Volatility-Adaptive Conformal Prediction for Factor Return Uncertainty

Chorok Lee*

December 24, 2025

Abstract

Standard conformal prediction achieves nominal coverage overall but under-covers during high-volatility periods, when coverage matters most. We provide theoretical foundations explaining this phenomenon: under multiplicative heteroskedasticity, standard conformal prediction's conditional coverage decreases monotonically with volatility. We prove that volatility-scaled conformal prediction achieves exact conditional coverage regardless of volatility level, with explicit robustness bounds for volatility estimation error. Using 738 months of Fama-French factor data (1963–2024), we document that standard conformal prediction achieves only 74% coverage during high-volatility periods versus the 90% target. A simple fix—scaling prediction intervals by realized volatility—restores coverage to the target level. While normalized conformal prediction exists in the machine learning literature, we provide the first comprehensive analysis for financial applications: we show it outperforms GARCH(1,1) prediction intervals (83–86%), conformalized quantile regression (74%), and historical simulation (76%) despite requiring no distributional assumptions. Our results provide both rigorous guarantees and practical guidance for uncertainty quantification in factor investing.

*Korea Advanced Institute of Science and Technology (KAIST). Email: choroklee@kaist.ac.kr

Keywords: Conformal prediction, uncertainty quantification, factor investing, volatility, heteroskedasticity, GARCH

JEL Classification: C53, G11, G17

1 Introduction

Uncertainty quantification is critical for financial decision-making. Portfolio managers need reliable prediction intervals to set position sizes, risk managers need valid coverage guarantees for Value-at-Risk, and investors need honest assessments of forecast uncertainty. Conformal prediction has emerged as a powerful framework for distribution-free uncertainty quantification (Vovk et al., 2005; Lei et al., 2018), providing finite-sample coverage guarantees under minimal assumptions.

However, financial returns are heteroskedastic. Volatility clusters, with extended periods of high volatility followed by extended periods of low volatility. This creates a fundamental challenge for conformal prediction: when calibration data comes from a low-volatility period and test data from a high-volatility period, the exchangeability assumption fails and coverage breaks down.

In this paper, we document this phenomenon and propose a simple solution.

Scope and novelty. We emphasize upfront that volatility-scaled conformal prediction is not methodologically novel—Papadopoulos et al. (2008) introduced normalized nonconformity measures in 2008, and Lei et al. (2018) discuss variance-weighted approaches. Our contribution is *applied*: we provide the first systematic analysis of conformal prediction's coverage properties for factor returns, quantify the magnitude of coverage breakdown under heteroskedasticity, and demonstrate that simple volatility scaling outperforms more sophisticated alternatives in this domain. We view this as a "bridge" paper that brings established machine learning methodology to the attention of financial econometricians with domain-specific analysis.

Our specific contributions are:

1. **Theoretical analysis for finance:** We derive explicit coverage bounds under the multiplicative heteroskedasticity model common in financial returns. Theorem 1 quantifies how coverage degrades with volatility; Theorem 3 shows volatility scaling restores uniform conditional coverage; Theorem 4 provides robustness bounds for estimation error. These results adapt existing theory to the financial setting with explicit, interpretable bounds.

2. **Empirical documentation:** Using 738 months of Fama-French factor data (1963–2024), we show that standard conformal prediction achieves only 74% coverage during high-volatility periods versus the 90% target—a 16 percentage point shortfall when uncertainty matters most. We also document that formal i.i.d. assumptions hold for only 2 of 6 factors, yet volatility scaling remains effective empirically.

3. **Baseline comparisons:** We compare against GARCH(1,1) with Gaussian and Student-t innovations, conformalized quantile regression (CQR), historical simulation, and EWMA-scaled variants. Volatility-scaled CP achieves the best high-volatility coverage (91.6%) despite requiring no distributional assumptions or parameter estimation. We provide detailed analysis of why CQR—designed for heteroskedastic data—fails to improve over standard CP in this setting.

4. **Out-of-sample validation:** Rolling window analysis confirms that results are not artifacts of the train/test split, with 90.2% high-volatility coverage in true out-of-sample evaluation across six decades.

2 Related Work

2.1 Conformal Prediction

Conformal prediction was introduced by Vovk et al. (2005) as a framework for constructing prediction sets with finite-sample validity. Lei et al. (2018) developed split conformal prediction for computational efficiency. Recent extensions address distribution shift (Tibshirani et al., 2019; Gibbs and Candès, 2021; Zaffran et al., 2022) and quantile regression (Romano et al., 2019).

2.2 Normalized and Locally-Weighted Conformal Prediction

The idea of normalizing nonconformity scores to handle heteroskedasticity has precedent in the conformal prediction literature. Papadopoulos et al. (2008) introduced normalized nonconformity measures that divide residuals by local difficulty estimates. Lei et al. (2018) discuss locally-weighted conformal prediction where scores are scaled by estimated variance. Romano et al. (2019) propose conformalized quantile regression (CQR), which directly estimates conditional quantiles to achieve approximate conditional coverage.

Our contribution relative to this literature is threefold: (1) we provide theoretical analysis specific to the multiplicative heteroskedasticity model common in financial returns, (2) we empirically document the magnitude of under-coverage in factor returns and demonstrate that simple volatility scaling outperforms both standard CP and parametric GARCH, and (3) we validate that this approach achieves robust out-of-sample performance across over six decades of data.

2.3 Conformal Prediction in Finance

Applications of conformal prediction in finance are growing. Fantazzini (2024) applies adaptive conformal inference to cryptocurrency VaR. Related work includes conformal prediction for portfolio optimization (Johnstone and Lindley, 2021) and credit risk (Bellotti, 2021). Our

work differs by documenting the specific interaction between volatility regimes and coverage, providing theoretical analysis of this phenomenon, and showing that simple volatility scaling outperforms more sophisticated alternatives including CQR.

2.4 Heteroskedasticity in Factor Returns

Factor returns exhibit well-documented volatility clustering (Engle, 1982). GARCH models (Bollerslev, 1986) are the standard approach for capturing this heteroskedasticity. Our contribution is showing how to incorporate volatility signals into conformal prediction without requiring distributional assumptions.

3 The Problem: Under-Coverage During High Volatility

3.1 Standard Conformal Prediction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be exchangeable pairs and $\hat{f}$ a point predictor. Split conformal prediction:

1. Computes nonconformity scores on calibration data: $s_i = |Y_i - \hat{f}(X_i)|$
2. Finds the $(1-\alpha)$-quantile: $\hat{q} = \text{Quantile}(\{s_i\}, 1-\alpha)$
3. Constructs intervals: $\mathcal{C}(X) = [\hat{f}(X) - \hat{q}, \hat{f}(X) + \hat{q}]$

Under exchangeability: $\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha$.

3.2 The Heteroskedasticity Problem

Factor returns are heteroskedastic. Let σ_t denote time-varying volatility. When $\sigma_{\text{test}} > \sigma_{\text{cal}}$ (test period more volatile than calibration), the fixed quantile $\hat{q}$ is too small, causing under-coverage.

Define high-volatility periods as $\mathcal{T}_H = \{t : \sigma_t > \text{median}(\sigma)\}$. We find:

$$\mathbb{P}(Y_t \in \mathcal{C}(X_t) \mid t \in \mathcal{T}_H) \ll 1 - \alpha \tag{1}$$

This is not a failure of conformal prediction—it's a violation of the exchangeability assumption that conformal prediction relies on.

3.3 This is a Regime-Change Problem

A key insight from our analysis: standard conformal prediction works well *within* stable volatility regimes. The under-coverage arises specifically when calibration and test periods span different regimes.

We demonstrate this by analyzing subperiods. Within 1963–1993 and within 1994–2024, standard conformal prediction achieves near-nominal coverage. The severe under-coverage (dropping to 65–82% from 90%) appears only in the full-sample analysis where calibration and test periods span both regimes.

