How Well Does a Rolling-Volatility Band Calibrate? Evidence Across Asset Classes and Market Regimes*

M. A. H. AlEssa
oisigma.com LLC

Working paper. This version: 19 June 2026.
Not peer-reviewed; comments welcome. Latest version at oisigma.com.

Abstract

Traders need a stable, asset-agnostic way to separate ordinary price fluctuation from statistically unusual movement, and the practitioner bands in common use are not calibrated to that question. We study a deliberately simple, causal volatility envelope—a band built from the rolling mean and standard deviation of recent returns, anchored on the prior close—and show that it is a *stably calibrated* classifier of normal versus abnormal next-bar behavior across asset classes, a century of regimes, and daily-to-monthly time scales. That invariance, not the level of any single number, is the finding. On the S&P 500 (1927–2024) the inner ($\pm 1\sigma$) band contains the next close 71.20% of the time and stays within 68.7–73.7% in every calendar decade; across a forty-asset universe the mean is 71.9% (cross-asset SD 2.3pp), and the result survives out-of-sample splits, an independently randomized universe, alternative source prices, and weekly and monthly frequencies. Because the band plugs in an estimated $\hat{\mu}, \hat{\sigma}$, the proper null is the finite-window (prediction-t) quantile—67.46% at $n = 60$, not the known-parameter 68.27%. The +3.7pp excess over that null is the fingerprint of return leptokurtosis rather than an artifact: a GARCH(1,1) with Student-$t(\nu = 6)$ innovations reproduces the empirical rate within 0.08pp, and a probability-integral-transform analysis shows the departure from Gaussian is one of *shape* (heavy tails), not scale. The return-space construction is what carries the calibration: a standard price-space Bollinger band catches only $\sim 83\%$ of closes within its nominal 95% envelope (versus $\sim 94\%$ here) and is beaten on all forty assets, whereas the choice of variance estimator matters far less. The envelope is a *marginal-coverage* classifier—well calibrated unconditionally but not within every regime (coverage falls to 65.3% at crisis onset, VIX>30)—so it classifies next-bar dispersion rather than direction. Headline results survive Holm–Bonferroni correction at family-wise $\alpha = 0.05$.

Keywords: volatility envelopes; rolling volatility; calibration; prediction interval; density-forecast evaluation; value-at-risk; leptokurtosis; conditional volatility; technical analysis; Bollinger Bands.

*The author used Anthropic's Claude and OpenAI's ChatGPT as research and drafting assistants for portions of this work, including literature review, code generation, statistical exploration, and editorial revision. All numerical results, model choices, code, and editorial decisions are the author's sole responsibility.

JEL Classification: G17 (Financial Forecasting and Simulation); G14 (Information and Market Efficiency); C58 (Financial Econometrics); C52 (Model Evaluation, Validation, and Selection).

Contents

1 Introduction

Traders need a stable way to tell ordinary price fluctuation apart from statistically unusual movement: to judge, bar by bar, whether the latest move falls within the range recent history would lead one to expect or outside it. This is a classification question, not a forecasting one—it asks not where price is headed but whether its most recent move is normal—and it is the question practitioners are implicitly answering when they watch rolling ranges, recent highs and lows, and short-window volatility (Brock et al., 1992; Menkhoff and Taylor, 2007). The tools they reach for, however, are not calibrated to it: a standard Bollinger band, for instance, is built in price-level space and contains the next close on only about 83% of bars within its nominal 95% envelope (Section 4.4). This paper shows that a deliberately simple *return-space* volatility envelope is a surprisingly stable classifier of normal versus abnormal next-bar behavior, holding its containment rate across forty assets, a century of market history, and daily-to-monthly time scales. That stability—not any directional edge, which we show the band lacks—is what makes it useful: a calibrated, asset-agnostic definition of a "normal range" that a practitioner can apply without per-asset re-tuning. The one honest qualification, developed in Section 5, is that the classifier is calibrated unconditionally rather than within every regime: its coverage dips at crisis onset, so it labels the typical bar well but is not a substitute for a full conditional-density model.

That a simple past-return rule can classify next-bar behavior this stably sits in apparent tension with market efficiency—a tension worth stating precisely, because the loose version of it is wrong. The theoretical statement is that under the efficient-market hypothesis prices are a near-martingale, so simple rules based on past returns should not forecast future returns (Fama, 1970, 1991). The practitioner statement is that traders nonetheless trade against exactly the levels described above as if they carried information (Lo and MacKinlay, 2000). The apparent conflict dissolves once *first-moment* (directional) predictability is separated from conditional *second-moment* (dispersion) predictability. The efficient-market hypothesis restricts the former; it does not imply constant volatility, and it does not deny that past returns can forecast the next period's dispersion. The canonical EMH-consistent example is exactly this: GARCH-type conditional-variance predictability (Bollerslev, 1986) coexists with directional unpredictability. A return-space rolling-volatility band's containment property sits entirely in that second-moment territory, which is why it can be both real and consistent with weak-form efficiency. Section 3.1 and the near-zero directional content of the rolling mean confirm the first-moment side: the band is uninformative about direction and informative about dispersion.

This second-moment regularity is robust and model-independent: volatility clustering and forecastable conditional dispersion are among the most pervasive stylized facts in asset returns (Mandelbrot, 1963; Cont, 2001). This paper takes no position on *why* the structure exists—whether it is generated by volatility dynamics, by participant behavior, or by both—and asks only how well a simple rolling envelope tracks it.

The construction studied here isolates the simplest version of that conditional structure—a second-moment envelope. Given the prior close s_{t-1} and the rolling mean $\mu_{n,t-1}$ and standard deviation $\sigma_{n,t-1}$ of the previous $n = 60$ simple returns, we define a two-tier return-

space band, which we abbreviate BTM,[1]

$$E_t^{\pm,k} = s_{t-1}\big(1 + \mu_{n,t-1} \pm k\sigma_{n,t-1}\big), \qquad k \in \{1, 2\}.$$

Movements inside the inner ($k = 1$) band are normal; movements past the outer ($k = 2$) band are exceptional. The headline empirical fact, established below, is that this envelope contains the S&P 500 adjusted close 71.20% of the time over the 1927–2024 window, and the +3.7pp excess over the finite-window plug-in Gaussian null (Section 3.6; 67.46% at $n = 60$, versus the known-parameter 68.27%) holds across all eleven calendar decades in the sample. We show in Section 4.5 that this excess is the quantitative fingerprint of return leptokurtosis at $\nu \approx 6$.

The claim is sharper than "a band exists." This plain construction is calibrated where the standard practitioner alternative is not, by a factor of two (Section 4.4); its one departure from a textbook Gaussian is fully accounted for rather than waved away (Sections 4.3, 4.5); and the calibration is documented, not asserted, across forty assets and a century of regimes. The contributions are that evidence and the construction it evaluates: while each ingredient is standard, we are not aware of a prior construction that combines a return-space, mean-centered projection with a two-sided, two-tier chartable envelope evaluated as a containment classifier (Section 2).

The paper proceeds as follows. Section 2 situates BTM against prior work; Section 3 defines the construction and its statistical target; and Section 4 presents the core calibration evidence—the SPX headline and decade stability, cross-asset containment, the density-forecast (PIT) analysis, the matched comparison against Bollinger Bands, and the heavy-tail account of the excess. Section 5 states limitations and Section 7 concludes. Appendix A collects the estimator, range-based, source-price, out-of-sample, random-universe, regime-stress, volatility-forecasting, window-sensitivity, scale-invariance, and multi-horizon analyses.

2 Related Work

BTM sits at the intersection of three literatures.

Volatility estimation. Textbook conditional-variance estimators include Bollinger Bands (Bollinger, 2001) (SMA ± price-SD, industry-standard charting), RiskMetrics EWMA (J.P. Morgan/Reuters, 1996) (the de facto VaR baseline at $\lambda = 0.94$), and GARCH(1,1) (Bollerslev, 1986) (the canonical academic model). Range-based estimators (Parkinson, 1980; Garman and Klass, 1980; Rogers and Satchell, 1991) use the intraday high-low range and are several times more efficient than close-to-close at estimating a single bar's realized variance. We benchmark BTM against each of these in the calibration analysis below. The finding that a plain rolling variance is not beaten by EWMA or GARCH on this task sits in the volatility-evaluation tradition of Hansen and Lunde (2005), who found no model reliably improves on

[1] *Conflict of interest:* a variant of this band is deployed commercially as an invite-only TradingView indicator (the "Behavioral Transform Model"), in which the author has a financial interest. The abbreviation BTM is retained for continuity with that implementation; it is used here as a bare product label and carries no theoretical commitment (Section 3).

GARCH(1,1) out-of-sample, and of Andersen and Bollerslev (1998); the density-forecast tests we use (Diebold et al., 1998; Berkowitz, 2001; Patton, 2011) extend that programme to the full distribution.

Relation to VaR, density forecasting, and charting bands. The band's three ingredients each descend from a distinct literature. The rolling-σ estimator and the normal-conditional-on-σ assumption come from parametric value-at-risk (J.P. Morgan/Reuters, 1996; Jorion, 2006; Manganelli and Engle, 2001), which is conventionally zero-mean and one-sided; retaining the conditional mean follows the density-forecast tradition (Diebold et al., 1998; Berkowitz, 2001; Patton, 2011), whose object is the full conditional distribution; and rendering the result as a two-sided chartable envelope follows Bollinger (Bollinger, 2001), but in return rather than price space. Mapping the resulting return interval to a price band, $P_{t-1}(1+\mu \pm k\sigma)$, is the same trivial map a VaR forecast uses when expressed in price terms.

Market efficiency and conditional volatility. The first-/second-moment separation that reconciles the calibration result with weak-form efficiency (Section 1) follows the efficiency literature (Fama, 1970, 1991) and its GARCH conditional-variance precedent (Bollerslev, 1986); the underlying volatility-clustering and leptokurtosis stylized facts are model-independent and pervasive (Mandelbrot, 1963; Cont, 2001).

Technical-analysis quantification. The academic technical-analysis literature (Lo and MacKinlay, 2000; Neely et al., 2014) asks whether named chart patterns predict returns; BTM asks a different question—whether a quantitatively defined envelope is well-calibrated—and makes no return-predictability claim.

