Conformal Prediction Intervals for Daily Equity Returns Under Volatility Regimes

Fathimah Al-Mashumah[1], Yasi Dani[1]
[1]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
Email: fathimah.al-mashumah@binus.ac.id | ORCID: 0009-0008-9261-5542
Email: yasi.dani@binus.ac.id | ORCID: 0000-0002-6712-664X

Abstrak. *Laporan risiko bergantung pada prakiraan interval, tetapi model volatilitas parametrik kehilangan cakupan nominalnya ketika pasar bergejolak. Kami menguji apakah prediksi konformal yang bebas distribusi dapat memberikan interval satu langkah ke depan yang andal untuk imbal hasil saham harian di berbagai rezim volatilitas. Di bawah satu protokol daring, kami membandingkan tolok ukur Gaussian tak bersyarat, interval GARCH(1,1) Gaussian dan Student-t, prediktor konformal terbelah berskala volatilitas, dan inferensi konformal adaptif, bersama lima metode interval berbasis pembelajaran mesin, pada empat indeks Eropa dan indeks S&P 500 dari 1990 hingga 2026. Tolok ukur tak bersyarat memiliki cakupan berlebih secara keseluruhan, namun hanya mencakup sekitar 85 persen hari bervolatilitas tinggi. Interval GARCH efisien dan terkalibrasi baik, sedangkan metode konformal memberikan kalibrasi marginal paling akurat dan cakupan terbaik pada hari bergejolak. Uji signifikansi dan analisis sensitivitas menegaskan bahwa koreksi konformal berskala volatilitas paling andal saat bergejolak, dengan biaya lebar yang kecil.*

Kata Kunci: *prediksi konformal; interval prediksi; imbal hasil saham; rezim volatilitas; uji-balik cakupan*

Abstract. *Risk reports rely on interval forecasts, yet parametric volatility models lose their nominal coverage when markets turn turbulent. We ask whether distribution-free conformal prediction can give dependable one-step-ahead intervals for daily equity returns across volatility regimes. Under one online protocol we compare an unconditional Gaussian benchmark, Gaussian and Student-t GARCH(1,1) intervals, a volatility-scaled split conformal predictor, and adaptive conformal inference, alongside five machine learning interval methods, on four European indices and the S&P 500 from 1990 to 2026. The unconditional benchmark overcovers yet covers only about 85 percent of high-volatility days. The GARCH intervals are efficient and well calibrated; the conformal methods give the most accurate marginal coverage and the best coverage on turbulent days. Significance tests and a sensitivity analysis confirm that volatility-scaled conformal correction is most reliable in turbulence, at a small cost in width.*

Keywords: *conformal prediction; prediction intervals; equity returns; volatility regimes; coverage backtesting*

1. Introduction

Good financial decisions need more than a point forecast of returns. They need an honest picture of the uncertainty around that forecast. For daily equity returns the predictable part of the conditional mean is small and unstable, and recent studies find that machine learning adds only modest gains to point forecasts, with the gains depending on the horizon and showing up in particular market segments [1], [2], [3]. Critical work on factor-based learning reaches the same conclusion about how fragile return predictability is [4], and a recent head-to-head comparison of classical and deep models for daily equity returns confirms that any predictive

signal is modest and only visible under strict leakage controls [5]. When the mean carries little signal, what matters in practice is the second moment and the reliability of the intervals a model reports. Well-calibrated intervals drive risk limits, margining, capital adequacy assessment, and portfolio construction. Interval quality is therefore a goal in its own right, not a by-product of point prediction.

Daily returns make interval forecasting hard. They are non-stationary and heavy-tailed, and they cluster in volatility and break structurally [6]. These same features have driven a large literature on instance normalization and adaptive differencing, which stabilize the input distribution before a model is fit [7], [8], [9], and on architectures that model trend and change points directly [10]. Classical interval methods usually assume a fixed parametric error distribution instead. A Gaussian interval of constant width ignores volatility that moves over time, and even a volatility-scaled Gaussian interval can get the tail risk wrong when returns are strongly leptokurtic. Researchers have studied regression-based interval forecasts for returns directly [11], and deep models keep being applied to index returns across markets [12]. Yet these papers rarely say whether the reported intervals actually reach their nominal coverage, or whether they hold up in both calm and turbulent periods.

Two gaps motivate this work. The conformal time series literature and the financial volatility literature have grown up mostly apart, so we still do not know whether distribution-free calibration beats a well-specified GARCH interval for daily equity returns, or what it costs in width. There is a second gap as well. Studies that evaluate interval forecasts usually report a single marginal coverage number, and that one number can hide large differences between calm and turbulent regimes [13], [14]. We tackle both gaps with a controlled benchmark rather than a new model. We put five interval constructions head-to-head on five equity return series under one strictly chronological online protocol, and we measure coverage both marginally and within the prevailing volatility regime. The work is empirical and deliberately incremental. It shows when a distribution-free calibration layer earns its width penalty, it puts numbers on the regime dependence that a single coverage figure hides, and it ships a fully reproducible pipeline on public data. We pursue four aims: comparing parametric and conformal one-step-ahead intervals on marginal coverage, mean width, and a combined sharpness coverage score; tracking how coverage moves across low, medium, and high-volatility regimes; checking that the findings hold across four European indices and a heavy-tailed United States index; and releasing a deterministic, one-click reproduction of every table and figure.

The contribution of this study is therefore not a new estimator but a controlled and reproducible answer to a practical question, and it has three specific parts. First, it provides a like-for-like evaluation of parametric and distribution-free intervals under one strictly causal online protocol with shared forecast origins, so that any difference in coverage is attributable to the interval construction alone. Second, it stratifies calibration by volatility regime and shows, with formal coverage tests, that a healthy marginal coverage can hide a severe conditional shortfall exactly when risk is greatest. Third, it broadens the comparison to five modern machine learning interval methods and supports every claim with tests of statistical significance and a sensitivity analysis, which together clarify when a light volatility-scaled conformal correction is worth its small width premium. The emphasis throughout is on calibrated risk under turbulence rather than on point accuracy.

2. Literature Review

Conformal prediction offers a different route. It builds prediction sets with finite-sample, distribution-free coverage guarantees under an exchangeability assumption, and it can wrap any underlying point predictor [15], [16]. The split, or inductive, variant calibrates a nonconformity score on a held-out sample and is computationally light [17], [18], [19]. Because raw absolute residuals produce intervals of constant width, normalized conformal predictors scale the score by a local difficulty estimate so that the interval widens where the model is less certain [20], and recent work weights the nonconformity score by data-dependent distances to adapt interval width and improve conditional coverage [21]. In a return-forecasting context the

natural difficulty estimate is a volatility forecast, which links conformal calibration directly to the heteroskedastic structure of the data.

