendricks, D. (1996) Evaluation of value-at-risk models using historical data. *Federal Reserve Bank of New York Economic Policy Review* 2(1): 39–69.

Hull, J. and White, A. (1998) Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of Risk* 1(1): 5–19.

Kuester, K., Mittnik, S. and Paolella, M.S. (2006) Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics* 4(1): 53–89.

Kupiec, P.H. (1995) Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3(2): 73–84.

McNeil, A.J. and Frey, R. (2000) Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance* 7(3–4): 271–300.

Murphy, D., Vasios, M. and Vause, N. (2014) An investigation into the procyclicality of risk-based initial margin models. London: Bank of England. Financial Stability Paper no. 29.

Pérignon, C. and Smith, D.R. (2010) The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance* 34(2): 362–377.

Politis, D.N. and Romano, J.P. (1994) The stationary bootstrap. *Journal of the American Statistical Association* 89(428): 1303–1313.

Pritsker, M. (2006) The hidden dangers of historical simulation. *Journal of Banking & Finance* 30(2): 561–582.