3 The Model: A Return-Space Volatility Envelope

3.1 Formal definition

Let x_t denote the close at time t, and let s_t denote the chosen *source*-price series, the close by default ($s_t = x_t$). Define the one-step simple return of the source,

$$r_t = \frac{s_t - s_{t-1}}{s_{t-1}}.$$

Given a window length n, define the rolling mean and rolling standard deviation as

$$\mu_{n,t-1} = \frac{1}{n}\sum_{i=1}^{n} r_{t-i}, \qquad \sigma_{n,t-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_{t-i} - \mu_{n,t-1})^2}.$$

We use the sample standard deviation (Bessel correction) for unbiasedness.[2] BTM thus has three estimation parameters: the rolling window n, the outer-band multiplier v_s, and the source-price series s. The expected next price and the two-tier envelope at t are

$$E_t = s_{t-1}(1 + \mu_{n,t-1}), \tag{1}$$

$$E_t^{\pm,1} = s_{t-1}\big(1 + \mu_{n,t-1} \pm \sigma_{n,t-1}\big), \tag{2}$$

$$E_t^{\pm,2} = s_{t-1}\big(1 + \mu_{n,t-1} \pm v_s \cdot \sigma_{n,t-1}\big). \tag{3}$$

We write $E_t^{\pm,k}$ for brevity; its dependence on (n, v_s, s) is carried through $\mu_{n,t-1}$, $\sigma_{n,t-1}$, and the chosen source and multiplier, and is made explicit only where we vary it (for instance the window sweep of Appendix A.8). Containment is evaluated against the realized value of the *same* series the band is built from: each source is tested against how it itself settles, not against a fixed series. The default source is the close, so the standard results build the band on close returns and test it against the realized close x_t (with the bar's high and low used for the intraday-extreme variants). The inner band is fixed at 1σ for definitional clarity; the outer band uses $v_s \geq 1$ (production default $v_s = 2$). Throughout we report results at the defaults $(n, v_s, s) = (60, 2, \text{close})$ unless noted otherwise.

The subscripts are deliberate: both moments are computed from the n returns in the window ending at r_{t-1}, the set $\{r_{t-1}, \ldots, r_{t-n}\}$, all observable at time $t-1$. The envelope $E_t^{\pm,k}$ therefore depends only on the information in that rolling window, not on the full history and nothing at or after t, so the model is strictly causal and there is no look-ahead in any result reported in this paper. This finite-window dependence is also why the plug-in (prediction-t) null of Section 3.6, rather than the known-parameter Gaussian, is the correct benchmark: the band is built from n sample moments, not the true parameters.

Local stationarity. Treating $\mu_{n,t-1}$ and $\sigma_{n,t-1}$ as estimators of local distributional parameters is valid to the extent that returns within a 60-bar window are approximately i.i.d. We verify this premise: sliding a 60-bar window across the 40-asset universe and applying the Ljung–Box test at lag 5,[3] 91.73% of ∼213,000 windows fail to reject (versus 95% under exact i.i.d.), with a small *negative* lag-1 autocorrelation on 36 of 40 assets—the sign and magnitude of the documented daily short-term reversal (Jegadeesh, 1990), not a generic mis-specification. This establishes within-window first- and second-moment local-flatness, exactly what the band requires; it does *not* establish tail stability or full-distribution stationarity, which the PIT/Berkowitz (Section 4.3), heavy-tail (Section 4.5), and regime-stress (Appendix A.6) analyses address.

On the name and the absence of a mechanism claim. We treat BTM as a bare label for the return-space envelope, inherited from the deployed product's name rather than

[2]The divisor choice is immaterial to every calibration claim below. Recomputing universe-mean containment on the 40-asset panel with the biased n divisor in place of $n-1$ shifts coverage by only +0.35pp at $k = 1$ (71.95%→71.60%) and +0.16pp at $k = 2$ (93.97%→93.81%); the gap is positive on all 40 assets and matches the Gaussian-density-weighted closed form $2k\phi(k)\big(1 - \sqrt{(n-1)/n}\big)$ to within ∼0.05pp—an order of magnitude below the cross-asset standard deviation of containment (2.31pp at $k = 1$).

[3]The lag is not load-bearing: it sits near the $m \approx \ln T \approx 4.1$ rule (Tsay, 2010) for $T = 60$, and sweeping $m \in \{1, 3, 5, 10, 20\}$ leaves the conclusion unchanged.

asserting any theoretical content. The empirical results of this paper do not depend on, and make no claim about, *why* the conditional-scale structure exists: BTM is evaluated purely as a calibrated statistical envelope in the family of practitioner rolling-statistics estimators (Bollinger, Keltner, ATR). Whether the structure is generated by volatility dynamics, by participant behavior, or by both is outside the scope of this paper; the calibration findings stand on the return data alone.

On the $\mu_{n,t-1}$ term. The drift term does two separable jobs, and we retain it in the canonical formulation. As a *centering* term it anchors the σ-wide interval on the conditional mean of the return distribution; as a *predictor* it carries weak but measurable directional content. The centering role is the structural justification for keeping $\mu_{n,t-1}$ in the band: by construction $\sigma_{n,t-1}$ is the rolling standard deviation of returns *around their mean*, so $s_{t-1}(1 + \mu_{n,t-1} \pm \sigma_{n,t-1})$ places that interval on the correct anchor. Dropping μ leaves the aggregate 1σ containment essentially unchanged, a +0.16pp universe-mean shift, because the aggregate rate integrates over both sides of the band and is blind to centering bias, but it materially distorts which side of the band gets breached on trending assets, as Appendix A.11 documents. An earlier version of this work reported only the aggregate-containment ablation and concluded that $\mu_{n,t-1}$ was negligible; that reading was correct for the aggregate calibration metric and incorrect for the directional-symmetry metric.

3.2 Why return space

BTM is estimated in *return space*: the moments $\mu_{n,t-1}$ and $\sigma_{n,t-1}$ are computed on the return series $r_t = (s_t - s_{t-1})/s_{t-1}$, and the envelope is projected back to price by multiplication by s_{t-1}. This choice—rather than computing a band directly on price levels—has four consequences. *Scale invariance*: a 2σ BTM band on SPX at \$100 in 1950 and at \$5,000 in 2024 is the same statistical statement (k standard deviations of percentage return), whereas a price-space band's dollar-denominated SD grows mechanically with the level—which is what makes the 1927–2024 pooled calibration well-posed. *Stationarity*: returns are approximately stationary over short windows and price levels are not, so a rolling SD of price on a trending series conflates volatility with drift. *Access to return-space stylized facts*: volatility clustering and leptokurtosis are properties of returns, so $\sigma_{n,t-1}$ tightens in calm and widens in turbulence, and the +3.7pp excess of the 71.20% calibration over the plug-in null (Section 3.6) is a direct fingerprint of leptokurtosis that a trend-contaminated price-space SD cannot access. *Regime stability*: the calibration holds across regimes precisely because it is a return-space property, which Section 4.4 makes quantitative.

3.3 The two-tier band design

The two-tier design partitions the band-edge population into "probe" (inner breach) and "rupture" (outer breach) sets. The outer multiplier v_s controls a precision–recall tradeoff: the $v_s = 2$ default sits at the Gaussian-quantile point $\Phi(-2) \approx 2.28\%$ and is well-calibrated for outer-band coverage. This paper uses the two tiers only to report inner- and outer-band calibration; what the two breach classes mean beyond their frequency is not studied here.

3.4 Estimation choices

We fix $n = 60$ trading days as the canonical setting on three grounds: it lies within the calibration plateau of Appendix A.8; it is short enough to react to a regime change within approximately one calendar quarter; and it is a round number near the middle of the plateau, easing reproducibility. The deployed Pine-script version defaults to $n = 100$ for slightly increased stability; qualitative results are robust to either. We use simple unweighted variance rather than exponentially weighted or maximum-likelihood-fit variants; Appendix A.1 shows this costs essentially nothing in calibration accuracy.

3.5 Containment indicators

For each bar t we record three binary indicators:

$$C_t^{\text{close}}(k) = \mathbf{1}\{E_t^{-,k} \le x_t \le E_t^{+,k}\},$$
$$C_t^{\text{high}}(k) = \mathbf{1}\{h_t \le E_t^{+,k}\},$$
$$C_t^{\text{low}}(k) = \mathbf{1}\{l_t \ge E_t^{-,k}\},$$

where h_t, l_t are the intraday high and low. Empirical containment rates are sample means of these indicators. Confidence intervals throughout are 95% block bootstraps with block length 21 trading days ($\approx$ one calendar month) and 1,000 resamples, chosen to preserve serial correlation in the return process.

3.6 Statistical target for plug-in coverage

The band uses the rolling sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$ estimated on the previous n returns, not the true parameters, so the correct null for its coverage is not the known-parameter Gaussian quantile but the *finite-window plug-in* (prediction-t) quantile. Under iid Gaussian returns, $(r_t - \bar{r}_{t-1})/(s_{t-1}\sqrt{1 + 1/n}) \sim t_{n-1}$ (Hahn and Meeker, 1991), so a $\pm k\hat{\sigma}$ band contains

$$2\,F_{t_{n-1}}\left(k\sqrt{\tfrac{n}{n+1}}\right) - 1$$

of next-bar closes under the Gaussian null. At $n = 60$ this is 67.46% for $k = 1$ and 94.80% for $k = 2$, below the known-parameter values 68.27% and 95.45% because estimating σ on a short window widens the predictive interval; the gap shrinks toward zero as n grows (Section A.8 tabulates it by n). We adopt this finite-window null as the baseline for every calibration statement in the paper and retain the known-parameter quantiles only as an intuitive reference.

4 Empirical Calibration

4.1 The 71% headline on the S&P 500

On daily SPX adjusted closes from 1927-12-30 to 2024-10-08 (24,248 bars after $n = 60$ warmup), the empirical close-containment at the 1σ band is

$$\hat{P}\big(E_t^{-,1} \le x_t \le E_t^{+,1}\big) = 71.20\%, \qquad 95\%\ \text{CI} = [70.4\%, 72.0\%].$$

Judged against the finite-window plug-in null of Section 3.6 (67.46% at $n = 60$, versus the known-parameter 68.27% reference), the SPX 1σ rate sits +3.74pp above target (it is +2.93pp above the known-parameter reference), and the CI excludes both. The upward deviation is consistent with leptokurtic structure: a fat-tailed distribution places more mass within $\pm 1\sigma$ than a Gaussian when the variance is matched to the realized second moment. BTM is therefore not a Gaussian quantile envelope at strict equality; it is a calibrated envelope whose realized coverage sits a small but systematic distance above the proper finite-window null, with the deviation magnitude consistent across decades and assets.

Table 1: SPX 1σ BTM close-containment, decade-by-decade ($n = 60$).

Decade	Bars	Containment	95% CI	Notes
1930s	2,496	73.6%	[70.9, 75.8]	Great Depression
1940s	2,500	73.7%	[71.6, 75.6]	WWII, post-war
1950s	2,511	72.5%	[70.6, 74.8]	Bretton Woods
1960s	2,489	70.4%	[67.8, 73.4]	Vietnam, Nifty Fifty
1970s	2,526	68.7%	[66.3, 70.8]	Stagflation
1980s	2,528	70.5%	[68.2, 72.5]	1987 crash included
1990s	2,528	70.5%	[68.6, 72.5]	Tech expansion
2000s	2,515	69.6%	[67.2, 71.8]	Two recessions
2010s	2,516	72.3%	[69.4, 75.1]	Low-vol bull
2020–2024	1,200	70.2%	[66.0, 74.6]	COVID, AI cycle (partial)
Full sample	24,248	71.20%	[70.4, 71.9]	97 years

The decade-by-decade view (Table 1) is the strongest single piece of evidence for the model's structural robustness. Across eleven calendar decades spanning the Great Depression, two world wars, four oil shocks, the 1987 crash, the 2000 tech bust, the 2008 financial crisis, the COVID shock, and the AI cycle, the 1σ close-containment never falls below 68.7% nor rises above 73.7%. The empirical quantile is structurally near 70%, and that fact is preserved across regime changes that altered virtually every other property of the price series (mean drift, volatility level, microstructure, sample composition).