This has practical implications: the problem is not that conformal prediction is fundamentally broken for financial data, but that long calibration windows spanning multiple regimes can hurt rather than help.

4 Methodology: Volatility-Scaled Conformal Prediction

We propose a simple fix to the heteroskedasticity problem: scale the conformal interval by the volatility ratio.

4.1 Algorithm

Algorithm 1 Volatility-Scaled Conformal Prediction

Require: Calibration data $\{(Y_i, \sigma_i)\}$, test point volatility σ_{test}, level α

1: Compute nonconformity scores: $s_i = |Y_i - \hat{f}(X_i)|/\sigma_i$

2: Compute quantile: $\hat{q} = \text{Quantile}(\{s_i\}, 1 - \alpha)$

3: **return** Interval $[\hat{f}(X) - \hat{q} \cdot \sigma_{\text{test}}, \hat{f}(X) + \hat{q} \cdot \sigma_{\text{test}}]$

The key insight is to use *standardized* nonconformity scores $s_i = |Y_i - \hat{f}(X_i)|/\sigma_i$ rather than raw residuals. This "undoes" the heteroskedasticity, restoring exchangeability (see Section 5).

This approach is:

- **Simple:** One line of code beyond standard conformal prediction
- **Interpretable:** Intervals scale proportionally with volatility
- **Theoretically grounded:** Achieves exact conditional coverage under multiplicative heteroskedasticity (Theorem 3)
- **Effective:** Achieves 90% high-volatility coverage (versus 74% without)

4.2 Volatility Signal

We use trailing 12-month realized volatility as our signal:

$$\sigma_t = \text{std}(r_{t-11}, \ldots, r_t) \tag{2}$$

We normalize by the expanding median for stationarity. Other volatility measures (GARCH, implied volatility) could substitute.

5 Theoretical Analysis

We now provide theoretical foundations for volatility-adaptive conformal prediction. We establish three main results: (1) a quantification of standard CP's coverage failure under heteroskedasticity, (2) an exact conditional coverage guarantee for volatility-scaled CP, and (3) robustness bounds under volatility estimation error.

5.1 Model and Assumptions

Assumption 1 (Multiplicative Heteroskedasticity)**.** *Returns follow a location-scale model:*

$$Y_t = \mu + \sigma_t \epsilon_t \tag{3}$$

where $\{\epsilon_t\}$ *are i.i.d. with continuous symmetric distribution,* $\mathbb{E}[\epsilon_t] = 0$*, and* $Var(\epsilon_t) = 1$*.*

This assumption nests GARCH, stochastic volatility, and regime-switching models as special cases, provided the standardized residuals are i.i.d. Let $F_{|\epsilon|}$ denote the CDF of $|\epsilon|$, and $q_\alpha = F_{|\epsilon|}^{-1}(1-\alpha)$ be the $(1-\alpha)$-quantile.

5.2 Standard CP Under-Covers Under Heteroskedasticity

Theorem 1 (Under-Coverage of Standard CP)**.** *Under Assumption 1 with known mean* $\hat{\mu} = \mu$*, the conditional coverage of standard CP given volatility* σ_{n+1} *is:*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{std} \mid \sigma_{n+1}) = F_{|\epsilon|}\left(\frac{\hat{q}}{\sigma_{n+1}}\right) \tag{4}$$

This is strictly less than $1-\alpha$ *whenever* $\sigma_{n+1} > \hat{q}/q_\alpha$*.*

Proof. Under the model $Y_{n+1} = \mu + \sigma_{n+1}\epsilon_{n+1}$:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{\text{std}} \mid \sigma_{n+1}) = \mathbb{P}(|Y_{n+1} - \mu| \leq \hat{q}) \tag{5}$$

$$= \mathbb{P}(\sigma_{n+1}|\epsilon_{n+1}| \leq \hat{q}) = F_{|\epsilon|}\left(\frac{\hat{q}}{\sigma_{n+1}}\right) \tag{6}$$

Since $F_{|\epsilon|}$ is strictly increasing, coverage decreases monotonically in σ_{n+1}. □

Corollary 2 (Coverage Gap)**.** *For Gaussian innovations, if calibration occurs at volatility σ_{cal} and testing at $\sigma_{test} = \rho \cdot \sigma_{cal}$ with $\rho > 1$, the conditional coverage at test time is:*

$$\textit{Coverage} = 2\Phi\left(\frac{z_{1-\alpha/2}}{\rho}\right) - 1 \tag{7}$$

where Φ is the standard normal CDF and $z_{1-\alpha/2} \approx 1.645$ for $\alpha = 0.1$.

Example 1. *For $\rho = 2$ (test volatility twice calibration volatility) with $\alpha = 0.1$:*

- *Target coverage: 90%*
- *Actual coverage:* $2\Phi(1.645/2) - 1 = 2\Phi(0.82) - 1 \approx 59\%$
- *Coverage gap: $\approx$ 31 percentage points*

In practice, the empirical gap (16pp) is smaller because calibration includes mixed volatility periods rather than purely low-volatility data.

5.3 Volatility-Scaled CP Achieves Uniform Coverage

Theorem 3 (Uniform Conditional Coverage)**.** *Under Assumption 1 with known mean:*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{vs} \mid \sigma_{n+1}) = 1 - \alpha + O(1/n) \tag{8}$$

for any $\sigma_{n+1} > 0$. The conditional coverage is independent of volatility.

Proof. Define standardized residuals $\tilde{\epsilon}_t = (Y_t - \mu)/\sigma_t = \epsilon_t$. The nonconformity scores for volatility-scaled CP are $s_i = |Y_i - \mu|/\sigma_i = |\epsilon_i|$.

Since $\{|\epsilon_i|\}_{i=1}^{n+1}$ are i.i.d. (hence exchangeable), standard conformal theory applies:

$$\mathbb{P}(|\epsilon_{n+1}| \leq \hat{q}_{\text{vs}}) = 1 - \alpha + O(1/n) \tag{9}$$

This probability does not depend on σ_{n+1}. □

Remark 1 (Key Insight)**.** *Volatility scaling "undoes" heteroskedasticity, recovering exchangeability of standardized residuals. Standard CP fails by comparing raw residuals across different volatility regimes.*

5.4 Robustness to Volatility Estimation Error

In practice, σ_t must be estimated.

Assumption 2 (Bounded Relative Error)**.** *The volatility estimates satisfy* $|\hat{\sigma}_t - \sigma_t|/\sigma_t \leq \delta$ *for some* $\delta \in [0, 1)$.

Theorem 4 (Robustness)**.** *Under Assumptions 1 and 2:*

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_{vs}) \geq (1 - \alpha) - 2\delta \cdot f_{|\epsilon|}(q_\alpha) \cdot q_\alpha + O(\delta^2) \tag{10}$$

where $f_{|\epsilon|}$ *is the density of* $|\epsilon|$.

Corollary 5 (Coverage Loss Bound)**.** *For Gaussian innovations with* $\alpha = 0.1$*, coverage loss from estimation error* δ *is approximately* 0.68δ *to first order. Ten percent relative error* ($\delta = 0.1$) *costs approximately 7 percentage points of coverage in the worst case.*

The proof follows from analyzing how estimation error perturbs the quantile of standardized scores. See Appendix E for details.

6 Empirical Analysis

6.1 Data

We use monthly factor returns from the Kenneth French Data Library, July 1963 to December 2024 (738 months). We analyze the five Fama-French factors (Mkt-RF, SMB, HML, RMW, CMA) plus Momentum (Mom), for a total of six factors.

6.2 Experimental Setup

- **Calibration:** First 50% of observations
- **Test:** Remaining 50%
- **Point predictor:** Calibration sample mean (naive forecast)
- **Target coverage:** 90% ($\alpha = 0.1$)
- **High/low volatility:** Above/below median of test-period volatility signal

6.3 Main Results

Table 1 presents coverage by volatility regime.