Financial time-series data violate exchangeability, and a growing body of work adapts conformal prediction to dependent, drifting data. Adaptive conformal inference updates the effective mis-coverage level online in response to recent errors, keeping long-run coverage close to nominal even under arbitrary distribution shift [22], [23]. Related methods automate the learning rate [24] or control risk under explicit non-exchangeability [25]. Recent theory also explains why split conformal remains effective despite temporal dependence, provided the evaluation respects the ordering of time [26]. These results make conformal calibration a credible candidate for daily return intervals, but they are largely established on generic regression tasks or simulation rather than in a head-to-head comparison with the volatility models that practitioners already trust.

Conformal and interval methods are seeing more use in finance and nearby fields. Researchers have paired temporal conformal prediction with deep learning to forecast equity indices [27], and they have compared conformal layers against quantile regression and recurrent networks for time series under uncertainty [28]. Conformal prediction has sharpened probabilistic forecasts of photovoltaic power [29], electricity prices [30], and atmospheric pollutant concentrations [31], and shape adaptive quantile regression has been put forward for probabilistic intervals [32]. A parallel line of work joins signal decomposition, such as variational mode decomposition, with deep learning to sharpen stock and volatility forecasts [33], [34], [35]. Related forecasting work pairs variational mode decomposition with regularized regression for stock prediction [36], weights least-squares estimates toward recent data to track structural breaks in returns [37] and builds modular deep networks with reversible normalization to stay robust across regime shifts [38]. Beyond finance, machine learning pipelines for classification and risk screening in the informatics literature combine feature selection, imputation, and ensemble models [39], [40], yet they too report point decisions without calibrated uncertainty, which is the gap conformal calibration fills. Most of these studies chase point accuracy or report only marginal coverage, and few look at the calibration layer alone on equity returns across several markets.

At the same time, parametric volatility models stay strong, and sometimes dominate, as benchmarks for interval and value at risk forecasting. One direct comparison across volatility states found that old-fashioned parametric models are still the best choice in many settings [41]. Risk management researchers have also built careful backtests that separate marginal coverage from the way violations behave over time, including duration-based tests [13] and expected shortfall backtests [42]. Recent work brings the same backtesting logic to conformal prediction and tests both marginal and conditional coverage for non-stationary series [14]. That link between conformal calibration and value at risk backtesting shapes the evaluation we use here.

3. Methodology

3.1. Research Design

We run an out-of-sample comparison of one-step-ahead prediction intervals for daily equity returns. We fix the target coverage at 90 percent, so the nominal miscoverage rate is alpha equal to 0.10. Every method runs under the same online protocol, which respects the time ordering of the data, and every method shares the same forecast origins. Any difference between methods, therefore, comes from the interval construction, not from the sample. We deliberately favor transparency and replicability over model complexity because our goal is to benchmark calibration, not to squeeze out the best point accuracy. Figure 1 lays out the whole design, from the input series through the five interval constructions to the coverage-based evaluation, and the rest of this section defines each part.

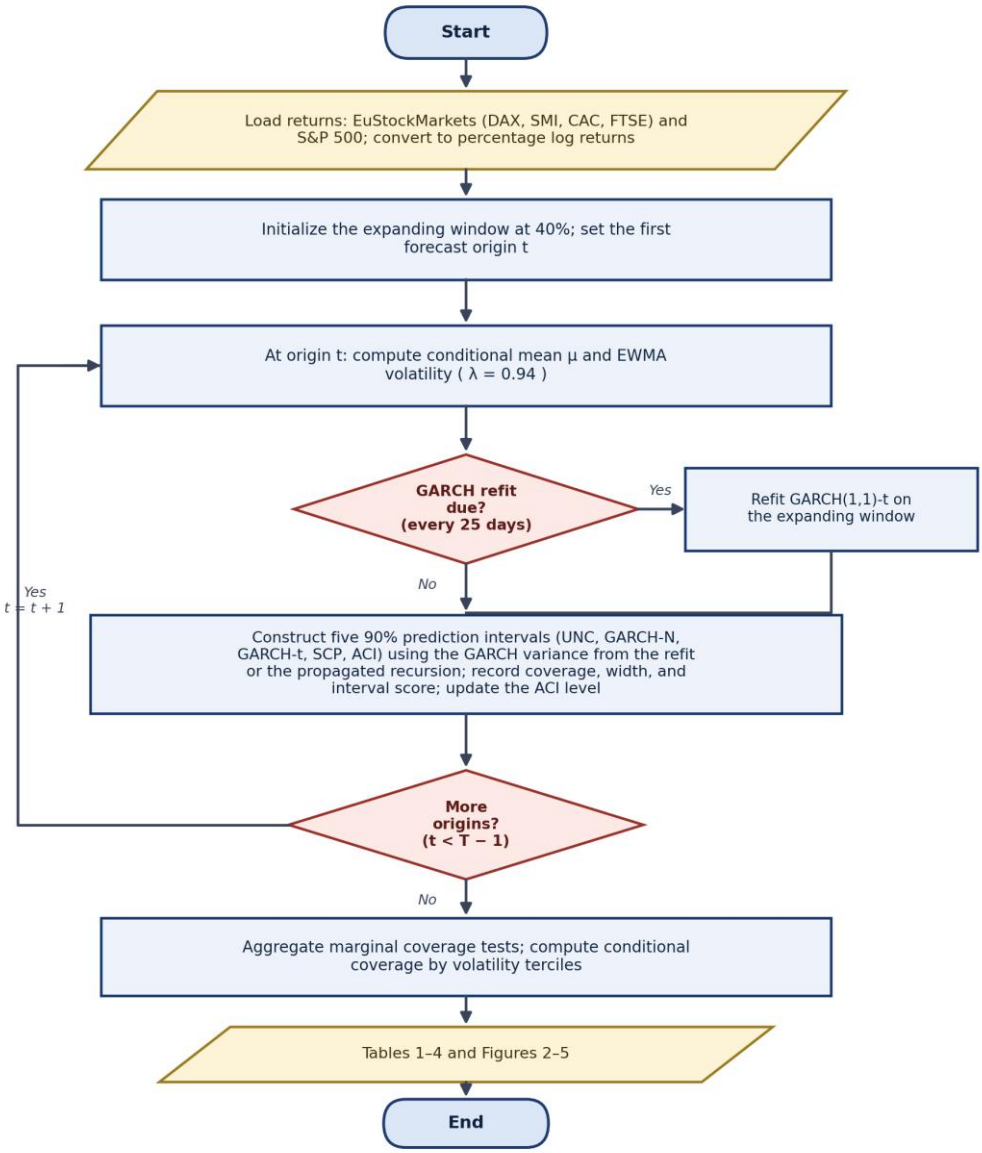


Figure 1. Research design of the study. Daily returns enter a strictly chronological online protocol; at each forecast origin a conditional mean and two volatility scales feed five interval constructions, which are then evaluated both marginally and within volatility regimes.