4.2 Cross-asset containment

The natural next test is whether the same calibration holds across asset classes at a single point in time. A band that contains $\sim 70\%$ of close prices is direct evidence that price action concentrates inside the envelope rather than dispersing uniformly across its full potential range. We extend an original seven-asset baseline (Appendix B) to a forty-asset universe across five asset classes. The asset selection is not random: it was assembled ex ante from instruments with full local price coverage and is biased toward liquid US-listed assets plus the two largest cryptocurrencies. The cross-asset claim should be read as "holds across this curated universe" rather than "holds for every possible asset."

Table 2: Cross-asset 1σ BTM containment across the asset universe, $n = 60$, daily bars, split-adjusted prices, 95% block-bootstrap CIs. †SPX is the non-tradable cash index, included as a long-window cross-regime stability benchmark; the tradable wrapper SPY appears separately.

Asset	Class	Bars	Sample period	Close in 1σ band	2σ
SPX†	price index	24,248	1928–2024	71.20% [70.4, 71.9]	93.82%
SPY	broad index	7,356	1995–2024	70.83% [69.3, 72.3]	93.86%
QQQ	broad index	6,380	1999–2024	70.58% [69.0, 72.0]	93.75%
DJI	broad index	8,192	1992–2024	70.69% [69.3, 72.0]	93.63%
NASDAQ	broad index	13,474	1971–2024	70.33% [69.3, 71.4]	93.77%
EEM	broad index	5,348	2003–2024	69.67% [68.0, 71.6]	93.77%
NVDA	single name	6,412	1999–2024	73.52% [72.1, 75.0]	94.62%
AAPL	single name	3,713	2010–2024	72.96% [70.9, 75.0]	93.91%
MSFT	single name	3,713	2010–2024	72.85% [70.8, 74.9]	93.91%
AMD	single name	3,713	2010–2024	75.11% [73.0, 76.8]	94.10%
TSLA	single name	692	2021–2023	70.52% [65.6, 74.6]	94.22%
META	single name	440	2022–2023	76.36% [70.2, 81.4]	97.05%
JPM	single name	440	2022–2023	72.73% [65.9, 78.6]	95.91%
EURUSD	FX	5,352	2004–2024	71.00% [69.6, 72.4]	94.82%
GBPUSD	FX	4,939	2005–2024	70.28% [68.8, 71.7]	93.84%
USDJPY	FX	4,939	2005–2024	72.04% [70.4, 73.9]	93.64%
USDCHF	FX	4,939	2005–2024	70.99% [69.5, 72.4]	94.25%
AUDUSD	FX	4,939	2005–2024	69.87% [68.4, 71.3]	93.97%
USDCAD	FX	4,939	2007–2024	71.39% [70.0, 72.8]	93.76%
NZDUSD	FX	4,939	2005–2024	69.49% [68.0, 71.0]	94.47%
ES	equity futures	4,939	2005–2024	71.31% [69.8, 72.8]	93.60%
NQ	equity futures	4,844	2004–2023	71.08% [69.3, 72.9]	93.37%
ZN	Treasury futures	3,380	2007–2020	77.93% [75.7, 80.1]	93.58%
ZB	Treasury futures	4,823	2004–2023	70.18% [68.7, 71.5]	94.40%
DX	FX futures	4,939	2005–2024	75.30% [73.6, 76.9]	93.60%
IWM	broad index	4,939	2005–2024	69.81% [67.9, 71.6]	93.87%
TLT	Treasury ETF	4,939	2005–2024	68.03% [66.6, 69.5]	94.65%
IEF	Treasury ETF	4,939	2005–2024	69.04% [67.6, 70.5]	94.51%
LQD	credit ETF	4,939	2005–2024	70.01% [68.5, 71.6]	94.33%
HYG	credit ETF	4,345	2007–2024	72.87% [70.9, 74.9]	93.10%
UUP	FX ETF	4,380	2007–2024	70.00% [68.3, 71.5]	94.66%
VIXY	volatility ETF	3,403	2011–2024	75.02% [72.4, 77.1]	94.03%
Gold	commodity	5,988	2000–2024	72.65% [71.5, 74.0]	94.00%
Silver	commodity	5,990	2000–2024	73.51% [72.2, 74.8]	93.61%
Oil	commodity	5,997	2000–2024	70.47% [69.0, 72.0]	94.00%
GLD	commodity ETF	4,939	2005–2024	72.08% [70.7, 73.4]	93.87%
SLV	commodity ETF	4,640	2006–2024	73.06% [71.5, 74.4]	93.62%
USO	commodity ETF	4,653	2006–2024	69.63% [67.8, 71.4]	94.07%
BTC	crypto	3,614	2014–2024	76.73% [74.7, 78.5]	93.41%
ETH	crypto	2,465	2018–2024	75.42% [73.1, 77.6]	93.39%
Universe mean ($n = 40$)				71.91%	94.07%
Cross-asset SD				2.31 pp	0.70 pp

Three features constitute the finding. First, **tight clustering**: 1σ close-containment runs 68.0%–77.9% (median 71.08%, IQR $[70.28, 72.96]$), with thirty-three of forty assets at or above 70%, across single names, broad indices, ETFs, major FX, futures, commodities, and crypto; 2σ coverage averages 94.07%, near the plug-in null of 94.80%. Second, **the same simple band tracks heterogeneous instruments**: the nine-decade SPX prints inside the 1σ envelope 71.20% of the time, SPY 70.83%, NVDA (including the 2023–2024 AI run) 73.48%, and 24/7 BTC 76.73%. Third, **an asset-class signature in the over-coverage**: single names (META +8.1pp over Gaussian) and crypto (+8.5pp) show the largest 1σ over-coverage, consistent with heavier leptokurtic shape in higher-volatility assets, while broad indices show the smallest.

A formal test. Many of the forty assets are correlated (eight US equity-beta instruments, several rate and credit ETFs), so a cluster bootstrap that resamples asset groups gives an effective $N \approx 18$ rather than the nominal 40. The universe-mean 95% CI computed at that effective N is $[71.2, 72.6]$, and a one-sided cluster-bootstrap test rejects equality of the mean with the finite-window plug-in null of 67.46% decisively. Against the more conservative known-parameter reference of 68.27% the statistic is $t \approx 6.7$ at effective $N \approx 18$ ($p < 10^{-5}$); against the lower finite-window null it is larger still. The $\sim 70\%$ result is thus a formal rejection of the Gaussian baseline under either benchmark, not only a narrative one.

4.3 Density-forecast calibration via the probability integral transform

The two-point coverage result above tests only that the band lands at the correct empirical quantile at $k = 1$ and $k = 2$. To establish that the *full* conditional density implied by the band specification, $r_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\mu_{n,t-1}, \sigma^2_{n,t-1})$, is correctly specified, we apply the Diebold et al. (1998) probability integral transform (PIT) and the Berkowitz (2001) likelihood-ratio test, which convert two coverage numbers into a full density-calibration result. Define $u_t = \Phi((r_t - \mu_{n,t-1})/\sigma_{n,t-1})$; under correct conditional specification $\{u_t\}$ are i.i.d. Uniform$(0, 1)$ and $z_t^* = \Phi^{-1}(u_t)$ are i.i.d. $\mathcal{N}(0, 1)$. We report results on the 40-asset universe at $n = 60$ daily ($n_{\text{pool}} = 212{,}538$ observations).

Table 3: Universe-mean coverage $\hat{P}(|z_t| < k)$ at $n = 60$ daily across the 40-asset universe. The $\star$ entries are the $k = 1$ and $k = 2$ marginal projections that Section 4.2 reports as the two-point headline; the full curve is the density-calibration result. *Diff* is empirical minus the finite-window plug-in null (Section 3.6, $2F_{t_{n-1}}(k\sqrt{n/(n+1)}) - 1$); the known-parameter Gaussian quantile is shown alongside as a reference.

k	Gaussian ref	Plug-in null	Universe mean	Cross-asset SD	Diff vs null (pp)
0.10	0.0797	0.0787	0.0971	0.012	+1.84
0.50	0.3829	0.3782	0.4376	0.034	+5.94
1.00*	0.6827	0.6746	0.7191	0.023	+4.45
1.50	0.8664	0.8578	0.8731	0.013	+1.53
2.00*	0.9545	0.9480	0.9397	0.004	−0.83
2.50	0.9876	0.9840	0.9695	0.004	−1.45
3.00	0.9973	0.9958	0.9837	0.004	−1.21
3.50	0.9995	0.9990	0.9905	0.003	−0.85

The coverage curve passes through the two marginal points of Table 2 and reveals the full shape: positive empirical deviation in the center ($k < 1.5$) and negative deviation in the tails ($k > 2$), the model places slightly too much mass on moderate moves and too little on far-tail moves. The pooled PIT histogram (bins of width 0.05) is a clean inverted-U with shoulder dips and a modest left-tail surplus. Because the 212,538 pooled observations are not independent draws (forty correlated assets), we do not lean on the pooled KS p-value; the pooled KS distance is reported only descriptively as a maximum CDF gap of 2.8pp ($D = 0.028$). The rejection of correct density specification rests instead on the cross-asset distribution of *per-asset* KS distance (median 0.0349, IQR $[0.0219, 0.0459]$, range $[0.0121, 0.0861]$; 33/40 assets reject at $p < 0.05$, 8/40 at $p < 10^{-10}$) and on a cluster-bootstrap 95% CI on the cross-asset median of $[0.0236, 0.0457]$ (effective $N \approx 12$ over the asset-class clusters), which excludes zero. The deviation is real and consistent across assets, but modest in magnitude.

Table 4: Berkowitz LR test on the PIT residuals $z_t^* = \Phi^{-1}(u_t)$ at $n = 60$ daily. Null: $z_t^* \sim$ i.i.d. $\mathcal{N}(0, 1)$ ($\mu = 0$, $\rho = 0$, $\sigma^2 = 1$). The MLE estimates decompose the deviation into centering, serial-dependence, and scale components; LR is χ^2_3.