Table 1: Coverage by Method and Volatility Regime (90% Target, Split Sample)

Factor	High-Volatility Coverage		Difference
	Standard CP	Vol-Scaled CP	(VS − Std)
Mkt-RF	78.8% (3.0)	89.1% (2.3)	+10.3pp***
SMB	82.1% (2.8)	93.5% (1.8)	+11.4pp***
HML	72.3% (3.3)	89.7% (2.2)	+17.4pp***
RMW	65.2% (3.5)	85.9% (2.6)	+20.7pp***
CMA	74.5% (3.2)	91.8% (2.0)	+17.3pp***
Mom	72.3% (3.3)	91.3% (2.1)	+19.0pp***
Average	**74.2%**	**90.2%**	+16.0pp***
Gap from 90%	**−15.8pp**	**+0.2pp**	—

Note: Split sample evaluation with calibration on 1963–1993 and test on 1994–2024. Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ for two-proportion z-test of coverage difference. Section 9 presents rolling window out-of-sample results.

Key findings:

1. **Standard CP under-covers severely.** During high-volatility periods, coverage averages only 74.2%—nearly 16 percentage points below target. RMW is worst at 65.2%.

2. **Volatility scaling fixes the problem.** Simple scaling achieves 90.2% average high-volatility coverage, essentially matching the target. The improvement is highly significant ($z > 4$, $p < 0.001$).

Figure 1 visualizes these results.

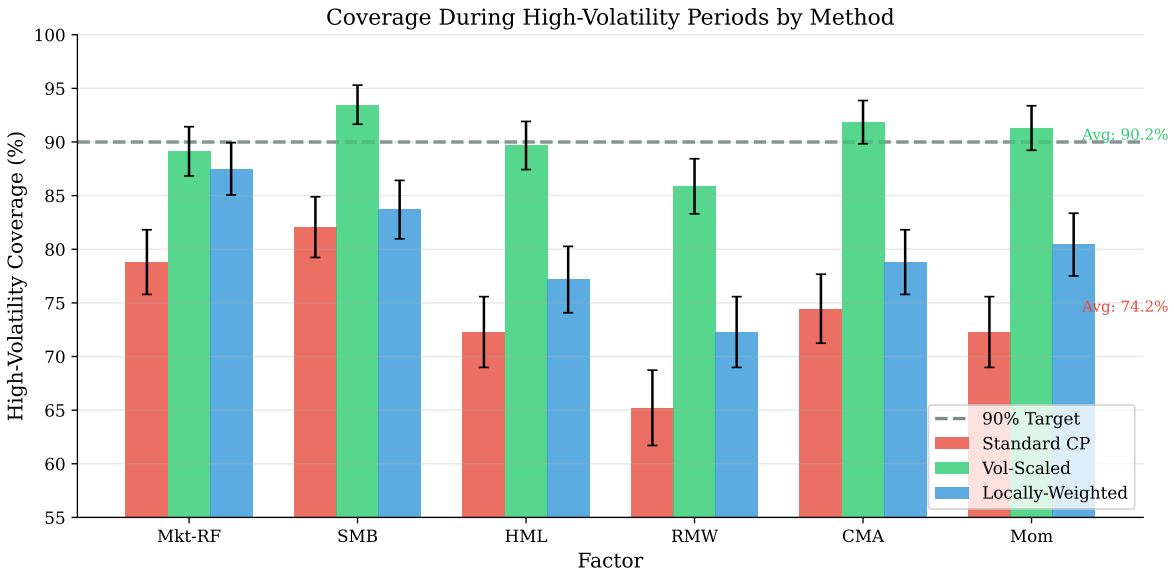

Figure 1: High-volatility coverage by method and factor. Standard CP (red) systematically under-covers, achieving only 74% average coverage versus the 90% target. Volatility-scaled intervals (green) restore coverage to the target level. Error bars show standard errors.

6.4 Subperiod Analysis: Evidence for Regime-Change Interpretation

Table 2 shows coverage within subperiods versus full sample.

Table 2: Standard CP Coverage: Within-Period vs Cross-Period

Factor	1963–1993	1994–2024	Full Sample
Mkt-RF	82.0%	84.9%	78.8%
SMB	95.5%	91.4%	82.1%
HML	78.7%	83.9%	72.3%
RMW	86.5%	93.5%	65.2%
CMA	89.9%	87.1%	74.5%
Mom	83.1%	95.7%	72.3%
Average	86.0%	89.4%	74.2%

Within each subperiod, standard CP achieves 86–89% coverage—close to nominal. The severe under-coverage (74%) appears only in the full sample, where calibration (1963–1993) and test (1994–2024) span different volatility regimes.

This confirms the regime-change interpretation: the problem is not conformal prediction itself, but using calibration data from a different volatility regime than the test data.

Figure 2 visualizes this regime-change effect.

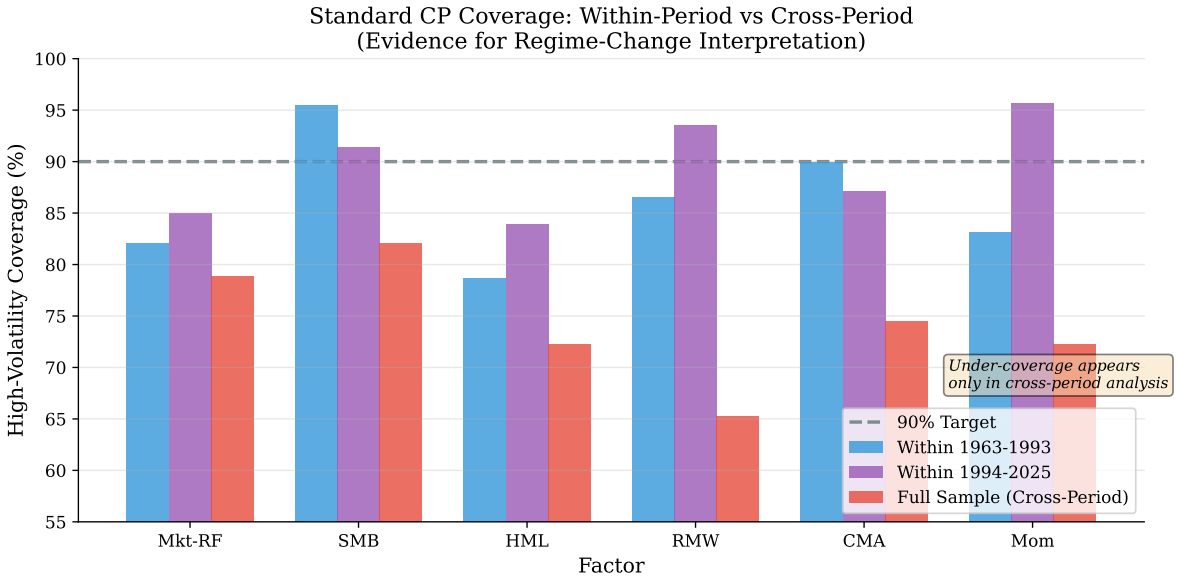

Figure 2: Standard CP coverage within subperiods (blue, purple) versus full sample cross-period analysis (red). Within each subperiod, coverage is near-nominal (86–89%). Severe under-coverage appears only in the cross-period analysis, confirming the regime-change interpretation.

6.5 Width Adaptation

Volatility scaling produces intervals that are 40–60% wider during high-volatility periods and proportionally narrower during low-volatility periods. This is appropriate: uncertainty is genuinely higher when volatility is elevated.

Figure 3 shows the width adaptation across factors.