3.2. Data and Returns

Five daily equity index series are used, namely the German DAX, the Swiss SMI, the French CAC 40, the British FTSE 100, and the United States S&P 500. Each series is the daily closing level from the start of 1990 through July 2026, except the FTSE 100, which begins in 2001; the four longer series therefore span more than thirty-five years and include the dot com collapse, the 2008 global financial crisis, and the 2020 pandemic shock. Returns are computed as one hundred times the natural logarithm of the ratio of consecutive closing levels, which yields percentage log returns. After differencing, the number of daily observations is 9,068 for the DAX; 8,730 for the SMI; 9,093 for the CAC; 6,311 for the FTSE; and 8,969 for the S&P 500, for more than forty-two thousand return observations in total. The DAX, SMI, and CAC levels are drawn from Yahoo Finance, and the FTSE 100 and S&P 500 levels from Investing.com, and the merged daily data are released with the reproducible code so that every result can be reproduced. Descriptive statistics are summarized in Table 1.

3.3. Notation and Conditional Moments

Let the return on day t be modelled given the information available at the previous close. Each method reports a symmetric interval centered on a forecast of the conditional mean, with a half-width equal to a multiplier times the square root of a forecast of the conditional variance. The conditional mean is estimated by the expanding sample mean, which is appropriate given the weak predictability of daily returns noted above. The methods differ in how they estimate the conditional variance and how they choose the multiplier. To avoid clutter, the variance is written as h and the mean as mu in Equation 1 below.

$$C_t = \left[\mu - m_t\sqrt{h_t},\ \ \mu + m_t\sqrt{h_t}\right] \tag{1}$$

Here, C is the prediction interval for the return on day t, mu is the conditional mean, the square root of h is the conditional volatility in percent, and m is a method-specific multiplier (dimensionless). Each symbol is defined where it first appears.

3.4. Volatility Scales

Two volatility forecasts are used. The first is an exponentially weighted moving average of squared deviations, with smoothing parameter lambda equal to 0.94, computed causally so that the variance for day t uses only returns observed before day t (Equation 2):

$$h_t = \lambda\, h_{t-1} + (1-\lambda)\,(r_{t-1} - \mu)^2 \tag{2}$$

Here, h is the one-step variance forecast for day t in percent squared, lambda equal to 0.94 is the RiskMetrics decay, mu is the conditional mean, and the squared deviation uses the previous return. This exponentially weighted scale is the volatility estimate used by the two conformal methods.

The second is a generalized autoregressive conditional heteroskedasticity model of order one in both terms, GARCH(1,1), with Student t innovations, whose conditional variance follows Equation 3:

$$h_t = \omega + a\,(r_{t-1} - \mu)^2 + b\,h_{t-1} \tag{3}$$

where omega (percent squared) is the constant, a is the ARCH coefficient on the most recent squared shock and b is the coefficient on the past variance, both dimensionless, with omega greater than zero, a and b non-negative, and a plus b below one for stationarity. The parameters omega, a, b, and the degrees of freedom nu are estimated by quasi-maximum likelihood. To keep the procedure online, the GARCH model is refit on an expanding window every 25 trading days, and between refits, the variance recursion is propagated one step at a time using the most recent parameters, so no future information enters any forecast. This GARCH variance forecast drives the GARCH-N and GARCH-t intervals and also defines the volatility terciles used later to classify regimes. Parametric volatility models of this type remain strong benchmarks for interval and value at risk forecasting [41], [35].

GARCH(1,1) is chosen because it is the standard and most parsimonious model for daily equity volatility. As a generalization of the ARCH model by Bollerslev [43], it captures the volatility clustering and persistence of daily returns with only three variance parameters. This parsimony is well supported empirically: in a comparison of 330 volatility specifications, Hansen and Lunde [44] found that no other methods can forecast the conditional variance of a broad market series significantly better than a GARCH(1,1), since the extra parameters of richer models add estimation error that outweighs their in-sample flexibility, although models with a leverage term can help for individual stocks. Student t innovations are used on the same grounds: a heavy-tailed error distribution markedly improves GARCH(1,1) variance forecasts for a major equity index, whereas allowing for skewness or time-varying higher moments does not [45]. Asymmetric and higher-order variants may refine the fit for some assets and are a natural extension, but a parsimonious GARCH(1,1) with Student-t errors is the transparent benchmark appropriate to a study focused on calibration rather than volatility model selection.

3.5. Interval Constructions

We compare five methods. The unconditional benchmark, UNC, takes the expanding sample mean and standard deviation and sets the multiplier to the standard normal quantile, which gives an interval of constant width. The two parametric volatility intervals are built on the GARCH variance forecast. GARCH-N applies the normal quantile to the GARCH standard deviation, while GARCH-t applies the scaled Student t quantile with the estimated degrees of freedom, so its interval allows for heavier tails. The volatility-scaled split conformal predictor, SCP, follows the normalized conformal idea [46], [17]. It builds a nonconformity score as the absolute deviation of the return from the conditional mean, divided by the exponentially weighted volatility, as defined in Equation 4,

$$s_i = \frac{|r_i - \mu|}{\sqrt{h_i}} \tag{4}$$

where the score s, which is dimensionless, measures how far a calibration return sits from the mean in units of its own exponentially weighted volatility. The method computes the conformal quantile q, also dimensionless, of these scores on a rolling calibration window of the most recent 500 points with the usual finite-sample correction and reports the conditional mean plus or minus q times the exponentially weighted volatility. The score and the quantile q are used by both the split conformal and the adaptive methods. The adaptive method, denoted ACI, implements adaptive conformal inference [22], [23]: it maintains a time-varying level, forms the interval from the corresponding quantile of the scaled scores, and updates the level after each observation according to Equation 5,

$$\alpha_{t+1} = \alpha_t + \gamma\,(\alpha - e_t), \qquad e_t = 1\{r_t \notin C_t\} \tag{5}$$

where the adaptive level on day t, the step size gamma equal to 0.01, and the violation indicator are all dimensionless, and the indicator equals one when the realized return falls outside the interval and zero otherwise. Only the adaptive method uses this online level update, which lets it defend its long-run coverage when exchangeability fails [25], [24].