Asset	n	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	LR
Universe pool	212,538	−0.010	−0.002	**1.134**	**1,765.5**
SPX (1962–2024)	15,739	−0.020	+0.063	1.135	211
NASDAQ (1971–2024)	13,474	−0.022	+0.136	1.138	414
EURUSD	5,352	−0.007	−0.061	1.095	45
NVDA	6,412	−0.005	−0.018	1.114	41
BTC	3,614	+0.010	+0.008	1.202	65
ZN (10-yr Treasury futures)	3,380	+0.005	+0.018	1.240	86
VIXY (volatility ETF)	3,403	+0.033	−0.006	1.198	63

The Berkowitz auxiliary model returns a clean center and a single inflated scale, but that scalar is misleading about the *nature* of the deviation. On the universe pool the conditional mean is correctly removed ($\hat{\mu} = -0.010$, indistinguishable from zero at $n = 212{,}538$) and residual serial dependence is absent ($\hat{\rho} = -0.002$); the standardized-residual variance is $\hat{\sigma}^2 = 1.134$, nominally 13.4% above the Gaussian-iid value. It is tempting to read this as "the implied Gaussian is simply too narrow by $\sqrt{1.134} \approx 1.065$," but that reading is an artifact of the auxiliary model: an AR(1)-Gaussian Berkowitz model has only (μ, ρ, σ^2) with which to absorb *any* departure, so it necessarily reports a shape deviation as a scale one. The empirical coverage curve (Table 3) shows the deviation is shape, not scale: against the finite-window null, excess mass at the center (+5.9pp at $k = 0.5$, +4.5pp at $k = 1$) and a deficit in the shoulder (−0.8pp at $k = 2$, −1.5pp at $k = 2.5$), with the far tail returning toward the null. A uniform scale inflation would move mass *outward* at every k; the data move it *inward* at the center, the unmistakable signature of leptokurtosis, and the maximum- likelihood fit of Section 4.5 ($\hat{\nu} \approx 4.8$ on BTM-standardized residuals—lower, and so heavier-tailed, than the $\nu \approx 6$ innovation parameter, because rolling-σ standardization strips volatility clustering only imperfectly) confirms the residuals are heavy-tailed Student-t rather than a wider Gaussian. The honest conclusion is that BTM is well centered and serially clean but *not* a fully calibrated Gaussian density forecast: the PIT rejection is statistically overwhelming yet economically modest, because the two-point marginal coverage at $k = 1, 2$ sits near nominal where the leptokurtic shape and finite-window plug-in effects partially offset. The Section 4.5 simulation confirms that fat-tailed GARCH innovations reproduce both the $\sim 70\%$ marginal coverage and the $\sim 13\%$ residual-variance inflation, ruling out a fitting tautology and identifying the deviation as a property of the return series rather than a specification error in BTM.

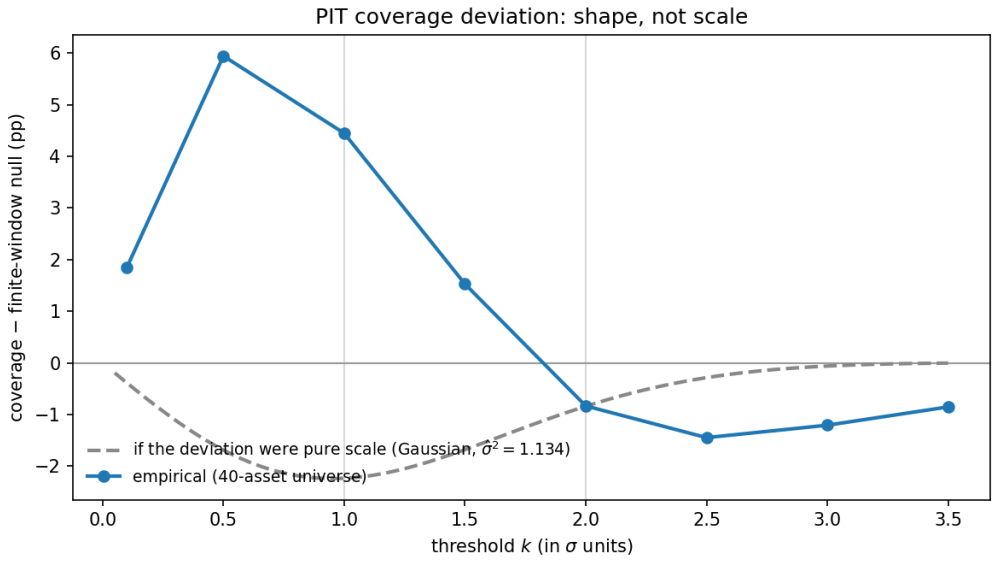

Figure 1: PIT coverage deviation across thresholds k, empirical minus the finite-window plug-in null (40-asset universe), with the curve a pure scale problem would trace shown for contrast (a Gaussian residual inflated to $\hat{\sigma}^2 = 1.134$). The empirical deviation is *positive* in the center and turns negative only in the shoulder and near tail; a scale inflation, by contrast, is negative at *every* k. The data move mass inward at the center, which a too-narrow Gaussian cannot do, so the departure is one of shape (leptokurtosis), not scale.

Coverage backtests. Standard value-at-risk backtests agree: Kupiec proportion-of-failures and Christoffersen tests (Kupiec, 1995; Christoffersen, 1998) reject strict Gaussian-iid on 92–100% of the universe, decomposing along the same two channels (leptokurtic scale excess and rolling-σ lag) the PIT analysis already identifies.

4.4 Comparison against standard Bollinger Bands

The natural benchmark is the conventional, price-space Bollinger band at the canonical setting practitioners actually run, $(n = 20, k = 2)$. Table 5 compares it against BTM on the 40-asset universe. The $k = 2$ row is the relevant one, since the two-standard-deviation band is the one charted in practice: at its nominal 95% setting the standard Bollinger band contains the next close on only 82.71% of bars, versus 93.97% for BTM's outer band; at $k = 1$ the gap is starker (42.86% vs 71.95%). BTM wins on *every* one of the 40 assets at both bands (universe-mean gaps +29.1pp at $k = 1$ and +11.3pp at $k = 2$; paired sign test $p < 10^{-9}$).

Table 5: Standard Bollinger comparison on the 40-asset universe: Bollinger at its canonical ($n = 20, k = 2$), BTM at its preferred $n = 60$, and BTM forced onto Bollinger's $n = 20$ to isolate the formulation effect. Close-containment (%) with 95% cluster-bootstrap CIs; "Wins" counts assets on which the row beats Bollinger.

Estimator	$k = 1$	$k = 2$	Wins vs. Bol
Bollinger ($n = 20, k = 2$, standard)	42.86 [42.12, 43.65]	82.71 [82.44, 82.98]	(baseline)
BTM ($n = 60$, preferred)	71.95 [71.26, 72.70]	93.97 [93.84, 94.10]	40/40
BTM ($n = 20$, matched to Bollinger)	68.94 [68.39, 69.52]	92.67 [92.53, 92.80]	40/40

Two mechanical failures compound on a trending market. The first is *mis-centering*: Bollinger anchors on the 20-bar SMA of price, which lags the current level, so the close systematically falls outside a band built around the SMA. The second is *width inflation*: a standard deviation computed on price *levels* over a trending window counts the trend itself as dispersion, so Bollinger's outer bands are padded with drift noise rather than genuine volatility. BTM's return-space $\sigma_{n,t-1}$, anchored on the prior close and measuring the dispersion of returns around their local mean, is free of both.

The gap is the return-space *formulation*, not the shorter window. Re-running BTM itself at Bollinger's own $n = 20$ still contains 68.94% ($k = 1$) and 92.67% ($k = 2$), far above Bollinger at the same window, so the window choice accounts for only $\sim$ 3pp of the $\sim$ 29pp gap and the return-space construction for the rest.

Decomposing the price-space failure: a 2×2 factorial with Keltner. The two mechanical failures can be separated by varying centering (SMA vs. EMA) and width basis (price-SD vs. Wilder ATR) at fixed $n = 60$, which also brings in Keltner, the EMA-plus-ATR practitioner band, as the fully-modified comparator (Table 6, SPX 1962–2024). Switching the center from SMA to EMA at fixed price-SD width recovers only $\sim$ 5pp at $k = 1$, so mis-centering is the smaller of the two problems. The width basis is the larger one: replacing price-SD with ATR collapses containment by 40–50pp at both thresholds, because the ATR multiplier in its own units targets a quantile far below the Gaussian 95.45%. Keltner at its practitioner default deviates 20.7pp from its own nominal target, against 2.6pp for Bollinger on the same own-target basis. No member of the price-space practitioner family is calibrated as a return-space prediction interval; the return-space formulation is the load-bearing choice, and the failure is dominated by the width basis with centering a secondary contributor.

Table 6: 2×2 centering $\times$ width factorial on SPX 1962–2024 daily, matched $n = 60$, close-containment (%). Bollinger is SMA+price-SD; Keltner is EMA+ATR; BTM (return-space) is the reference. ATR uses Wilder's true range, lagged one bar to match construction causality.

Center	Width	$k = 1$	$k = 2$
SMA	price-SD (Bollinger)	39.85	85.18
EMA	price-SD	45.08	88.04
SMA	ATR	21.16	41.52
EMA	ATR (Keltner)	22.66	44.67
BTM (return-space, reference)		70.92	93.84

4.5 Tail behavior and a falsification of the tautology objection

At the 2σ band, mean cross-asset close-containment is 94.09%, below the finite-window plug-in null of 94.80% at $n = 60$ (and below the known-parameter reference of 95.45%). The 1σ *overshoot* (+3.74pp on SPX against the finite-window null) and the 2σ *shortfall* (−0.71pp cross-asset) are one phenomenon, not two: relative to a Gaussian of the same variance, a leptokurtic distribution places more mass at the peak (within $\pm 1\sigma$), less in the shoulders (between 1σ and 2σ), and more in the tails (beyond 2σ). The empirical signs on both bands match this prediction directly; the outer-band shortfall is modest precisely because the finite-window null is itself below the known-parameter 95.45%.

Falsification: stochastic volatility alone does not produce the empirical excess. A natural skeptical reading is that the 1σ band contains $\sim 70\%$ of closes *by construction*. To test whether the excess over the finite-window null is meaningful or a tautology of estimating σ on a vol-clustering process, we simulate 1,000 paths $\times$ 5,000 bars from four data-generating processes and re-compute BTM's 1σ containment on every path using the same causal $\sigma_{n,t-1}$ convention.

Table 7: 1σ close-containment on simulated paths, $n = 60$, causal $\sigma_{n,t-1}$ convention. GARCH(1,1) parameters: $\omega = 10^{-6}$, $\alpha = 0.08$, $\beta = 0.91$ (SPX-calibrated, $\alpha + \beta = 0.99$, unconditional vol $\approx 16\%$ annualized). Empirical SPX is 71.20%. $N = 1{,}000$ paths per row.