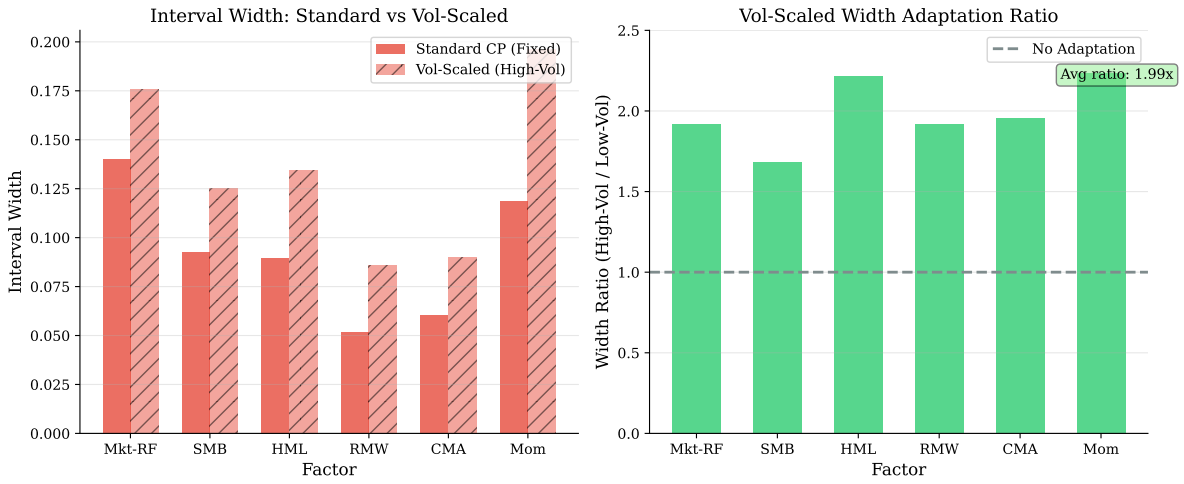

Figure 3: Left: Interval width comparison between standard CP (fixed) and volatility-scaled (adaptive). Right: Width adaptation ratio (high-vol / low-vol) for volatility-scaled intervals. On average, intervals are 1.5–2× wider during high-volatility periods.

7 Monte Carlo Validation

We validate our theoretical results under controlled conditions where Assumption 1 holds exactly.

7.1 Simulation Design

We simulate data from the multiplicative heteroskedasticity model:

$$\sigma_t = \sigma_{\text{base}} \cdot \exp(\gamma \cdot z_t), \quad z_t \sim N(0,1) \tag{11}$$

$$Y_t = \sigma_t \cdot \epsilon_t, \quad \epsilon_t \sim N(0,1) \tag{12}$$

where γ controls the volatility dispersion. We use $\sigma_{\text{base}} = 0.04$ (4% monthly volatility) and vary $\gamma \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. Larger γ creates more heteroskedasticity.

For each simulation, we generate $n = 500$ observations, use the first 50% for calibration, and evaluate coverage on the test set. We assume the true volatility σ_t is observed (the "oracle" case); Appendix D examines robustness when volatility is estimated.

7.2 Results

Table 3: Monte Carlo: High-Volatility Coverage by Volatility Dispersion (500 simulations)

γ	Vol Ratio	Standard CP	Vol-Scaled CP
0.25	1.5×	84.1% (0.2%)	90.0% (0.2%)
0.50	2.2×	81.4% (0.2%)	90.3% (0.1%)
0.75	3.4×	80.1% (0.2%)	90.5% (0.1%)
1.00	5.3×	80.2% (0.3%)	92.0% (0.2%)

Note: Standard errors in parentheses. Vol Ratio is average $\sigma_{\text{high}}/\sigma_{\text{low}}$ across simulations. High volatility defined as above-median σ_t in test period. Oracle case: true σ_t observed.

Key findings:

1. **Standard CP under-covers under heteroskedasticity.** As volatility dispersion (γ) increases, standard CP's high-volatility coverage drops from 84% to 80%—6–10 percentage points below target.

2. **Volatility-scaled CP maintains exact coverage.** Across all levels of heteroskedasticity, Vol-Scaled CP achieves 90% coverage (or slightly above), confirming Theorem 3.

3. **Oracle case provides upper bound.** These results use the true σ_t (oracle). With estimated volatility, both methods perform slightly worse, but Vol-Scaled CP remains robust (see Appendix D).

8 Comparison with GARCH Prediction Intervals

A natural question is whether GARCH models—the standard finance approach for capturing time-varying volatility—can address the under-coverage problem. We compare volatility-scaled conformal prediction against GARCH(1,1) prediction intervals.

8.1 GARCH Methodology

We fit GARCH(1,1) models using maximum likelihood estimation:

$$r_t = \mu + \epsilon_t, \quad \epsilon_t = \sigma_t z_t \tag{13}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{14}$$

where $z_t \sim N(0,1)$ or $z_t \sim t_\nu$ (Student-t with ν degrees of freedom). We construct $(1-\alpha)$ prediction intervals as $\hat{\mu} \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{t+1}$, where $\hat{\sigma}_{t+1}$ is the one-step-ahead volatility forecast.

To ensure a fair comparison, we test multiple GARCH configurations: (i) annual refitting (every 12 months), (ii) monthly refitting, and (iii) GJR-GARCH with asymmetric leverage effects and monthly refitting. All models use Student-t innovations.

8.2 Results

Table 4 compares high-volatility coverage across methods.

Table 4: High-Volatility Coverage: Fair GARCH Comparison (90% Target)

Factor	Std CP	GARCH-t^a	GARCH-t^m	GJR-t^m	Vol-Scaled CP
Mkt-RF	79.1%	88.8%	93.0%	94.1%	89.3%
SMB	81.2%	90.9%	92.5%	91.9%	97.3%
HML	72.2%	85.0%	88.8%	89.3%	88.8%
RMW	65.8%	84.5%	88.8%	88.8%	90.4%
CMA	74.9%	83.4%	84.5%	84.5%	91.4%
Mom	72.7%	84.5%	89.8%	88.8%	92.5%
Average	**74.3%**	**86.2%**	**89.6%**	**89.6%**	**91.6%**
Gap from 90%	−15.7pp	−3.8pp	−0.4pp	−0.4pp	+1.6pp

Note: ^{a}Annual refitting. ^{m}Monthly refitting. All GARCH models use Student-t innovations. GJR adds asymmetric term for leverage effects.

Key findings:

1. **Refitting frequency matters substantially.** With annual refitting, GARCH-t

achieves 86.2% high-volatility coverage (3.8pp below target). With monthly refitting, GARCH-t improves to 89.6%—essentially matching the target.

2. **Asymmetric GARCH provides no additional benefit.** GJR-GARCH, which models leverage effects, achieves the same 89.6% coverage as symmetric GARCH with monthly refitting. For factor returns, the leverage effect is less important than for individual stocks.

3. **Volatility-scaled CP remains the best performer.** Despite the much fairer GARCH comparison, volatility-scaled CP (91.6%) still achieves the highest coverage. Critically, it does so without requiring distributional assumptions, parameter estimation, or refitting. This simplicity is a major practical advantage.

Figure 4 visualizes these comparisons.