To position the parametric and conformal constructions against the current interval forecasting literature, five machine learning benchmarks are evaluated under the same online protocol and the same forecast origins. Each benchmark predicts the next return from a common causal feature vector built only from past information, namely recent lagged returns, the absolute value of the previous return, rolling standard deviations over twenty and sixty days, a short rolling mean, and the exponentially weighted volatility. Conformalized quantile regression (CQR) pairs gradient-boosted quantile regressors for the lower and upper 5 percent quantiles with a split conformal correction estimated on the most recent calibration window, which restores finite-sample coverage. Its adaptive variant (Adaptive CQR) applies the same online level update used by adaptive conformal inference to the conformity quantile, so the effective level tracks realized coverage. Ensemble batch prediction intervals (EnbPI) build a bootstrap ensemble of regression trees and form intervals from the quantiles of the leave-one-out residuals, refreshed as new observations arrive. Quantile regression forests (QRF) read the lower and upper quantiles directly from the empirical leaf distribution of a random forest, and natural gradient boosting (NGBoost) fits a probabilistic model that returns a predictive mean and variance from which a Gaussian interval is formed. The tree and boosting models are refitted every one hundred test steps on a rolling window of the most recent fifteen hundred observations and used to predict the intervening origins, which keeps the comparison strictly out-of-sample and bounds its cost on the long samples.

The default conformal settings follow common practice in online conformal literature. A calibration window of five hundred observations is long enough to estimate with low variance the 90 percent conformity quantile that a 90 percent interval requires yet short enough to adapt to slow changes in the return distribution, and a learning rate of gamma equal to 0.01 is the small value recommended for adaptive conformal inference so that the level adjusts steadily without overreacting to isolated violations. Because any fixed choice is open to question,

Section 4.6 reports a sensitivity analysis over calibration windows of two hundred fifty, five hundred, and one thousand days and learning rates of 0.001, 0.01, and 0.05.

3.6. Online Evaluation Protocol

For each series, the first 40 percent of observations initialize the models, and intervals are then produced for every subsequent day in strict chronological order, with the estimation window expanding as new data arrives. No observation is used to forecast itself, and the calibration window for the conformal methods always precedes the forecast origin. This walk-forward design avoids the look-ahead leakage that inflates apparent performance in back-tests [47], [48]. Pooling across the five series yields more than twenty-five thousand out-of-sample interval evaluations.

3.7. Evaluation Metrics

Three aspects of interval quality are measured. Coverage is the empirical frequency with which the realized return falls inside the interval, which should equal the nominal 90 percent. Sharpness is the mean interval width, where narrower is better for a given coverage. The two are combined in the interval score, a proper scoring rule that rewards narrow intervals but penalizes a missed observation in proportion to the distance by which it falls outside, as shown in Equation 6 [16]:

$$\mathrm{IS} = (u - l) + \frac{2}{\alpha}(l - y)\,1\{y < l\} + \frac{2}{\alpha}(y - u)\,1\{y > u\} \tag{6}$$

Here, l and u are the lower and upper interval limits, and y is the realized return being scored, all in percent, with y equal to the next day's return; the indicator is one when its condition holds. The interval score, also in percent, is the efficiency number we report in the pooled results and in the coverage and width figure, and a lower score is better. We test calibration formally in two ways: a likelihood ratio test of unconditional coverage, and a Markov-style test of whether violations cluster in time. Both follow standard value at risk back-testing practice [13] and its recent move into conformal prediction [14]. Finally, to surface behavior that marginal numbers hide, we sort each forecast day into a low, medium, or high-volatility regime by terciles of the GARCH variance forecast, and we recompute coverage inside each regime.

To judge calibration formally, coverage is assessed with the likelihood ratio tests of Christoffersen. The unconditional coverage test compares the observed violation rate, where a violation is a day on which the realized return falls outside the interval, against the nominal rate of 0.10 through a likelihood ratio statistic that is asymptotically chi-squared with one degree of freedom, so a large p-value indicates that the empirical coverage is statistically indistinguishable from the ninety percent target. The independence test examines whether violations cluster in time by modelling the violation sequence as a first-order Markov chain and testing the null of serially independent exceedances, again with a chi-squared statistic on one degree of freedom. These two tests, reported as the LR coverage and LR independence p-values, separate methods that merely hit the average rate from methods that also place their exceedances without dangerous clustering.

Because several methods attain similar average scores, pairwise differences in predictive performance are also tested for statistical significance rather than reading from point estimates alone. The Diebold and Mariano test is applied to the per-day loss differential of the interval score between two methods, using a heteroskedasticity-consistent long-run variance, to test the null of equal predictive accuracy. The same paired differences are examined with the non-parametric Wilcoxon signed-rank test, which avoids the normality assumption, and a moving-block bootstrap with a block length of twenty days places a 95 percent confidence interval on the mean score difference while respecting serial dependence. Differences in coverage between two methods on the same forecast origins are assessed with a paired McNemar test on the two-by-two table of joint hit and miss outcomes. These procedures are reported in Section 4.6.

3.8. Implementation and Reproducibility

We run everything in Python with a fixed random seed, and we estimate the GARCH models with a standard volatility modeling library. One notebook downloads the data, fits every model, runs the evaluation, and rebuilds all tables and figures on its own, so anyone can reproduce the whole study from raw data in a single pass.

The volatility models are estimated with the arch package for Python, which fits the GARCH(1,1) specification with Student t innovations by maximum likelihood, so the parametric intervals use the same estimator throughout. The conditional mean is a constant, the variance follows the standard GARCH(1,1) recursion, and the model is refitted every twenty-five steps on the expanding window with its variance recursion propagated between refits. The order (1,1) is adopted because it is the conventional and parsimonious choice for daily equity volatility and because the study targets calibration under a fixed and transparent specification rather than model selection; richer orders and asymmetric variants are a natural extension. Quantiles and the coverage tests use SciPy, and the machine learning benchmarks use scikit-learn for gradient boosting, the quantile-forest package for quantile regression forests, and the NGBoost package for natural gradient boosting. Rather than tuning the conformal hyperparameters to the data, their sensitivity is examined directly in Section 4.6.

4. Results and Discussion

4.1. Return Characteristics

Table 1 summarizes the five series. Each one has a small daily mean relative to its standard deviation, which confirms the weak signal in the conditional mean. Every series has heavy tails, with excess kurtosis between about 6 and 11, and every series is negatively skewed, so large negative returns arrive more often than a Gaussian model expects. The fattest-tailed series is the S&P 500, with excess kurtosis near 11, and the largest single-day losses, close to thirteen percent, occur in the DAX and the CAC during the 2008 and 2020 crises. These features, heavy tails and dispersion that move over time, are exactly where a constant-width Gaussian interval should struggle, and where volatility scaling and distribution-free calibration should help.

Table 1. Descriptive statistics of daily percentage returns for the five series.