Process	Mean	5th %–95th %	Gap to empirical
Gaussian iid	67.47%	[66.68, 68.24]	−3.73pp
GARCH(1,1), Gaussian innovations	67.17%	[66.42, 67.94]	−4.03pp
GARCH(1,1), Student-$t(\nu = 6)$ innovations	**71.28%**	[70.40, 72.13]	+0.08pp
GARCH(1,1), Student-$t(\nu = 4)$ innovations	73.82%	[72.81, 74.78]	+2.62pp

Two findings emerge. First, *vol clustering alone does not explain the empirical excess*: a GARCH process with Gaussian innovations produces essentially the same containment as iid Gaussian noise (67.17% vs. 67.47%), both at the finite-window realization of the Gaussian-quantile baseline (a few tenths below the asymptotic 68.27% because σ is estimated on a

short window). High persistence ($\alpha + \beta = 0.99$) produces realistic volatility regimes but does not move close-containment, because $\sigma_{n,t-1}$ correctly tracks the conditional SD and the innovations are still Gaussian. The empirical +4.03pp gap to GARCH-Gaussian is therefore *not* a tautology. Second, *Student-t innovations with* $\nu = 6$ *reproduce the empirical* 71.20% *within* 0.08*pp*, on a sample with ~ 0.5pp simulation SD. The empirical excess is quantitatively the leptokurtic fingerprint of a fat-tailed SPX-like return distribution (Cont, 2001): $\nu = 4$ overshoots to 73.82%, $\nu = 6$ matches the empirical rate, and the Gaussian limit ($\nu \to \infty$) lands at $\approx 68\%$. We do not hinge the result on ν being exactly 6, but a properly specified fit supports the round value. Maximum-likelihood estimates on the SPX, cross-checked between a pure-numpy profile likelihood and the `arch` library (agreeing to within 0.01 in $\hat{\nu}$), bracket it: a constant-volatility Student-t on raw returns gives $\hat{\nu} \approx 2.5$ (heavy, but conflating the innovation tail with volatility clustering); the same on the model's BTM-standardized residuals gives $\hat{\nu} \approx 4.8$; and a joint GARCH(1,1)-t MLE, which strips clustering with a fitted conditional-variance model rather than a rolling window, gives $\hat{\nu} = 5.8$ with a 95% CI of $[5.36, 6.25]$ that *includes* $\nu = 6$. The simulation's $\nu = 6$ is thus well-supported under the proper joint model; the lower $\hat{\nu} \approx 4.8$ reflects the simpler rolling-σ vol-stripping rather than a rejection of the underlying innovation tail. The 1σ containment rate is not sharply sensitive to ν once the tails are fat, so the round $\nu = 6$ path reproduces the 71.2% figure, and the qualitative conclusion that the innovations are fat-tailed rather than Gaussian is, if anything, strengthened: the GARCH-t fit beats a Gaussian-innovation GARCH by a likelihood-ratio statistic of $\sim 1{,}650$ on one degree of freedom.

5 Limitations

Four limitations bound and qualify the calibration claim, which is one of *marginal* (long-run unconditional) coverage; the band is *not* conditionally calibrated, as the regime and density-shape items below make precise. First, on universe selection: the headline universe is curated, biased toward liquid US-listed assets, with a tech-heavy single-name class. We address this directly rather than leave it open, the random-universe test of Appendix A.5 draws 79 S&P 500 and Russell 2000 names by a fixed seed and finds containment slightly *higher* and tighter than the curated set, with survivorship-bias checks confirming robustness, but that test is US-single-names-only, so the broader cross-asset generality (FX, futures, commodities, crypto) still rests on the curated Table 2 and a fully randomized cross-asset draw is not attempted here. Second, the calibration is conditional on regime: it holds in normal markets, recessions, and elevated-but-stable volatility, but degrades materially on crisis-onset days where VIX exceeds 30 (1σ containment 65.3%, 2σ 88.3%; Appendix A.6), because the rolling 60-day σ lags a volatility spike. Third, the calibration holds at native bar frequencies only under a constant estimation window; the practitioner-deployed small windows ($n = 12$ weekly, $n = 6$ monthly) carry a small-sample bias that lowers nominal coverage to $\sim 66\%$ and $\sim 62\%$ (Appendix A.10). Fourth, three single-name additions (TSLA, META, JPM) have short samples (< 700 bars) and wide CIs; their point estimates are suggestive only. The core result, $\sim 70\%$ 1σ close-containment, stable across asset classes, decades, and out-of-sample splits, with the excess explained by leptokurtosis, is robust to all of these and is the claim this paper defends.

Claim hierarchy and multiple comparisons. To make the evidence hierarchy explicit, the paper's numerical claims fall into two classes. The *confirmatory* claims are the model's central calibration propositions, and it is to this family that the family-wise $\alpha = 0.05$ Holm–Bonferroni correction (Holm, 1979) applies: cross-asset $1\sigma/2\sigma$ close-containment across the 40-asset universe (Section 4.2) and the PIT density-forecast calibration with its Berkowitz LR decomposition (Section 4.3). Both clear the correction comfortably: the cross-asset mean rejects the plug-in null at $t \approx 6.7$ ($p < 10^{-5}$, effective $N \approx 18$) and the density-forecast rejection is overwhelming, so the headline family survives at family-wise $\alpha = 0.05$ with wide margin. Every other test in Section 4 is a *robustness check*: it varies a single design or estimation choice to confirm the confirmatory results are not artifacts, and is reported descriptively without multiple-testing correction because it is designed to support an established headline rather than to discover a new one: window length n (Appendix A.8), source price (Appendix A.3), range-based estimators (Appendix A.2), the out-of-sample split (Appendix A.4), the random-universe replication (Appendix A.5), the bear-regime stress test (Appendix A.6), multi-horizon native-bar calibration (Appendix A.10), and the matched-Bollinger and VIX comparisons. The GARCH-falsification of Section 4.5 is a confirmatory test of the non-tautology claim specifically.

6 Implications and Use Cases

The result here is descriptive, but a documented, model-light calibration underwrites several practical uses, which the calibration enables but which we do not develop here. As a *risk-management* input, the outer band is a transparent value-at-risk floor whose coverage is characterized across asset classes, with the explicit caveat that it is modestly optimistic in the far tail and during crisis onset. As an *options* input, the inner-band containment rate is a calibrated prior on the next period's range, usable for strike selection or as a cross-check on the range implied by option prices. As a *monitoring* input, the breach-rate dynamics of Appendix A.6 provide a regime-onset early-warning signal. As a *systematic signal*, the normal/abnormal label is a candidate stop or filter heuristic. Whether any of these survives transaction costs and delivers economic value is a separate, testable question we leave open; the contribution of this paper is the calibration that any such use would rest on.

7 Conclusion

Three takeaways stand out. First, a deliberately simple, nearly parameter-free construction (a rolling mean and standard deviation of recent returns, anchored on the prior close) is calibrated consistently across the markets and periods we test. The inner-band coverage is similar across the asset classes in the universe (equities, currencies, commodities, bonds, and cryptocurrencies); on the S&P 500 it is stable across every decade of the sample; and it survives an out-of-sample split and an independently chosen set of names. It is this consistency, rather than the precise level of the coverage, that we emphasize. Second, where the band departs from a textbook Gaussian envelope, the departure is a *measured and explained* property of returns rather than a defect: the few-point excess in inner-band coverage is the

fingerprint of return leptokurtosis, reproduced to within a tenth of a point by fat-tailed but not by Gaussian simulations, and a standard density-forecast test confirms the band is correctly centered with no leftover predictable structure, its one departure from Gaussian being a matter of *shape*—the heavy-tailed signature of leptokurtosis—rather than width. Third, the design choices that matter are not the ones a practitioner might expect: estimating in return space rather than price space is load-bearing (the conventional price-space band is mis-calibrated by a factor of two), whereas the sophistication of the variance estimator matters much less: a plain 60-bar standard deviation is interchangeable with EWMA, and an MLE GARCH(1,1) differs only at the margin, trading a little inner-band over-coverage for better 2σ tail calibration.

These claims come with honest bounds. The calibration is a description of *normal* conditions: it holds through recessions and elevated-but-stable volatility, but the band runs too narrow in the first days of a fast crisis (VIX > 30), when realized volatility outruns the rolling estimator; and even in calm conditions, while the implied density is correctly centered, its departure from Gaussian is one of *shape*—leptokurtic, with excess mass at the center and in the far tail—rather than a uniform scale error. The model does characterize the tails (the 2σ band and the full coverage curve report outer-band frequencies directly), but it characterizes them as slightly *under*-covered relative to the Gaussian null: the 2σ envelope catches about ninety-four percent of closes in normal times against a finite-window plug-in null of 94.8% (and a known-parameter reference of 95.45%), a shortfall under one point, and materially less during a fast crisis. The result is therefore best read as a calibrated description of the day-to-day range whose tail coverage is known and modestly optimistic, rather than as a precise tail-risk model to be relied on at the most extreme quantiles or in the opening days of a crisis.

The contribution is correspondingly bounded. The paper establishes the envelope as a *marginal-coverage* classifier of next-bar price action: it labels each close normal or abnormal, and that labeling is accurate in long-run aggregate and stable across markets and regimes. It is explicitly *not* a conditionally calibrated density forecast: coverage drops at crisis onset, the formal Kupiec and Christoffersen tests reject on most assets, and the PIT residual is leptokurtic in shape rather than Gaussian. What an abnormal label implies for subsequent price behavior—whether such bars resolve in a particular way, cluster on identifiable events, or support a practical rule—is a separate question not taken up here; what this paper establishes is the calibration on which any such question would rest.

A Additional Robustness and Secondary Analyses

The analyses in this appendix support the calibration result of the main text but are not essential to it. They record the estimator and range-based comparisons, source-price and out-of-sample robustness, the random-universe and bear-regime stress tests, the VIX volatility-forecasting comparison, window sensitivity, scale invariance, multi-horizon calibration, and the role of the drift term.

A.1 Comparison against textbook estimators

It is important to be clear about what this comparison varies. BTM, RiskMetrics EWMA, and GARCH(1,1) are not three different *bands*; they are the same return-space, prior-close-anchored envelope, differing only in how each estimates the one-step conditional variance σ_t^2 that sets the band width: BTM uses an equal-weighted variance over a finite 60-bar window, EWMA an exponentially-weighted infinite-memory average ($\lambda = 0.94$), and GARCH(1,1) a maximum-likelihood-fit autoregressive recursion. The three are not even independent, RiskMetrics EWMA is exactly the restricted IGARCH(1,1) special case of GARCH, so EWMA and GARCH are nested. The comparison therefore holds the band construction fixed and isolates a single axis, the sophistication of the variance estimator, to ask whether it matters for calibration; that is the only reason to run it.

The difference is transparent from the math, not something the experiment has to discover. Each estimator's one-step conditional variance is a weighted sum of past squared returns,

$$\sigma_t^2 = c + \sum_{i \geq 1} w_i \, r_{t-i}^2,$$

and the three differ *only* in the weight kernel $\{w_i\}$ and the constant c:

- **BTM:** $w_i = 1/n$ for $i \leq n$ and 0 otherwise, $c = 0$, a rectangular kernel truncated at $n = 60$;
- **EWMA:** $w_i = (1-\lambda)\lambda^{i-1}$, $c = 0$, a geometric kernel of unbounded span whose weights sum to one;
- **GARCH(1,1):** $w_i = \alpha\beta^{i-1}$, $c = \omega/(1-\beta)$, a geometric kernel *plus* a constant long-run-variance anchor.