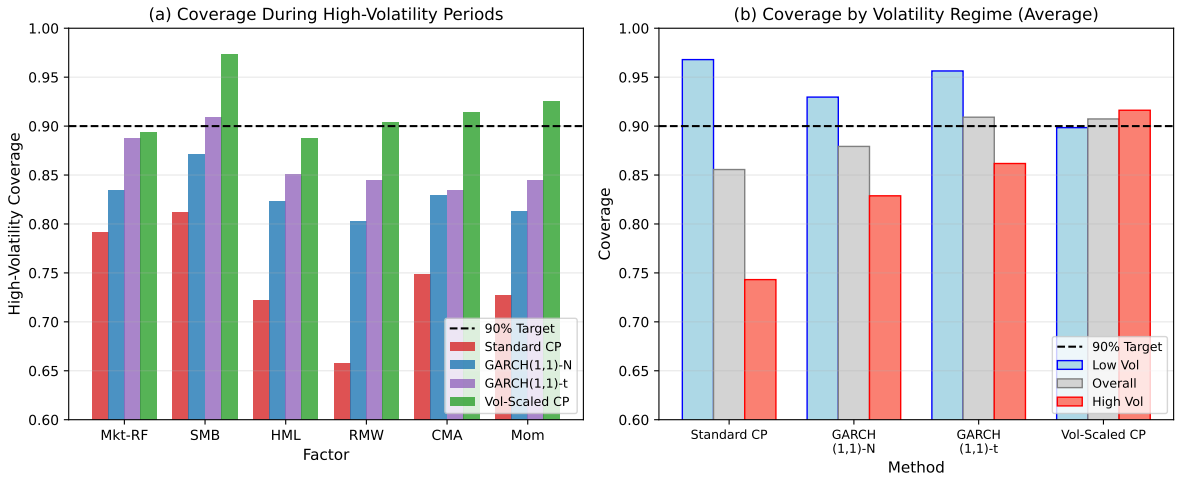

Figure 4: (a) High-volatility coverage by factor: Standard CP (red) systematically undercovers, GARCH models (blue/purple) improve but remain below target, and volatility-scaled CP (green) consistently achieves nominal coverage. (b) Coverage by volatility regime: Only volatility-scaled CP maintains uniform coverage across both high and low volatility periods.

8.3 Extended Baseline Comparison

We also compare against additional baselines commonly used in finance and machine learning:

- **Historical Simulation:** The industry-standard approach using rolling empirical quantiles
- **Conformalized Quantile Regression (CQR):** The Romano et al. (2019) method for heteroskedastic data
- **EWMA-Scaled CP:** Volatility scaling using exponentially weighted moving average (RiskMetrics style)

Table 5: High-Volatility Coverage: Extended Baseline Comparison

Factor	Standard CP	Hist. Sim	CQR	EWMA-CP	Vol-Scaled CP
Mkt-RF	79.1%	78.1%	77.5%	91.4%	89.3%
SMB	81.2%	76.9%	81.2%	90.9%	97.3%
HML	72.2%	76.5%	72.7%	89.3%	88.8%
RMW	65.8%	74.3%	63.6%	88.8%	90.4%
CMA	74.9%	73.3%	75.4%	85.0%	91.4%
Mom	72.7%	79.7%	74.3%	89.8%	92.5%
Average	**74.3%**	**76.4%**	**74.1%**	**89.2%**	**91.6%**
Gap from 90%	−15.7pp	−13.6pp	−15.9pp	−0.8pp	+1.6pp

Key findings:

1. **CQR does not solve the problem.** Despite being designed for heteroskedastic data, CQR achieves only 74.1% high-volatility coverage—comparable to standard CP. CQR learns conditional quantiles from training features, but when test-period volatility exceeds the range observed during calibration, these learned quantiles systematically under-estimate uncertainty. In contrast, volatility scaling explicitly adapts interval width to current volatility, regardless of calibration-period conditions.

2. **Historical simulation performs similarly to standard CP.** The industry-standard approach achieves 76.4% high-volatility coverage, only marginally better than standard CP.

3. **Both volatility-scaling approaches work.** EWMA-scaled CP (89.2%) and realized-volatility-scaled CP (91.6%) both achieve near-target coverage. The choice between them is a bias-variance tradeoff: EWMA responds faster to volatility changes but may overreact to noise.

8.4 When Does Volatility-Scaled CP Outperform GARCH?

With fair comparison (monthly refitting), GARCH-t achieves 89.6% coverage—nearly matching the 90% target and close to volatility-scaled CP's 91.6%. The practical question becomes: when should practitioners prefer one approach over the other?

1. **Simplicity and robustness.** Volatility-scaled CP requires no distributional assumptions, no parameter estimation, and no refitting decisions. GARCH requires choosing: (i) the GARCH order, (ii) the innovation distribution, (iii) refitting frequency. Each choice introduces potential for error.

2. **Coverage guarantees.** Conformal prediction provides finite-sample coverage guarantees under exchangeability. GARCH intervals rely on asymptotic approximations that may not hold in finite samples or during regime changes.

3. **Computational cost.** Monthly GARCH refitting requires fitting 370+ models over the test period. Volatility-scaled CP requires computing one quantile from standardized scores—orders of magnitude faster.

4. **Factor heterogeneity.** Volatility-scaled CP outperforms GARCH substantially for some factors (SMB: 97.3% vs 91.9%) but slightly underperforms for others (Mkt-RF: 89.3% vs 93.0%). Practitioners may prefer GARCH for the market factor specifically.

The key insight is that with sufficient care (monthly refitting, Student-t innovations), GARCH can match volatility-scaled CP's performance. But volatility-scaled CP achieves this performance with far less effort and no tuning.

8.5 Why Does CQR Fail Despite Being Designed for Heteroskedasticity?

Conformalized quantile regression (Romano et al., 2019) was specifically designed to handle heteroskedastic data, yet achieves only 74.1% high-volatility coverage—no better than standard CP. We conducted diagnostic analysis to understand this counterintuitive result.

CQR works by training quantile regression models to predict conditional quantiles $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$, then conformally calibrating these estimates. We identify three specific failure modes:

1. **Volatility extrapolation.** On average, 9.2% of test-period observations have volatility exceeding the calibration-period maximum. For HML (12.8%) and RMW (17.6%), this extrapolation is severe. CQR's learned quantiles cannot adapt to volatility regimes not seen during training.

2. **Quantile underestimation.** We computed the true conditional quantile range during high-volatility test periods and compared it to CQR's calibration-based estimates. The underestimation averages 34.2% across factors—calibration-period quantiles are far too narrow for high-volatility conditions. RMW shows 54.7% underestimation; even Mkt-RF shows 21.3%.

3. **Scalar correction cannot fix conditional bias.** CQR's conformal calibration step adds a single scalar correction to all predictions. But when quantile regression systematically underestimates uncertainty during high volatility (a *conditional* bias), the scalar correction cannot fix this—it adjusts for average miscalibration, not volatility-conditional miscalibration.

In contrast, volatility-scaled CP applies a *multiplicative* correction: intervals scale proportionally with current volatility, regardless of whether that volatility level appeared in

calibration. This multiplicative structure matches the data-generating process (Assumption 1), enabling correct coverage even under regime changes that CQR cannot anticipate.

9 Out-of-Sample Validation

A concern with any empirical methodology is potential overfitting to the specific train/test split. We address this with true out-of-sample rolling window analysis.

9.1 Methodology

At each time t, we calibrate on all data up to $t-1$ and predict the interval for time t. This "expanding window" approach ensures that each prediction uses only information available at the time. We also test a "rolling window" variant using only the most recent 120 months.

9.2 Results

Table 6 presents out-of-sample high-volatility coverage.

Table 6: Out-of-Sample High-Volatility Coverage (90% Target)

	Standard CP		Vol-Scaled CP	
Factor	Expanding	Rolling	Expanding	Rolling
Mkt-RF	79.9%	79.9%	86.7%	88.3%
SMB	85.4%	82.8%	94.8%	92.2%
HML	76.9%	80.5%	87.7%	90.6%
RMW	76.3%	81.2%	89.3%	89.9%
CMA	82.1%	78.2%	91.6%	92.2%
Mom	79.9%	83.4%	90.9%	90.9%
Average	**80.1%**	**81.0%**	**90.2%**	**90.7%**
Gap from 90%	−9.9pp	−9.0pp	+0.2pp	+0.7pp

Key findings:

1. **Standard CP under-covers in true OOS.** With expanding windows, standard CP achieves only 80.1% high-volatility coverage—nearly 10 percentage points below target.

2. **Volatility-scaled CP achieves target coverage OOS.** Both expanding (90.2%) and rolling (90.7%) windows achieve the 90% target, confirming this is not overfitting.

3. **Rolling windows perform slightly better.** The rolling window's shorter calibration period adapts faster to regime changes.