Series	n	Mean	Std	Skewness	Exc. kurtosis	Min	Max
DAX	9,068	0.023	1.365	−0.206	5.954	−13.055	10.797
SMI	8,730	0.022	1.087	−0.336	7.511	−10.134	10.788
CAC	9,093	0.013	1.328	−0.187	5.974	−13.098	10.595
FTSE	6,311	0.007	1.123	−0.387	9.060	−11.512	9.384
S&P 500	8,969	0.035	1.132	−0.358	11.101	−12.766	10.964

4.2. Marginal Coverage and Efficiency

Pooled across all five markets and more than twenty-five thousand forecasts, the methods separate clearly on calibration, even though their average coverage is similar. Table 2 reports pooled coverage, mean width, the interval score, and the p-value of the unconditional coverage test, and Figure 2 plots coverage against the 90 percent target. The unconditional benchmark overcovers on average, at 0.935, and its coverage test rejects the nominal hypothesis, a first sign that a constant width does not fit a market whose dispersion moves over time. The Gaussian GARCH interval is close to nominal at 0.907, but its coverage test rejects the nominal hypothesis decisively, with a p-value below 0.001; with more than twenty-five thousand evaluations, a deviation of less than one percentage point is small in magnitude yet still statistically clear. The Student t GARCH interval is the only parametric method that passes the coverage test, at 0.902 with a p-value of 0.379, and it is also the most efficient method in the study, with the lowest interval score of 4.653 and the smallest mean width among the calibrated

methods, 3.426. The two conformal methods are the best calibrated marginally: split conformal reaches 0.900 with a p value of 0.849, and adaptive conformal inference reaches 0.900 with a p value of 0.985, the closest to nominal of all, at the cost of slightly wider intervals of 3.53 and 3.61, respectively.

Table 2. Pooled out-of-sample performance across all five markets. The coverage target is 0.90.

Method	Coverage	Mean width	Interval score	LR coverage p
UNC	0.935	4.130	5.580	<0.001
GARCH-N	0.907	3.495	4.658	<0.001
GARCH-t	0.902	3.426	4.653	0.379
SCP	0.900	3.533	4.722	0.849
ACI	0.900	3.613	4.770	0.985

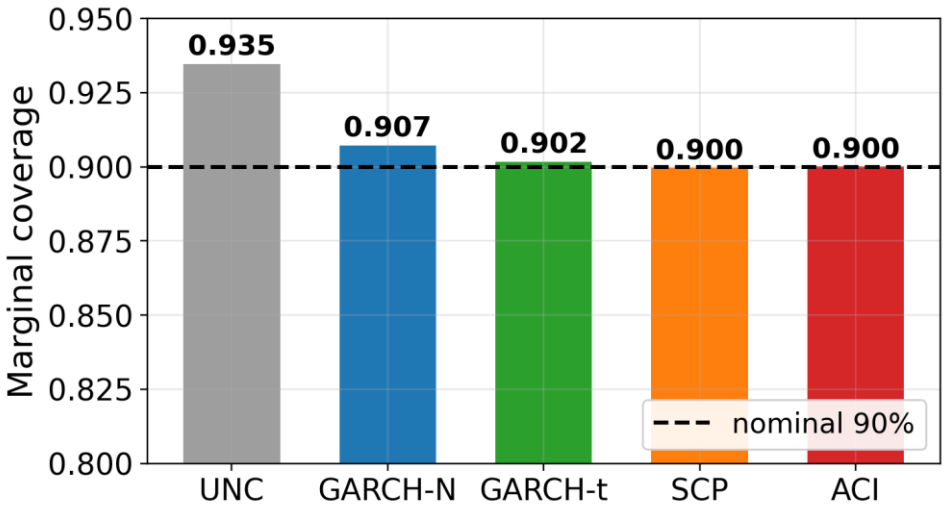


Figure 2. Pooled marginal coverage of each method against the 90 percent target. Bars near the dashed line are well calibrated. The unconditional interval sits well above the target, while the Student t GARCH interval and the two conformal methods lie closest to it.

4.3. Consistency Across Markets

Table 3 shows that a single index does not drive these patterns. The unconditional benchmark runs from 0.915 on the United States series to 0.954 on the FTSE, always at or above the target, which reflects a constant width that a long history of crises has inflated. The conformal methods stay within about one percentage point of the target in every market, and the two GARCH intervals are almost stable. The point is that a healthy marginal number, whether close to the target or comfortably above it, hides how coverage is distributed across calm and turbulent days. Only a regime-aware analysis exposes that distribution.

Table 3. Marginal coverage by market. Values close to 0.90 indicate accurate calibration.

Method	DAX	SMI	CAC	FTSE	S&P 500
UNC	0.939	0.939	0.932	0.954	0.915
GARCH-N	0.900	0.911	0.908	0.911	0.908
GARCH-t	0.894	0.906	0.905	0.907	0.898
SCP	0.900	0.899	0.900	0.900	0.899
ACI	0.900	0.900	0.900	0.900	0.900

4.4. Coverage Across Volatility Regimes

The central result of the study appears in Table 4 and Figure 3, which breaks coverage down by volatility regime. The unconditional benchmark is badly miscalibrated conditionally, even though it looked acceptable, indeed generous, marginally: it covers 0.987 of low-volatility days but only 0.847 of high-volatility days, a regime gap of about 14 percentage points, and a serious shortfall precisely when risk is greatest. Conditioning on the GARCH variance forecast removes this gap almost entirely, since the Gaussian GARCH interval covers 0.905, 0.910, and 0.906 across the low, medium, and high regimes. The conformal methods go one step further in the turbulent regime: split conformal and adaptive conformal inference reach 0.912 and 0.920 coverage on high-volatility days, the best of all methods, while accepting mild undercoverage on calm days at 0.884 and 0.878. In effect, the distribution-free methods reallocate coverage from quiet periods, where a small shortfall is harmless, toward turbulent periods, where coverage matters most for risk control. Figure 4 shows how this works on a turbulent stretch of the DAX. The adaptive band widens as volatility climbs, while the Gaussian GARCH band, whose width follows the same variance forecast but whose Gaussian tail shape is fixed, is breached by a cluster of large moves.

Table 4. Coverage within each volatility regime, set by terciles of the variance forecast.

Method	Low volatility	Medium volatility	High volatility
UNC	0.987	0.970	0.847
GARCH-N	0.905	0.910	0.906
GARCH-t	0.900	0.905	0.900
SCP	0.884	0.903	0.912
ACI	0.878	0.902	0.920

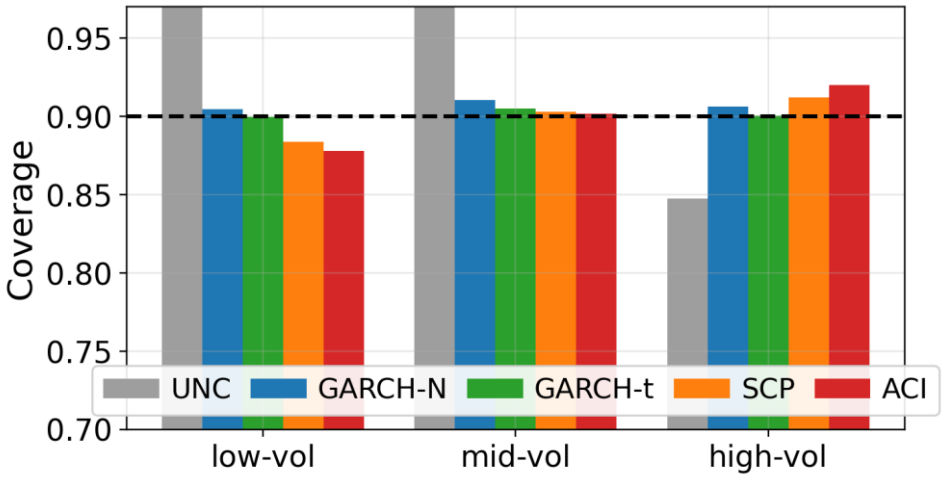


Figure 3. Coverage within the low, medium, and high-volatility regimes. The unconditional interval collapses when volatility is high, while the conformal methods cover best there.