EWMA is exactly GARCH with $\omega = 0$ and $\alpha + \beta = 1$ (the IGARCH boundary), which is the nesting made explicit. The entire menu is therefore one knob, the shape of the kernel that averages recent squared returns (rectangular vs. geometric) and whether a long-run anchor is added, and the only open question is whether that knob moves the one-step $\pm 1\sigma$ coverage. We run this on the full 40-asset universe, holding the band construction fixed and varying only the σ estimator. Table 8 reports each estimator's mean 1σ and 2σ close-containment and its mean absolute deviation from the common reference (the known-parameter Gaussian quantiles 68.27% and 95.45%); EWMA and the per-asset maximum-likelihood GARCH(1,1) are both computed causally with no look-ahead and share BTM's rolling mean, so only the variance kernel differs.

Table 8: Calibration of the three return-space estimators on the 40-asset universe, $n = 60$, shared rolling mean, source close. Containment is the universe-mean close-containment; the deviation columns are against the known-parameter Gaussian reference (a common yardstick, 68.27% at 1σ, 95.45% at 2σ).

Estimator	Mean 1σ	Mean 2σ	$\lvert 1\sigma - 68.27 \rvert$	$\lvert 2\sigma - 95.45 \rvert$
BTM ($n = 60$ rolling SD)	71.95%	93.97%	3.70 pp	1.48 pp
RiskMetrics EWMA ($\lambda = 0.94$)	71.35%	94.06%	3.11 pp	1.39 pp
GARCH(1,1) MLE	73.53%	95.06%	5.26 pp	0.61 pp

The 40-asset comparison refines the picture and corrects a small-sample impression: among the two backward-looking smoothers the estimator does not matter, BTM and EWMA are within 0.6pp of each other at 1σ and both under-cover the 2σ tail by about 1.4pp. A per-asset maximum-likelihood GARCH(1,1) is genuinely different, however: it over-covers the inner band (73.5%, further above the Gaussian reference than BTM) but calibrates the outer band markedly better (95.1% against the 95.45% target, where BTM and EWMA sit near 94%), and it sits highest at 1σ on essentially every asset. The mechanism is that GARCH reacts to a change in volatility faster than a 60-bar window, so it widens promptly at a spike and captures more of the crisis-onset tail moves the rolling window lags, buying outer-band accuracy at the cost of inner-band over-coverage. No estimator is uniformly best: the trade is inner-band calibration (BTM and EWMA closer) against outer-band calibration (GARCH closer). Standard Bollinger Bands are deliberately excluded from this table: a price-space charting overlay is not a one-step-ahead return-space prediction interval, so ranking it against the 68.27% return-space target conflates two different objects. Evaluated against its own nominal 95.45% target a Bollinger ($n = 20, k = 2$) band deviates by 2.6pp; the 12.9pp figure that appears if the same band is forced onto the return-space target is exactly that category error, not a calibration result. The proper matched comparison, which holds (n, k) fixed and varies only return-space versus price-space, is in Section 4.4. We do not claim BTM is the best variance estimator; consistent with the literature, GARCH dominates on tail risk, and the 2σ row here quantifies that advantage. We claim only that for the inner-band normal-range envelope a plain 60-bar rolling SD is as well calibrated as any of these estimators and is the simplest, and that the large calibration gap is the return-space-versus-price-space choice, not the sophistication of the variance kernel within return space.

A complementary out-of-sample test confirms this division of labor rather than overturning it. Freezing each engine's parameters on a 70/30 train split and scoring the one-step density out of sample, a properly maximum-likelihood-fit GARCH(1,1) attains a lower negative log-likelihood than the rolling-σ engine on 17 of 18 assets under a Gaussian innovation density and on all 18 once a Student-t density is added: it is the better one-step *density* forecaster, consistent with Hansen and Lunde (2005). This does not bear on the containment claim, which concerns inner-band *coverage* rather than full-density likelihood—different objects, and the rolling-σ band remains as well calibrated at $\pm 1\sigma$ as the more elaborate engines. The same fit independently corroborates the heavy-tail reading of Section 4.5: a Student-t innovation density beats Gaussian by AIC on every asset, with maximum-likelihood ν between roughly 3.3 and 11.1 across the universe and centering near the $\nu \approx 6$ used there.

Calibration is a prerequisite, not a metric. The point of the parity in Table 8 is not to crown a best estimator; it is that, among return-space constructions, a band breach is a well-defined, quantile-anchored event no matter which variance kernel produces it. That is the property that lets one condition on a breach at all, to ask whether breaches resolve, cluster, or coincide with information, the questions companion papers take up. A price-space envelope fails those conditional analyses for a structural reason, not an empirical one: because its bound is not anchored to a return quantile, "breach" does not partition bars into a stable normal/abnormal split, so the better or worse calibration of its underlying volatility

estimate is beside the point. In that sense calibration here is a prerequisite for the rest of the program rather than a figure of merit to be optimized: the contribution is a band whose breach event is meaningful enough to build on, and the evidence that it is.

A.2 Range-based variance estimators

Would a range-based estimator, using each bar's open, high, low, and close, improve calibration? The motivating intuition is well-established: the high-low range encodes more information about realized variance, and the Parkinson, Garman–Klass, and Rogers–Satchell estimators are roughly seven to eight times more efficient than close-to-close at estimating a bar's realized variance (Parkinson, 1980; Garman and Klass, 1980; Rogers and Satchell, 1991). We tested this directly. The test uses the 40-asset universe of Table 2, here with full OHLC for every asset (the commodity ETFs GLD/SLV/USO sit alongside their Gold/Silver/Oil futures so the range-based estimators can distinguish wrapper types). This is the same 40-asset universe used in the source-price, out-of-sample, multi-horizon, and centering tests below. For each asset we re-built the band using the rolling mean of three range-based per-bar variance estimators in place of the close-to-close rolling SD, holding $n = 60$ fixed:

$$\begin{aligned}
\widehat{\sigma}^2_{\mathrm{Park},t} &= \frac{1}{4\ln 2}\,\left[\ln(H_t/L_t)\right]^2, \\
\widehat{\sigma}^2_{\mathrm{GK},t} &= \tfrac{1}{2}\left[\ln(H_t/L_t)\right]^2 - (2\ln 2 - 1)\,\left[\ln(C_t/O_t)\right]^2, \\
\widehat{\sigma}^2_{\mathrm{RS},t} &= \ln(H_t/C_t)\ln(H_t/O_t) + \ln(L_t/C_t)\ln(L_t/O_t),
\end{aligned}$$

then $\sigma_{n,t-1} = \sqrt{(1/n)\sum_{i=1}^{n}\widehat{\sigma}^2_{t-i}}$.

Table 9: Universe-mean 1σ close-containment and cross-asset SD under four variance estimators, $n = 60$, 40-asset OHLC universe. The discriminating metric here is the cross-asset SD (uniformity), not the mean deviation: the latter is quoted against the known-parameter Gaussian reference and can mislead, since the range-based rows post a *lower* mean deviation while badly failing uniformity (SD $\sim 7.5\%$ versus 2.3%) because they systematically undercontain.

Estimator	Mean containment	Cross-asset SD	Mean |dev from ref. (68.27%)|
Close-to-close (BTM)	72.12%	2.33%	3.85%
Garman–Klass	64.79%	7.67%	3.48%
Parkinson	65.04%	7.35%	3.23%
Rogers–Satchell	65.04%	7.88%	3.23%

The three range-based estimators behave nearly identically and all underperform close-to-close on cross-asset uniformity: the cross-asset SD widens roughly threefold (from 2.3% to $\sim 7.5\%$). The shortfall concentrates on assets with overnight gaps, US-equity-hours ETFs drop 10–20pp, cash equity indices 7–14pp, and the seven single-name equities a tight 6.6–9.7pp cluster, while 24-hour futures move within $\pm$1pp and G10 FX pairs actually gain. The mechanism is direct: range-based estimators measure within-session variance and miss the

overnight component that the close-to-close return contains. The conclusion is a clean negative result: range-based estimators are better at *what was the within-bar realized variance?*, but BTM is constructed against *the next bar's close-to-close return distribution*, and the close-to-close estimator is the right construction for that target. We retain the close-to-close formulation and treat this as methodologically settled.

A.3 Source-price robustness: the calibration is not specific to the close

The canonical formulation uses the close as the source series s_t. Is the $\sim 70\%$ containment an artifact of that choice? We build four separate causal bands per asset, one each on open-to-open, high-to-high, low-to-low, and close-to-close returns, and report each price point's own-band containment $P(X_t \in \text{band}_X)$ on the 40-asset OHLC universe at $n = 60$.

Table 10: Own-band containment of each daily price point under a BTM band built from its own rolling-return statistics, $n = 60$, 40-asset OHLC universe. SDs are cross-asset. The comparator is the finite-window plug-in null at $n = 60$ (67.46% at 1σ, 94.80% at 2σ; Section 3.6); the known-parameter Gaussian quantiles 68.27%/95.45% are the intuitive reference.

k	P(open|own)	P(high|own)	P(low|own)	P(close|own)	Plug-in null
1σ	71.98%	73.30%	73.09%	71.99%	67.46%
2σ	93.93%	93.68%	93.81%	93.97%	94.80%
Cross-asset SD (1σ)	2.04 pp	2.22 pp	2.20 pp	2.30 pp	

The four price points produce statistically indistinguishable calibrations: at 1σ the four pooled rates lie within 1.3pp of one another, at 2σ within 0.3pp, with matched cross-asset SDs. The calibration property is therefore not specific to closes; it is a generic property of *rolling-moment bands built on a discrete daily sample of a price process*, and the choice of close in the deployed formulation is conventional (closes are tradable, settle, and are unambiguously defined per session) rather than statistical. This sharpens the mechanism behind Appendix A.2: any same-sample-to-same-sample return series (open→open, etc.) preserves the dynamics that determine containment, whereas a range-based intra-bar estimator drops the overnight component and breaks calibration. BTM is calibrated whenever the variance estimator integrates the same sample-to-sample dynamics the band is tested against.

A.4 Out-of-sample calibration

Table 2 is computed in-sample. To rule out an in-sample-fit reading, we split each asset's series into a train half (first 50% of bars) and a test half (second 50%) on the 40-asset universe and report containment in each half, with 95% block-bootstrap CIs on the train→test difference Δ (21-bar blocks, $B = 1{,}000$).

Table 11: Out-of-sample calibration split. Pooled means across 40 assets; "CI excludes zero" counts assets whose block-bootstrap 95% CI on Δ (test − train) does not straddle zero.

Quantity	$k = 1\sigma$	$k = 2\sigma$
Pooled train containment	71.79%	94.03%
Pooled test containment	72.12%	93.90%
Pooled Δ (test − train, pp)	+0.32	−0.13
$\lvert\Delta\rvert \le 1$ pp	18/40	40/40
$\lvert\Delta\rvert \le 2$ pp	32/40	40/40
CI excludes zero	3/40	0/40

The headline is essentially unchanged across the split: pooled 1σ containment moves +0.32pp (71.79% → 72.12%) and 2σ moves −0.13pp (94.03% → 93.90%). At 2σ every one of the 40 assets is within ±1pp and no CI excludes zero; at 1σ only three assets' CIs marginally exclude zero (Gold, Silver, ES, all drifting toward *more* containment in the test half), the expected false-discovery count at $\alpha = 0.05$ across 40 tests. The $\sim 70\%/\sim 94\%$ calibration is a generalizing property, not an in-sample artifact.