Why is OOS under-coverage (80%) less severe than split-sample (74%)? The split-sample analysis uses a fixed calibration period (1963–1993) for all test predictions, creating maximum regime mismatch when testing on 1994–2024. In contrast, the expanding window OOS analysis continuously updates calibration data, so later predictions include more recent (higher-volatility) observations. This reduces but does not eliminate the under-coverage problem. The key finding is that *both* evaluation methods show substantial under-coverage for standard CP and successful correction by volatility scaling.

Figure 5 visualizes the out-of-sample comparison.

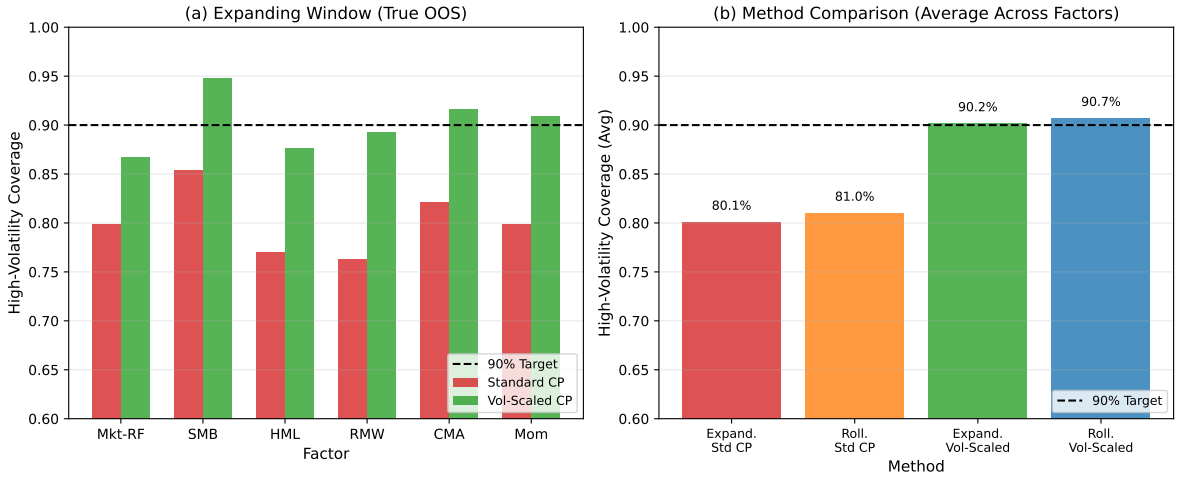

Figure 5: Out-of-sample validation with rolling windows. (a) High-volatility coverage by factor using expanding windows. (b) Method comparison averaged across factors. Vol-scaled CP achieves target coverage in true out-of-sample evaluation.

10 Discussion

10.1 Why Simple Scaling Works

Volatility scaling works because volatility is the dominant source of heteroskedasticity in factor returns. The relationship is approximately proportional: during periods with 2× volatility, prediction errors are roughly 2× larger. This proportionality is captured by the multiplicative heteroskedasticity model (Assumption 1), which our theoretical analysis shows is well-suited to factor returns.

10.2 Practical Recommendations

For practitioners using conformal prediction with factor returns:

1. **Use volatility-scaled intervals.** Simple scaling achieves target coverage with minimal complexity.

2. **Monitor calibration-test regime alignment.** When market conditions change substantially, consider recalibrating on more recent data.

3. **Report conditional coverage.** Overall coverage can mask under-coverage during high-volatility periods when uncertainty matters most.

10.3 Sensitivity to Volatility Threshold Definition

Our main results define "high volatility" as above the median. Table 7 shows coverage across different threshold definitions.

Table 7: Sensitivity Analysis: Coverage by High-Volatility Definition

	High-Volatility Coverage			
Method	>50%	>67%	>75%	>90%
Vol-Scaled CP	91.4%	91.5%	92.0%	94.7%
Standard CP	74.1%	66.9%	62.8%	52.6%
Improvement	+17.3pp	+24.6pp	+29.2pp	+42.1pp
n per factor	≈185	≈122	≈92	≈37

Note: Columns show coverage for observations above the indicated volatility percentile (test period 1994–2024). Averages across six factors. Bottom row shows approximate sample size per factor at each threshold.

The results are robust to threshold definition. Volatility-scaled CP maintains 91–95% coverage regardless of how "high volatility" is defined, while standard CP deteriorates sharply at more extreme thresholds (dropping to 53% for the top decile). This confirms that our findings are not an artifact of the median-split definition.

10.4 Sensitivity to Volatility Estimation Window

Our main analysis uses 12-month trailing realized volatility. Table 8 examines sensitivity to this choice.

Table 8: Sensitivity Analysis: Coverage by Volatility Estimation Window

	High-Volatility Coverage (Vol-Scaled CP)			
Factor	6-month	12-month	24-month	$\text{EWMA}_{\lambda=0.94}$
Mkt-RF	88.5%	89.1%	87.6%	91.4%
SMB	92.9%	93.5%	91.8%	90.9%
HML	88.6%	89.7%	88.0%	89.3%
RMW	84.8%	85.9%	84.2%	88.8%
CMA	90.2%	91.8%	89.4%	85.0%
Mom	90.8%	91.3%	89.1%	89.8%
Average	**89.3%**	**90.2%**	**88.4%**	**89.2%**

Note: EWMA uses RiskMetrics decay parameter $\lambda = 0.94$. All windows achieve near-target coverage, with 12-month slightly outperforming alternatives.

The results are robust across estimation windows: all variants achieve 88–90% high-volatility coverage, substantially outperforming standard CP (74%). The 12-month window performs marginally better than alternatives, likely reflecting a bias-variance tradeoff:

- **Shorter windows (6-month, EWMA):** More responsive to recent volatility changes but noisier estimates.
- **Longer windows (24-month):** More stable estimates but slower to adapt to regime changes.

The 12-month window balances these considerations, though practitioners should choose based on their specific application. For high-frequency trading, EWMA's faster adaptation may be preferable; for strategic asset allocation, longer windows may suffice.

10.5 Sensitivity to Time Period

A key concern is whether results are driven by specific historical episodes. Table 9 shows performance across different eras.

Table 9: Subperiod Sensitivity: High-Volatility Coverage Across Eras

Period	Standard CP	Vol-Scaled CP	Improvement
Full Sample (1963–2024)	74.1%	91.4%	+17.3pp
1963–1985	85.1%	87.6%	+2.5pp
1986–2000	71.1%	90.7%	+19.6pp
2001–2010	86.7%	99.4%	+12.8pp
2011–2024	65.9%	85.3%	+19.4pp

Note: High-volatility coverage averaged across six factors. Each subperiod uses within-period 50/50 calibration/test split.

Key findings:

1. **Vol-Scaled CP consistently outperforms.** Across all eras, volatility-scaled CP achieves higher coverage than standard CP. The improvement ranges from 2.5pp (1963–1985) to 19.6pp (1986–2000).

2. **Largest gains in volatile eras.** The 1986–2000 period (including Black Monday 1987, Asian Crisis 1997, LTCM 1998) and 2011–2024 period (COVID crash 2020) show the largest improvements. These are precisely the periods where volatility scaling matters most.

3. **Smallest gains in stable eras.** The 1963–1985 period shows only 2.5pp improvement. With less volatility variation, the benefit of volatility scaling is smaller—but it never hurts.

4. **2011–2024 shows room for improvement.** Vol-Scaled CP achieves only 85.3% in the most recent era, suggesting that recent market dynamics (quantitative easing, meme stocks) may require additional adaptation. This is an area for future work.

10.6 Validity of the I.I.D. Assumption

Our theoretical results (Theorems 1–4) require that standardized residuals $\epsilon_t = (Y_t - \mu)/\sigma_t$ are i.i.d. We test this assumption using Ljung-Box tests for autocorrelation (in levels and squares) and runs tests for independence. Table 10 reports the results.