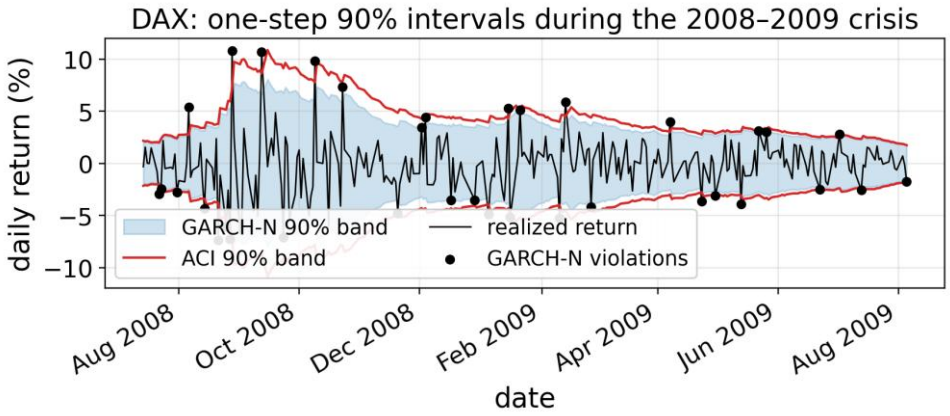


Figure 4. One-step-ahead interval over a turbulent stretch of the DAX. The adaptive band widens with volatility, and black markers flag realized returns that fall outside the GARCH-N band.

The three regimes are defined by the terciles of the one-step-ahead variance forecast, which split the forecast origins into equally sized low, medium, and high volatility groups. Equal-sized bins keep the coverage estimates comparable in precision across regimes and avoid imposing external thresholds. To confirm that the conclusions do not depend on this choice, the same analysis is repeated with quartiles and quintiles in Section 4.6, and the ordering of the methods is unchanged. A fully model-based regime classifier, such as a Markov switching volatility model, is a promising alternative that we leave to future work.

4.5. The Coverage and Width Trade-off

Figure 5 places each method on a plane of pooled coverage against mean width. The unconditional benchmark sits to the right of the target line and overcovers at the largest width, so it is dominated. The two GARCH intervals sit close to the target line at the smallest widths among the calibrated methods, the Student t version being the most efficient of all, while the two conformal methods sit on the line, at a width premium of only about one to three percent. No method wins on every count. The honest summary is simple: efficiency favors the parametric GARCH intervals, while exact marginal calibration and turbulent regime coverage favor the conformal methods, which match the GARCH calibration without any distributional assumption.

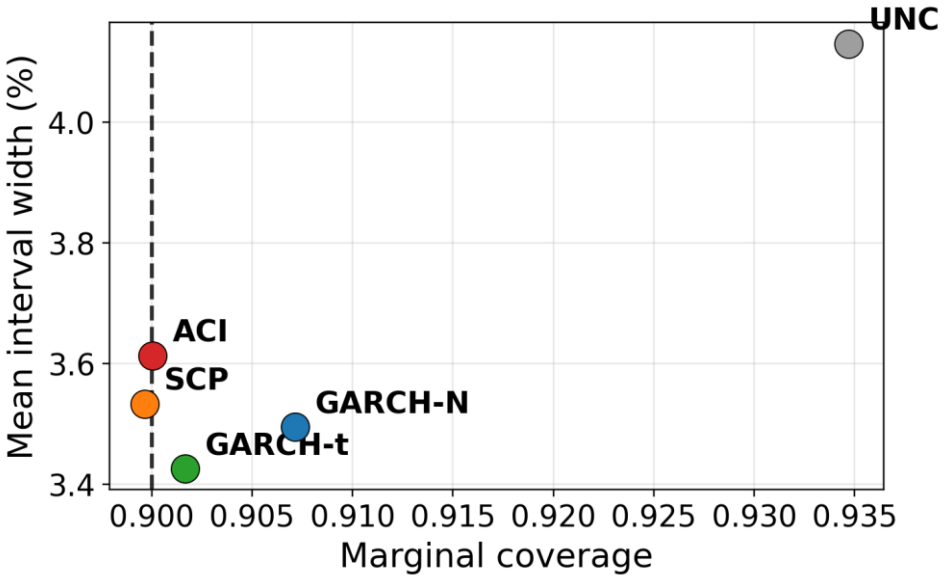


Figure 5. Pooled coverage against the mean interval width. Methods on the vertical target line are well calibrated; points to its left indicate undercoverage. The Student t GARCH interval is the most efficient method, and the conformal methods lie closest to the target line.

4.6. Extended Benchmarks, Sensitivity, and Statistical Significance

This subsection broadens the comparison and subjects it to formal tests. Table 5 reports the pooled evaluation of the five machine learning benchmarks alongside the five interval constructions introduced earlier. The machine learning intervals miscalibrate the heavy tails of daily returns: conformalized quantile regression reaches 0.885 coverage, quantile regression forests 0.886, and natural gradient boosting only 0.862, each rejected by the unconditional coverage test, while ensemble batch prediction intervals overcover at 0.914, so none matches the nominal target. The adaptive variant of conformalized quantile regression recovers to 0.898 coverage with a coverage test p-value of 0.241, which is the key comparison: the same quantile regression base learner passes the coverage test only once the online adaptive level update is added. This shows that the calibration gain comes from the adaptive conformal mechanism rather than from the choice of base learner, and it places the volatility-scaled split conformal and adaptive conformal methods, both at 0.900 coverage, among the best calibrated of all ten methods.

Table 5. Extended pooled comparison including five machine learning benchmarks. The coverage target is 0.90; LR cov. is the p value of the unconditional coverage test.