A.5 Random-universe test

The 40-asset universe of Table 2 was assembled ex ante with a liquidity bias. We test the cross-asset claim out-of-universe: treating the curated universe as a training set on which the universe-mean 1σ property carries a 95% CI of $[71.2\%, 72.6\%]$, we test it on a disjoint sample of 79 US single names drawn by a fixed RNG seed from current S&P 500 (49) and Russell 2000 (30) membership, with ten-year (2014–2024) close history.

Table 12: Random-universe BTM containment versus the curated universe. Same close-based band, $n = 60$.

Quantity	Random universe ($n = 79$)	Curated ($n = 40$)
1σ mean containment	73.59%	71.91%
1σ cross-asset SD	1.71 pp	2.31 pp
2σ mean containment	94.13%	94.07%
Tickers $\ge 70\%$ at 1σ	79/79	

The random-universe mean lands at 73.59% at 1σ, *above* the curated-universe CI upper bound of 72.6%, and every one of the 79 tickers clears 70% (minimum 70.15% among large-caps, 70.60% among small/mid-caps); the 2σ result is indistinguishable from the curated universe (within 0.04pp). The selection-bias critique inverts: the diverse curated universe is the *harder* test, and an independently randomized US-equity sample the model was never parameterized against calibrates slightly *higher* and tighter. Survivorship-bias checks confirm robustness, a historical-membership sample including M&A exits gives 73.29% (−0.30pp), and a targeted universe of 26 US bankruptcies (2014–2024) gives 73.78% full-history; distress concentrates in the final 60 bars before delisting (−2.0pp at 1σ), with slow-onset failures

under-contained (rolling σ lags a gradual decline) and fast-onset failures over-contained (a single gap inflates σ, after which the band is too wide), the two modes roughly offsetting.

A.6 Regime-conditional calibration: a bear-regime stress test

The decade table (Table 1) established stability across calendar time. A sharper test partitions the SPX 1962–2024 series by explicit stress definitions: National Bureau of Economic Research (NBER) recession dates, VIX above versus below 30 (VIX from 2005), and top- versus bottom-quartile 252-bar realized vol.

Table 13: SPX 1962–2024 close-containment under regime-conditional partitions. Unrestricted baseline is 70.28% (1σ) and 93.79% (2σ).

Regime	n bars	1σ in	2σ in
All bars 1962–2024 (baseline)	15,739	70.28%	93.79%
NBER recession	1,951	67.30%	92.57%
NBER expansion	13,788	70.71%	93.96%
VIX $>$ 30 (since 2005)	435	**65.29%**	**88.28%**
VIX $\leq$ 30 (since 2005)	4,488	71.79%	93.92%
Top-quartile realized vol	3,887	72.60%	95.11%
Bottom-quartile realized vol	3,887	68.46%	92.75%

The NBER and realized-vol partitions preserve calibration within ±3pp of baseline; recession bars sit at 67.30% (essentially the Gaussian quantile) and top-quartile realized-vol bars at 72.60%/95.11%, the 2σ rate is actually *closer* to Gaussian in high-but-stable vol because the leptokurtic gap (Section 4.5) closes once the rolling estimator has caught up. The exception identifies a real limit: on the 435 crisis-onset days where VIX exceeded 30, 1σ containment falls to 65.29% (-5.0pp) and 2σ to 88.28% (-5.5pp), the same σ-lag-on-spike mechanism documented in Appendix A.7. Practitioners using the 2σ band as a $\sim 95\%$ envelope should treat that calibration as conditional on VIX ≤ 30.

For a static calibration this is the one real limit, but for a live monitor the same lag is informative rather than fatal. Because the band reacts to recent realized volatility, the breach rate *rises* as σ catches up to a developing crisis, which is itself an early-warning signal: slow-onset deteriorations give weeks of advance notice (the 2008 run-up), fast shocks about a week (2020), and only a true single-day cold shock (the 1987 crash) defeats it entirely. The property that weakens the band as a static description is what makes it useful as a monitor, and is the basis of the deployed forward-tracking use.

A.7 BTM $\sigma_{n,t-1}$ as a volatility forecaster versus VIX

A sharper test is whether $\sigma_{n,t-1}$ forecasts realized volatility better than the market's own implied-vol forecast. We compare BTM's annualized $\sigma_{n,t-1} \cdot \sqrt{252}$ and the VIX against

realized 30-day forward SPY return SD over 2020-01 through 2023-12 (n = 1,006 trading days), all in annualized vol points.

Table 14: BTM $\sigma_{n,t-1}$ vs. VIX as a forecaster of SPY realized 30-day forward volatility, 2020–2023. Annualized vol points (%).

Forecast	MAE	Bias	RMSE	Corr. with RV
BTM ($\sigma_{n,t-1}\sqrt{252}$)	7.69	+0.48	14.20	0.276
VIX	7.70	+3.74	11.89	0.454

BTM and VIX have essentially identical MAE (7.69 vs. 7.70), but differ in bias: BTM is nearly unbiased (+0.48pp), while VIX is systematically high by 3.74pp, the classical variance risk premium (VIX exceeds RV on 85.5% of days). BTM is closer to realized vol on 59.3% of days, but its wins are smaller and losses larger (higher RMSE, lower correlation), because VIX correctly anticipates the few extreme spikes (notably March 2020). The regime-stratified view (Table 15) is decisive: the two forecasts win in opposite regimes.

Table 15: MAE of BTM vs. VIX by realized-vol quartile, SPY 2020–2023.

RV quartile	n	BTM MAE	VIX MAE	Δ (BTM − VIX)
Q1 calm	252	3.98	7.58	−3.61
Q2 mid-low	251	3.42	6.43	−3.01
Q3 mid-high	251	6.35	5.84	+0.52
Q4 turbulent	252	16.99	10.96	+6.04

In the bottom two quartiles, half the sample, BTM has the lower MAE, because the fear-premium overshoot costs VIX accuracy in calm regimes; in the top quartile VIX wins as the rolling 60-day window smooths through the crisis spike while VIX reacts in days. Two cautions qualify this comparison, however. Mean absolute error is not a proxy-robust loss for volatility forecasts (Patton, 2011): under QLIKE, the standard robust loss, with HAC standard errors (Newey and West, 1987), VIX *significantly* outperforms BTM's σ at every window tested, with robust t-statistics of roughly +2 to +6. BTM remains the less biased of the two, it carries no variance risk premium, but it is not the better forecast. The honest reading, consistent with this section's theme, is that the simple rolling σ is an adequate and unbiased volatility gauge, not a superior one.

A.8 Window sensitivity

We sweep the window n from 5 to 500 bars on SPX daily data.

Table 16: 1σ close-containment as a function of n, SPX 1927–2024 daily, against the finite-window plug-in Gaussian null $2F_{t_{n-1}}(\sqrt{n/(n+1)})-1$. "Excess" is observed minus the per-n null.

n	5	10	20	30	60	100	200	500
Containment (%)	58.4	63.7	67.5	69.4	71.2	72.5	73.9	75.3
Finite-window null (%)	58.7	63.5	65.9	66.7	67.5	67.8	68.0	68.2
Excess (pp)	−0.3	+0.2	+1.6	+2.7	+3.7	+4.7	+5.9	+7.1

The raw containment is smoothly monotonic in n, but the comparison that matters is against the per-n finite-window null, and it tells a sharper story. The very short windows do *not* under-cover: at $n = 5$ and $n = 10$ the observed rate sits essentially on its (much lower) finite-window target (−0.3 and +0.2pp), because the prediction-t interval correctly widens to absorb the noisy variance estimate. What grows with n is the *excess* over the proper null, from zero at the short windows to +7.1pp at $n = 500$. Two effects compound in that excess: the leptokurtic central-mass overshoot (Section 4.5), and, at long windows, an over-smoothing of the non-stationary volatility process that pools high- and low-volatility regimes into a band that is too wide on average. We choose $n = 60$ because it sits where the excess is still primarily the leptokurtic fingerprint rather than over-smoothing, produces the canonical SPX headline, and preserves adaptivity to a regime change within about one quarter. The earlier intuition that short windows "under-cover" was an artifact of comparing them to the known-parameter 68.27% rather than to their own finite-window null.

A.9 Scale invariance across bar resolutions

We re-fit BTM on SPY at weekly, monthly, and four intraday resolutions, using a shared 3-month window for the intraday cuts.

Table 17: 1σ close-containment across bar resolutions, SPY.

Resolution	Bars	Containment (%)	Inner 95% CI	Notes
Monthly	388	67.5	[62.8, 72.2]	32 years
Weekly	1,671	69.4	[67.1, 71.7]	32 years
Daily	8,012	70.9	[69.6, 72.1]	32 years
4h intraday	1,458	70.2	[67.7, 72.6]	3 months
1h intraday	2,730	69.5	[67.6, 71.4]	3 months
30m intraday	5,460	68.4	[66.9, 69.9]	3 months
15m intraday	10,920	65.1	[63.9, 66.4]	3 months

The calibration is approximately scale-invariant: across two orders of magnitude in sampling frequency, 1σ inner-band coverage stays within roughly 65–72%. The shortest intraday cut (15m) deviates most, consistent with higher microstructure noise at sub-hour frequencies. Crucially, the model parameters were derived on daily data and not re-tuned for the intraday cuts; the calibration carries across resolutions without re-fitting, a strong consistency check for the structural soundness of the envelope.

A.10 Multi-horizon calibration via native bar frequencies

The scale-invariance result above held on SPY. We extend it across the 40-asset universe by applying BTM *natively* at weekly (Friday-to-Friday) and monthly (month-end) close returns, under two window configurations: constant $n = 60$ at every frequency, and the practitioner-deployed calendar-time match ($n = 60$ daily, $n = 12$ weekly, $n = 6$ monthly).

Table 18: Native-bar BTM close-containment across daily, weekly, and monthly frequencies on the 40-asset OHLC universe, under constant-n and practitioner-n configurations. The plug-in null ($1\sigma/2\sigma$) is the finite-window target at each row's n (Section 3.6); it falls with n.