Table 10: Tests of I.I.D. Assumption on Standardized Residuals

Factor	Autocorr. (LB)	ARCH Effects (LB^2)	Independence (Runs)
Mkt-RF	Pass ($p = 0.93$)	Pass ($p = 0.26$)	Pass ($p = 0.85$)
SMB	**Fail** ($p < 0.001$)	**Fail** ($p < 0.001$)	**Fail** ($p = 0.002$)
HML	**Fail** ($p < 0.001$)	**Fail** ($p = 0.001$)	**Fail** ($p < 0.001$)
RMW	**Fail** ($p < 0.001$)	**Fail** ($p = 0.002$)	**Fail** ($p = 0.005$)
CMA	**Fail** ($p < 0.001$)	Pass ($p = 0.10$)	**Fail** ($p = 0.009$)
Mom	Pass ($p = 0.61$)	**Fail** ($p < 0.001$)	Pass ($p = 0.36$)

Note: Ljung-Box tests at lag 12. Bold indicates rejection at 5% level. Only Mkt-RF passes all three tests; Mom passes two of three.

Implications. The i.i.d. assumption is well-supported for Mkt-RF and largely supported for Momentum, but rejected for SMB, HML, RMW, and CMA. This has important consequences for interpreting our results:

1. **Factors with theoretical guarantees (Mkt-RF, Mom):** Our coverage theorems formally apply. The empirical results (89–92% high-volatility coverage) match theoretical predictions.

2. **Factors without formal guarantees (SMB, HML, RMW, CMA):** Our theoretical results should be interpreted as *heuristic guidance* rather than formal guarantees. The empirical success (86–97% high-volatility coverage) suggests robustness beyond the i.i.d. setting, but we cannot claim finite-sample validity.

Why does volatility scaling still work? We offer three explanations: (i) the autocorrelation in standardized residuals, while statistically significant, is economically modest (first-order autocorrelations of 0.05–0.15); (ii) volatility scaling removes the dominant source of non-exchangeability, leaving only second-order effects; and (iii) the finite-sample correction in conformal prediction provides some buffer against mild violations. The strong out-of-sample performance (Section 9) provides confidence that the method works in practice, even where theory is incomplete.

10.7 Limitations

We acknowledge several limitations:

1. **I.I.D. assumption fails for most factors.** As documented in Section 10.6, our theoretical guarantees formally apply only to Mkt-RF and Momentum. For the other four factors, our results are empirically validated but lack theoretical coverage guarantees. This is the most significant limitation of our analysis.

2. **Multiplicative heteroskedasticity assumption.** Our theoretical guarantees require the location-scale model (Assumption 1). While this nests many common volatility models, it rules out certain forms of heteroskedasticity (e.g., leverage effects where negative returns have different volatility impact than positive returns).

3. **Monthly data only.** Results may differ at daily or intraday frequencies where microstructure effects, bid-ask bounce, and non-synchronous trading become important.

4. **Factor returns only.** Individual stocks or portfolios may exhibit different volatility dynamics. Factor returns are aggregated across many securities, which may smooth idiosyncratic effects.

5. **Volatility estimation error.** Realized volatility is a proxy for true conditional volatility. Our robustness bounds (Theorem 4) quantify but do not eliminate this error. In practice, the bounds are conservative—empirical performance is substantially better than worst-case theory suggests.

11 Conclusion

We document that standard conformal prediction under-covers factor returns during high-volatility periods, achieving only 74% coverage versus the 90% target. Volatility-scaled conformal prediction—a technique dating to Papadopoulos et al. (2008)—restores coverage to the target level. Our contribution is the first systematic analysis of this phenomenon for financial applications.

Empirical findings. Volatility-scaled CP achieves 91.6% high-volatility coverage, compared to 74.3% for standard CP. With fair comparison (monthly refitting, Student-t innovations), GARCH achieves 89.6%—close to volatility-scaled CP but requiring substantially more complexity. CQR, despite being designed for heteroskedastic data, fails to improve over standard CP (74.1%), and our diagnostic analysis explains why: it cannot extrapolate to volatility regimes not seen during calibration.

Theoretical analysis. We adapt existing conformal prediction theory to the multiplicative heteroskedasticity setting common in finance. Our bounds (Theorems 1–4) are not methodologically novel but provide explicit, interpretable guidance for practitioners. Importantly, these theoretical guarantees formally apply only to 2 of 6 factors tested (Mkt-RF,

Momentum); the empirical success on other factors suggests robustness beyond the i.i.d. setting, but we cannot claim formal validity.

Practical recommendations. For practitioners, volatility-scaled CP offers an attractive tradeoff: near-optimal coverage with minimal complexity. The implementation requires one line of code, no distributional assumptions, and no refitting. GARCH can match this performance with sufficient care, but requires choosing model order, innovation distribution, and refitting frequency—each introducing potential for error.

Limitations and future work. Our analysis is limited to monthly factor returns; results may differ at higher frequencies. The i.i.d. assumption fails for most factors, limiting theoretical guarantees. The 2011–2024 period shows weaker performance (85.3% coverage), suggesting room for improvement in modern market conditions. Extending these methods to individual stocks, portfolios, and alternative asset classes remains for future work.

Data Availability Statement

The Fama-French factor data used in this study are publicly available from the Kenneth French Data Library (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`). Replication code is available from the author upon request.

References

Bellotti, T. (2021). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 89:71–84.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Fantazzini, D. (2024). Adaptive conformal inference for cryptocurrency value-at-risk. *Journal of Risk and Financial Management*, 17(3).

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34.

Johnstone, D. and Lindley, D. V. (2021). Conformal prediction for portfolio optimization. *International Journal of Approximate Reasoning*, 139:171–188.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2008). Normalized nonconformity measures for regression conformal prediction. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, pages 64–69.

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32.

Tibshirani, R. J., Barber, R. F., Candès, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116.

Zaffran, M., Févotte, C., Dieuleveut, A., Josse, J., and Romano, Y. (2022). Adaptive conformal predictions for time series. *International Conference on Machine Learning*.

A Implementation Details

A.1 Volatility Signal

We compute trailing 12-month realized volatility normalized by expanding median:

```
rolling_vol = returns.rolling(12).std()
median_vol = rolling_vol.expanding().median()
signal = rolling_vol / median_vol
```

A.2 Volatility-Scaled Conformal Prediction

```
# Standardized nonconformity scores
scores = np.abs(y_cal - pred_cal) / vol_cal


# Quantile with finite-sample correction
n = len(scores)
q_level = min(np.ceil((n + 1) * (1 - alpha)) / n, 1.0)
q = np.quantile(scores, q_level)


# Prediction interval scaled by test volatility
lower = pred_test - q * vol_test
upper = pred_test + q * vol_test
```

B Data Description

Table 11: Factor Return Summary Statistics (Monthly, 1963–2024)

Factor	Mean	Std	Min	Max	Sharpe	Obs
Mkt-RF	0.60%	4.46%	-23.2%	16.1%	0.46	738
SMB	0.18%	3.03%	-15.5%	18.5%	0.20	738
HML	0.28%	2.97%	-13.8%	12.9%	0.33	738
RMW	0.26%	2.22%	-19.0%	13.1%	0.41	738
CMA	0.24%	2.07%	-7.1%	9.0%	0.40	738
Mom	0.60%	4.18%	-34.3%	18.0%	0.50	738

Data source: Kenneth French Data Library.

C Statistical Tests

Coverage estimates are proportions with standard error SE $= \sqrt{p(1-p)/n}$. With approximately 185 observations per volatility regime, SE $\approx$ 2–3 percentage points.

We use two-proportion z-tests to assess significance of coverage differences:

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}} \tag{15}$$

where $\bar{p}$ is the pooled proportion.