Method	Coverage	Mean width	Interval score	LR cov.
UNC	0.935	4.130	5.580	<0.001
GARCH-N	0.907	3.495	4.658	<0.001
GARCH-t	0.902	3.426	4.653	0.379
SCP	0.900	3.533	4.722	0.849
ACI	0.900	3.613	4.770	0.985
CQR	0.885	3.192	4.874	<0.001
ACQR	0.898	3.444	4.735	0.241
EnbPI	0.914	3.796	5.654	<0.001
QRF	0.886	3.460	5.226	<0.001
NGBoost	0.862	2.986	5.054	<0.001

Table 6 tests whether the differences from adaptive conformal inference are statistically significant. Against the four machine learning benchmarks the differences are large and highly

significant on every test, with Diebold and Mariano p values of 0.001 or below and bootstrap confidence intervals for the score difference that exclude zero, so adaptive conformal inference is a significantly better interval forecaster than off-the-shelf machine learning here. The comparison with the parametric and conformal methods is reported honestly: the Gaussian and Student t GARCH intervals and split conformal carry a small but significant efficiency edge over adaptive conformal inference, with score differences of about 0.11, 0.12, and 0.05 in their favor, while the evidence on the adaptive conformalized quantile method is mixed: the Diebold and Mariano test does not reject equal predictive accuracy ($p = 0.162$) and the bootstrap interval straddles zero, yet the Wilcoxon signed-rank test does reject ($p < 0.001$), which reflects a small difference that is consistent in sign rather than large in magnitude. No single method dominates; the value of the adaptive conformal layer is its reliable coverage, in particular on turbulent days, at a modest and quantified cost in width.

Table 6. Statistical significance of the difference between adaptive conformal inference and each competing method, pooled across markets. A positive score difference favors adaptive conformal inference; the interval is a moving-block bootstrap 95 percent confidence interval; the DM, Wilcoxon, and McNemar columns report p values.

Comparison	Score diff.	95% CI	DM p	Wilcoxon p	McNemar p
ACI vs UNC	+0.810	[+0.681, +0.953]	<0.001	<0.001	<0.001
ACI vs GARCH-N	-0.112	[-0.136, -0.089]	<0.001	<0.001	<0.001
ACI vs GARCH-t	-0.117	[-0.142, -0.092]	<0.001	<0.001	0.130
ACI vs SCP	-0.048	[-0.064, -0.033]	<0.001	<0.001	0.635
ACI vs CQR	+0.105	[+0.009, +0.214]	0.001	<0.001	<0.001
ACI vs ACQR	-0.035	[-0.100, +0.042]	0.162	<0.001	0.132
ACI vs EnbPI	+0.884	[+0.719, +1.070]	<0.001	<0.001	<0.001
ACI vs QRF	+0.456	[+0.376, +0.538]	<0.001	<0.001	<0.001
ACI vs NGBoost	+0.284	[+0.167, +0.412]	<0.001	<0.001	<0.001

Table 7 examines the sensitivity of adaptive conformal inference to its two design choices. Across all nine combinations of the calibration window and the learning rate, pooled coverage stays between 0.899 and 0.900, so coverage is essentially insensitive to these settings. Efficiency responds in the expected direction: a larger learning rate widens the intervals and raises the interval score, while a longer calibration window is marginally more efficient. The default of a five-hundred-day window and a learning rate of 0.01 sits in the stable interior of this grid, which supports the choice made in the main analysis.

Table 7. Sensitivity of adaptive conformal inference to the calibration window and the learning rate, pooled across markets. The coverage target is 0.90.

Window	Learning rate	Coverage	Mean width	Interval score
250	0.001	0.899	3.556	4.740
250	0.01	0.900	3.670	4.796
250	0.05	0.899	3.822	4.913
500	0.001	0.900	3.546	4.727
500	0.01	0.900	3.613	4.770
500	0.05	0.899	3.782	4.888
1000	0.001	0.900	3.533	4.718
1000	0.01	0.900	3.608	4.759
1000	0.05	0.899	3.786	4.894

Finally, Table 8 repeats the regime analysis with quartiles instead of terciles. The ranking is unchanged: the unconditional and machine learning intervals lose coverage as volatility rises, falling to between 0.78 and 0.83 in the most turbulent quartile, whereas the split conformal and adaptive conformal methods gain coverage there, reaching 0.915 and 0.926. A quintile split gives the same picture, with adaptive conformal inference covering 0.928 of days in the most turbulent fifth. The regime conclusions are therefore robust to the number of bins.

Table 8. Coverage within volatility quartiles of the variance forecast, pooled across markets. Q1 is the calmest quartile and Q4 the most turbulent. The coverage target is 0.90.

Method	Q1	Q2	Q3	Q4
UNC	0.988	0.980	0.951	0.820
GARCH-N	0.903	0.910	0.907	0.908
GARCH-t	0.898	0.905	0.903	0.901
SCP	0.881	0.899	0.903	0.915
ACI	0.874	0.895	0.905	0.926
CQR	0.919	0.906	0.885	0.830
ACQR	0.901	0.904	0.896	0.889
EnbPI	0.980	0.968	0.930	0.778
QRF	0.954	0.926	0.877	0.787
NGBoost	0.898	0.882	0.869	0.800

4.7. Interpretation

The results reconcile two strands of literature. On one hand, they confirm that a marginal coverage number can be dangerously misleading for non-stationary returns, which is the core warning of the value at risk backtesting tradition [13], [14] and of work on distribution shift in time series [6]. The unconditional interval passes a casual marginal check yet fails badly when volatility is high. On the other hand, they support the view that well-specified parametric volatility models are hard to beat on efficiency [41], since the Student t GARCH interval is the most efficient method here and is well calibrated both marginally and across regimes, with the Gaussian version close behind on efficiency and well calibrated across regimes. The contribution of distribution-free calibration is therefore specific rather than universal. Conformal scaling and adaptive conformal inference deliver the most accurate marginal coverage and the best protection in the high-volatility regime, consistent with the design goals of online conformal methods under non-exchangeability [22], [23], [25], and the success of the simple split construction agrees with recent theory on why split conformal stays valid under temporal dependence [26]. The value of scaling the conformal score by a volatility forecast, rather than using raw residuals, matches the normalized conformal literature [46].

Three mechanisms explain these patterns. Adaptive conformal inference is most stable in the high-volatility regime because it does not rely on any distributional form: when a cluster of large moves pushes realized coverage below target, the online update widens the interval within a few days, and when calm returns it narrows again, so the method tracks the persistent swings in dispersion that define volatility clustering. Estimated on more than three decades of data, the Gaussian and Student t GARCH models are themselves well calibrated across regimes, which confirms that a well-specified volatility model already captures most of the conditional structure; the distribution-free layer then matches that calibration without any parametric assumption and adds a further gain on the most turbulent days. The unconditional benchmark, by contrast, fails conditionally: volatility clustering makes large and small moves arrive in runs, and a constant width interval is at once too wide in calm runs, where it overcovers, and too narrow in turbulent runs, where it covers only 0.847 of days, which is exactly the regime dependent miscoverage seen in Table 4. The machine learning benchmarks miscalibrate for a related reason: a single global quantile or variance estimate, however flexible the mean model, cannot track the fast changes in tail risk, so they undercover or overcover exactly when dispersion is shifting.