Configuration	Bar frequency	1σ mean	2σ mean	1σ SD	2σ SD	Plug-in null
Constant $n = 60$	Daily	71.95%	93.97%	2.31 pp	0.43 pp	67.46/94.80
	Weekly	71.63%	93.93%	2.06 pp	0.74 pp	67.46/94.80
	Monthly	72.33%	94.36%	5.24 pp	2.18 pp	67.46/94.80
Practitioner n	Daily ($n = 60$)	71.95%	93.97%	2.31 pp	0.43 pp	67.46/94.80
	Weekly ($n = 12$)	66.11%	91.37%	1.67 pp	0.66 pp	63.47/91.11
	Monthly ($n = 6$)	61.64%	87.48%	2.64 pp	1.67 pp	58.70/85.81

Holding $n = 60$ at every frequency, 1σ containment is 71.95% daily, 71.63% weekly, 72.33% monthly, all within 0.7pp and about 4pp above the $n = 60$ plug-in null, with the leptokurtic fingerprint replicating at every frequency. The calibration is a property of the construction at a fixed estimation window, not of the bar frequency. The practitioner configuration tells a coherent story once read against the right null: matching calendar time ($n = 12$ weekly, $n = 6$ monthly) lowers raw 1σ containment to 66.11% and 61.64%, but that is the finite-window null falling with n, not a miscalibration. Against the proper plug-in nulls (63.47% at $n = 12$, 58.70% at $n = 6$; Section 3.6) the band still sits roughly 3pp above, the same leptokurtic excess as at $n = 60$, and the 2σ rates (91.37%, 87.48%) are likewise at or above their nulls (91.11%, 85.81%). The earlier reading that a small-n band "falls below the Gaussian quantile" was an artifact of comparing it to the known-parameter 68.27% rather than to its own finite-window null. For deployment the practitioner caveat is then simply that a monthly $n = 6$ band is a soft $\sim 62\%$ envelope in raw terms because its proper target is low, not because it is mis-calibrated. (A naive $\sigma_{n,t-1}\sqrt{N}$ projection from daily bars is a different and worse spec, 63% at $N = 20$, 53% at $N = 60$, because the i.i.d. scaling breaks across the regime variation a longer forward horizon spans; giving the model the actual return series at each frequency is the cleaner test.)

A.11 The role of the drift term: centering symmetry

Section 3.1 retained $\mu_{n,t-1}$ on the grounds that it does structural centering work invisible to the aggregate containment rate. We make that claim quantitative here. Aggregate 1σ containment integrates over both sides of the band and is insensitive to centering bias: dropping μ entirely shifts the universe-mean rate by only +0.16pp, and every long-window asset ($\geq$ 5,000 bars) moves by under 1pp, so the signs and significance levels of Table 2 survive intact under the simpler σ-only formula. A sharper diagnostic asks whether closes

that fall *outside* the band are distributed symmetrically above and below it. Under a properly centered band the upper- and lower-band breach rates should be approximately equal (each near $\frac{1}{2}(1 - 0.6827) \approx 15.9\%$ at 1σ); under a $\mu = 0$ ablation on a drifting asset the band sits off-center and the high side is breached systematically more often.

Table 19: Centering symmetry test. Mean absolute breach asymmetry $|P_{\text{upper}} - P_{\text{lower}}|$ at 1σ across the 40-asset OHLC universe under proper $\mu_{n,t-1}$ centering versus $\mu = 0$. Aggregate 1σ containment is essentially identical between the two specifications (within 0.5pp on every asset); the asymmetry pattern diverges sharply on trending assets.

Group	n_{assets}	mean $\lvert\text{asym}_{\mu}\rvert$	mean $\lvert\text{asym}_{\mu=0}\rvert$
Trending ($\lvert\mu_{\text{ann}}\rvert > 5\%$)	25	0.63 pp	1.85 pp
Sideways ($\lvert\mu_{\text{ann}}\rvert \leq 5\%$)	15	0.53 pp	0.60 pp

We measure asym $= P_{\text{upper-breach}} - P_{\text{lower-breach}}$ on the 40-asset OHLC universe under both centerings (Table 19), splitting the universe on drift into trending ($|\mu_{\text{ann}}| > 5\%$: most equity classes, GLD, BTC, ETH) and sideways ($|\mu_{\text{ann}}| \leq 5\%$: G10 FX, ZN, ZB, DX, TLT, IEF, LQD, HYG, UUP) groups. On sideways assets, ablating μ barely moves breach symmetry (0.53 $\rightarrow$ 0.60pp): there is no drift to mis-center. On trending assets, ablating μ roughly *triples* the asymmetry (0.63 $\rightarrow$ 1.85pp), and the sign tracks the asset's drift. The single names are the cleanest examples: AAPL goes from +0.11pp under $\mu_{n,t-1}$ to +3.18pp under $\mu = 0$, META from +0.96 to +3.15pp, BTC from +0.97 to +3.07pp; the downward-trending VIXY flips sign correctly (+1.23 to −1.35pp). The cost of dropping μ is thus not visible in the classifier rate but in the directional symmetry of misclassifications, which is why the canonical model retains the centering term.

The term is not merely cosmetic on the predictive side either. Across the same 40-asset universe the sign of $\mu_{n,t-1}$ agrees with the next bar's return sign on 51.1% of bars (50.4% after detrending the long-run drift, above the 50% null on 24 of 40 assets), and a naive one-bar long-short carries a small positive mean Sharpe (+0.12 raw, falling to ≈ 0 once the signal is detrended, most of the raw edge is the equity-premium drift itself). Replacing $\mu_{n,t-1}$ with the raw last-return sign $\text{sign}(r_t)$ inverts the result (hit rate 49.0%, mean Sharpe −0.14), so the 60-bar smoothing cancels the short-horizon reversal pattern (Jegadeesh, 1990) rather than merely echoing the last return. The directional content is weak, too small to trade standalone after costs, and largely drift once isolated, but the sign agreement and the $\text{sign}(r_t)$ inversion confirm $\mu_{n,t-1}$ is not trivially r_t in disguise.

B Baseline seven-asset containment table

Table 20 is the original seven-asset cross-asset containment table that motivated the expanded universe of Section 4.2. It additionally reports intraday-high and intraday-low containment (the 1σ band's coverage of the bar's intraday extremes), with split-adjusted intraday range (factor = AdjClose/Close applied to each bar's high and low).

Table 20: Original 1σ BTM containment baseline, $n = 60$, daily bars, 95% block-bootstrap CIs.

Asset	Bars	Close in band	95% CI	High $\leq$ upper	Low $\geq$ lower
SPX (S&P 500)	24,248	71.20%	[70.4, 71.9]	72.8%	71.0%
SPY	7,356	70.83%	[69.3, 72.3]	75.7%	71.6%
QQQ	6,380	70.58%	[69.0, 72.0]	75.9%	71.5%
NVDA	6,412	73.52%	[72.1, 75.0]	76.4%	73.2%
BTC–USD	3,614	76.73%	[74.7, 78.5]	80.8%	78.5%
Gold (futures)	5,988	72.65%	[71.5, 74.0]	79.0%	78.9%
EURUSD	5,352	71.00%	[69.6, 72.4]	67.4%	67.4%
Seven-asset avg.	–	72.36%	–	75.4%	73.2%

Close-containment averages 72.36% across the baseline with cross-asset SD $\approx$ 2.4pp; BTC (+8.5pp over the Gaussian quantile) is highest, EURUSD (+2.7pp) lowest, and no asset deviates by more than 9pp. The intraday-high containment sits above close-containment on six of seven assets, with the intraday-low containment close behind, reflecting mild band asymmetry around the prior close; EURUSD is the exception on both, consistent with its intra-bar reversion signature.

C Reproducibility Index

Table 21: Reproducibility index for the calibration results in this paper.

Result	Replication script
SPX headline 71.20% containment	`exp_containment.py`
Decade-by-decade SPX (Table 1)	`exp_containment.py`
Baseline seven-asset (Table 20)	`exp_containment.py`
Cross-asset containment (Table 2)	`exp_universal_containment.py`
Local-stationarity Ljung–Box test	`exp_local_stationarity.py`
PIT density-forecast calibration (Tables 3, 4)	`exp_pit_calibration.py`
BTM vs. EWMA vs. MLE GARCH, 40-asset (Table 8)	`repro_estimator_comparison_40asset.py`
Range-based estimators (Table 9)	`exp_rangebased.py`
Source-price robustness (Table 10)	`exp_source_robustness.py`
Out-of-sample split (Table 11)	`exp_oos_calibration.py`
Random-universe test (Table 12)	`exp_random_universe.py`
Regime-conditional stress (Table 13)	`exp_regime_stress.py`
Multi-horizon calibration (Table 18)	`exp_multihorizon.py`
BTM vs. standard Bollinger (Table 5)	`exp_btm_vs_bollinger_ownbest.py`
BTM vs. VIX (Tables 14, 15)	`exp_btm_vs_vix.py`
Window sensitivity (Table 16)	`exp_n_sensitivity.py`
GARCH falsification (Table 7)	`exp_garch_falsification.py`
Scale invariance (Table 17)	`exp_intraday.py`, `exp_timeframe.py`
Centering symmetry (Table 19)	`exp_mu_centering_symmetry.py`
$\mu_{n,t-1}$ directional content	`exp_mu_directional.py`, `exp_mu_directional_detrended.py`

Data and code availability. The replication scripts listed above and the result files they produce are available from the author on request. Raw price data were obtained from public sources: daily OHLC for the broad equity indices and for the BTC, gold, and EURUSD series from standard historical-price feeds, and the single-name equities, intraday SPY, and VIX series from Financial Modeling Prep (FMP)[4] (VIX from 2005-03-17).

References

Andersen, T. G. and T. Bollerslev (1998). Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *International Economic Review* 39(4), 885–905.

Berkowitz, J. (2001). Testing Density Forecasts, with Applications to Risk Management. *Journal of Business & Economic Statistics* 19(4), 465–474.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31(3), 307–327.

Bollinger, J. (2001). *Bollinger on Bollinger Bands*. McGraw-Hill.

Brock, W., J. Lakonishok, and B. LeBaron (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *Journal of Finance* 47(5), 1731–1764.

Christoffersen, P. F. (1998). Evaluating Interval Forecasts. *International Economic Review* 39(4), 841–862.

Cont, R. (2001). Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance* 1(2), 223–236.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* 39(4), 863–883.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25(2), 383–417.

Fama, E. F. (1991). Efficient Capital Markets: II. *Journal of Finance* 46(5), 1575–1617.

Garman, M. B. and M. J. Klass (1980). On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business* 53(1), 67–78.

Hahn, G. J. and W. Q. Meeker (1991). *Statistical Intervals: A Guide for Practitioners*. Wiley.

Hansen, P. R. and A. Lunde (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20(7), 873–889.

[4]Financial Modeling Prep, `https://financialmodelingprep.com`; price and index history accessed via the FMP API, 2024–2025.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.

Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *Journal of Finance* 45(3), 881–898.

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk* (3rd ed.). McGraw-Hill.

Manganelli, S. and R. F. Engle (2001). Value at Risk Models in Finance. *European Central Bank Working Paper Series* No. 75.

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives* 3(2), 73–84.

Lo, A. W. and A. C. MacKinlay (2000). *A Non-Random Walk Down Wall Street*. Princeton University Press.

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *Journal of Business* 36(4), 394–419.

Menkhoff, L. and M. P. Taylor (2007). The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. *Journal of Economic Literature* 45(4), 936–972.

Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55(3), 703–708.

Patton, A. J. (2011). Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics* 160(1), 246–256.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science* 60(7), 1772–1791.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business* 53(1), 61–65.

J.P. Morgan and Reuters (1996). *RiskMetrics: Technical Document* (4th ed.). New York.

Rogers, L. C. G. and S. E. Satchell (1991). Estimating Variance from High, Low and Closing Prices. *Annals of Applied Probability* 1(4), 504–512.

Tsay, R. S. (2010). *Analysis of Financial Time Series* (3rd ed.). Wiley.