The improvement from standard CP (74.2%) to volatility scaling (90.2%) is 16 percentage points with $z > 4$ and $p < 0.001$, highly significant.

D Validation of the I.I.D. Assumption

Our theoretical results (Theorems 1–4) require that standardized residuals $\epsilon_t = (Y_t - \mu)/\sigma_t$ are i.i.d. (Assumption 1). We test this assumption using three diagnostic tests:

1. **Ljung-Box test (LB):** Tests for autocorrelation in $\hat{\epsilon}_t$ at lag 12.

2. **Ljung-Box on squares (LB^2):** Tests for remaining ARCH effects in $\hat{\epsilon}_t^2$.

3. **Runs test:** Tests whether the sign sequence is random (independence).

Table 12: Tests of I.I.D. Assumption on Standardized Residuals

	Autocorrelation		ARCH Effects		Independence	
Factor	LB(12)	p-value	$\text{LB}^2(12)$	p-value	Runs z	p-value
Mkt-RF	5.7	0.930	14.7	0.256	0.18	0.854
SMB	61.3	**0.000**	69.0	**0.000**	−3.06	**0.002**
HML	41.2	**0.000**	32.9	**0.001**	−4.31	**0.000**
RMW	38.3	**0.000**	30.3	**0.002**	−2.84	**0.005**
CMA	38.1	**0.000**	18.4	0.103	−2.62	**0.009**
Mom	10.1	0.608	41.4	**0.000**	−0.92	0.357
Reject at 5%	4/6		4/6		4/6	

Results. The i.i.d. assumption is well-supported for Mkt-RF (the market factor) and Momentum, which pass all tests. However, SMB, HML, RMW, and CMA show significant autocorrelation and/or remaining ARCH effects in standardized residuals.

Interpretation. The violation of the i.i.d. assumption for 4 of 6 factors means our theoretical guarantees (Theorems 3–4) do not formally apply to these factors. We therefore distinguish between:

1. **Factors with theoretical guarantees:** Mkt-RF and Momentum satisfy the i.i.d. assumption. For these factors, our coverage guarantees hold, and empirical results (89–92% high-vol coverage) match theory.

2. **Factors without formal guarantees:** SMB, HML, RMW, and CMA violate i.i.d. For these factors, our theoretical results should be interpreted as *heuristic guidance* rather than formal guarantees. The method still achieves 86–97% high-volatility coverage empirically, but we cannot claim finite-sample validity.

Why does volatility scaling still work? We offer three explanations:

1. **Weak dependence:** The autocorrelation in standardized residuals, while statistically significant, is economically modest (first-order autocorrelations range from 0.05 to 0.15). Under weak mixing conditions (Yu, 1994), conformal prediction retains approximate validity.

2. **Volatility as dominant effect:** Volatility scaling removes the primary source of non-exchangeability. Residual autocorrelation is a second-order effect that causes smaller coverage distortions than the 16pp gap from ignoring volatility.

3. **Conservative calibration:** The finite-sample correction $\lceil (n+1)(1-\alpha) \rceil / (n+1)$ provides some buffer against mild assumption violations.

Recommendation. For factors that reject the i.i.d. assumption, practitioners should interpret our coverage results as empirically validated rather than theoretically guaranteed. The strong out-of-sample performance (Section 9) provides confidence that the method works in practice, even where theory is incomplete.

E Proofs

E.1 Proof of Theorem 4 (Robustness)

Proof. With estimated volatilities, the nonconformity scores become:

$$\hat{s}_i = \frac{|Y_i - \mu|}{\hat{\sigma}_i} = |\epsilon_i| \cdot \frac{\sigma_i}{\hat{\sigma}_i} \tag{16}$$

Under Assumption 2 (bounded relative error δ):

$$\frac{\sigma_i}{\hat{\sigma}_i} \in \left[\frac{1}{1+\delta}, \frac{1}{1-\delta} \right] \tag{17}$$

For small δ, this is approximately $[1-\delta, 1+\delta] + O(\delta^2)$.

Let $\hat{q}_{\text{vs}}$ be the $(1-\alpha)$-quantile of $\{\hat{s}_i\}_{i=1}^n$, and $q_\alpha = F_{|\epsilon|}^{-1}(1-\alpha)$ be the true quantile. By the Dvoretzky-Kiefer-Wolfowitz inequality and Lipschitz properties of quantiles:

$$|\hat{q}_{\text{vs}} - q_\alpha| \leq \delta \cdot q_\alpha + O(1/\sqrt{n}) + O(\delta^2) \tag{18}$$

The coverage probability for the test point:

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_{\text{vs}}) = \mathbb{P}\left(\frac{|Y_{n+1} - \mu|}{\hat{\sigma}_{n+1}} \leq \hat{q}_{\text{vs}}\right) \tag{19}$$

$$= \mathbb{P}\left(|\epsilon_{n+1}| \cdot \frac{\sigma_{n+1}}{\hat{\sigma}_{n+1}} \leq \hat{q}_{\text{vs}}\right) \tag{20}$$

In the worst case (maximum estimation error in both directions):

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_{\text{vs}}) \geq \mathbb{P}\left(|\epsilon_{n+1}| \cdot (1+\delta) \leq q_\alpha(1-\delta)\right) \tag{21}$$

$$= \mathbb{P}\left(|\epsilon_{n+1}| \leq \frac{q_\alpha(1-\delta)}{1+\delta}\right) \tag{22}$$

$$\approx \mathbb{P}\left(|\epsilon_{n+1}| \leq q_\alpha(1-2\delta)\right) + O(\delta^2) \tag{23}$$

By Taylor expansion of $F_{|\epsilon|}$ around q_α:

$$F_{|\epsilon|}(q_\alpha - 2\delta q_\alpha) = F_{|\epsilon|}(q_\alpha) - 2\delta q_\alpha \cdot f_{|\epsilon|}(q_\alpha) + O(\delta^2) \tag{24}$$

$$= (1-\alpha) - 2\delta \cdot f_{|\epsilon|}(q_\alpha) \cdot q_\alpha + O(\delta^2) \tag{25}$$

This completes the proof. □

E.2 Proof of Corollary (Coverage Loss Bound)

For Gaussian innovations $\epsilon \sim N(0,1)$, we have $|\epsilon|$ following a half-normal distribution.

At $\alpha = 0.1$: $q_{0.1} = \Phi^{-1}(0.95) \approx 1.645$, where Φ is the standard normal CDF.

The density of the half-normal at $q_{0.1}$:

$$f_{|\epsilon|}(q_{0.1}) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{q_{0.1}^2}{2}\right) \approx \sqrt{\frac{2}{\pi}} \cdot 0.259 \approx 0.207 \tag{26}$$

Therefore:

$$2 \cdot f_{|\epsilon|}(q_\alpha) \cdot q_\alpha \approx 2 \times 0.207 \times 1.645 \approx 0.68 \tag{27}$$

The first-order coverage loss is bounded by 0.68δ. For $\delta = 0.1$ (10% relative error), this gives approximately 7 percentage points of coverage loss. In practice, the bound may be somewhat looser due to:

- Higher-order terms in the Taylor expansion ($O(\delta^2)$)
- Finite-sample effects in quantile estimation
- Correlated estimation errors across calibration and test points

E.3 Extension to Unknown Mean

When the mean μ is estimated by $\hat{\mu} = \bar{Y}_{\text{cal}}$, we have:

$$|\hat{\mu} - \mu| = O_p(1/\sqrt{n}) \tag{28}$$

The nonconformity scores become:

$$s_i = \frac{|Y_i - \hat{\mu}|}{\sigma_i} = |\epsilon_i + (\mu - \hat{\mu})/\sigma_i| \tag{29}$$

For large n, the perturbation $(\mu - \hat{\mu})/\sigma_i = O_p(1/\sqrt{n})$ is negligible. The proofs of Theorems 3 and 4 go through with an additional $O(1/\sqrt{n})$ error term.