4.8. Practical Implications

When average interval quality and efficiency matter most, a Student t GARCH interval is an excellent and cheap default. When coverage in turbulent periods matters most, the small width premium of a conformal or adaptive conformal interval buys clearly better high-volatility coverage and exact marginal calibration. For capital adequacy and limit setting, that exact calibration is often the more important property [42]. The conformal layer wraps any volatility forecast, so the two ideas work together rather than against each other. A practitioner can keep a trusted GARCH or exponentially weighted volatility model and bolt on a conformal calibration step that defends coverage when the distributional assumptions break. This pairing also matters for portfolio choices under uncertainty, where the reliability of the reported interval feeds straight into position sizing [49].

These calibration gains translate directly into risk management practice. The lower endpoint of a one-step 90 percent interval is, up to sign, a one day 95 percent value at risk estimate, so a method that holds its conditional coverage on turbulent days produces value at risk numbers that are neither systematically breached when risk is high nor needlessly conservative when it is low. At the portfolio level, better calibrated interval widths feed more accurate volatility and tail inputs for risk aggregation and for position sizing, where exposure is scaled inversely to forecast risk, so that leverage is trimmed precisely when the adaptive interval widens. The same intervals inform capital reserve and margin estimation, where both undercoverage, which leaves losses uncovered, and chronic overcoverage, which locks up capital, are penalized, so a method that reallocates coverage toward high-volatility days eases both concerns at once. Because the conformal layer wraps any base forecaster, these benefits can be added to an existing GARCH or machine learning pipeline without replacing it.

4.9. Contribution

Earlier conformal work in finance and energy mostly chases point accuracy or reports only marginal coverage [27], [28], [29], [30]. We do something different: we isolate the calibration layer and study it through a regime-conditional lens across several markets. The wider literature shares the same gap: adaptive conformal methods with data-dependent weights give valid, well calibrated intervals on scientific and engineering data [21], and careful comparisons of classical and deep models forecast daily returns under strict leakage controls [5], yet the former is never tested for coverage inside a high-volatility regime on equity returns, and the latter returns point forecasts with no finite-sample guarantee that any interval will contain the next return. That is the case our method handles and earlier ones cannot: a distribution-free interval whose coverage holds within the turbulent regime, recovered from a light volatility-scaled calibration step rather than a new forecasting model. Next to the decomposition and deep learning literature for stock and volatility forecasting [33], [34], [35], our results show that a light volatility scaled conformal step recovers much of the practical gain in interval reliability, without a heavy model. The payoff is concrete. Better calibrated intervals mean more trustworthy risk reporting for the pension funds, insurers, and asset managers that depend on daily value at risk style measures.

The contribution is concrete on three fronts at once: a like for like regime-conditional benchmark, a significance tested comparison against five modern machine learning interval methods, and a sensitivity analysis, which together turn a single empirical comparison into evidence about when the adaptive conformal layer is worth its cost.

4.10. Validity and Reliability

A few features make the findings credible. The evaluation runs in strict chronological order and avoids look-ahead leakage [47]. The conclusions hold across four European indices and a separately sourced United States index. The calibration claims rest on formal coverage tests, not on visual inspection alone. The whole pipeline is deterministic and reproducible from public data, so anyone can check the numbers. One caveat is worth stating plainly. The conditional analysis builds its volatility terciles from the same forecast that some methods use,

so the regime comparison is internal and descriptive rather than a formal independence test. We treat that as a deliberate design choice.

4.11. Limitations

We kept the study narrow on purpose. It covers only one-step-ahead intervals, so it does not examine multi-horizon behavior, where conformal calibration is known to be trickier. The intervals are symmetric, which may not suit skewed return distributions, and we use only one family of nonconformity scores. The benchmark also stops short of an economic objective, such as trading profit net of transaction costs or expected shortfall, which the literature flags as an open and important gap. These limits keep the comparison clean and set a clear agenda for future work.

Several limitations qualify these findings and set an agenda for future work. The empirical base is five liquid indices from developed markets, so the results may differ for emerging market indices, individual stocks, cryptocurrencies, and commodity or foreign exchange series, where liquidity and tail behavior are different; extending the same protocol to these assets is a natural next step. The parametric benchmarks rest on a single GARCH(1,1) specification, and asymmetric or higher-order volatility models could sharpen the parametric intervals. The conformal guarantees assume approximate exchangeability of the conformity scores, which holds only approximately under the serial dependence of returns, so the finite-sample coverage reported here is empirical rather than exactly guaranteed. The evaluation is also statistical rather than economic: it does not measure trading profit net of costs or run a full value at risk backtest under a regulatory capital framework, which would translate the calibration gains into monetary terms.

5. Conclusions and Suggestions

We set out to learn whether distribution-free conformal prediction can deliver dependable one-step-ahead intervals for daily equity returns, and how it stacks up against parametric volatility intervals across volatility regimes. We tested five interval constructions, together with five machine learning interval methods, on four European indices and the S&P 500 from 1990 through 2026 under one strictly chronological online protocol, and three findings stand out. The first is that marginal coverage alone is not enough. The unconditional Gaussian interval looked generous on average yet covered only about 85 percent of high-volatility days against a 90 percent target, a regime gap of roughly 14 percentage points. The second is that conditioning on a volatility forecast fixes the problem. The Student t GARCH interval was the most efficient method and the only parametric one to pass the marginal coverage test, and both GARCH intervals were well calibrated across regimes, which confirms that well-specified parametric models are strong benchmarks. The third is that distribution-free calibration matches that calibration without any distributional assumption and adds a further gain where it matters most. Volatility scaled split conformal and adaptive conformal inference reached the most accurate marginal coverage and the highest coverage in the high-volatility regime, moving protection toward the periods that matter most, at a width premium of only about one to three percent. They also outperformed four of the five machine learning benchmarks, whose intervals do not cope well with heavy tails. The fifth, adaptive conformalized quantile regression, stayed close to adaptive conformal inference (Diebold and Mariano $p = 0.162$), which is unsurprising because it is the only benchmark that also updates its level online.

The practical advice is straightforward. Analysts who care most about efficiency can use a Student t GARCH interval with confidence. Analysts who care most about coverage in turbulent periods can add a light conformal calibration step on top of the same volatility model, which defends coverage when distributional assumptions fail. The conformal layer is model agnostic, so the two approaches combine rather than compete. The limitations point to several extensions. Multi-step and multi-horizon intervals are worth studying, since calibration gets harder as the horizon grows. Asymmetric and multi-score conformal constructions could fit skewed returns better, and semantic or weighted conformal schemes may adapt more quickly to

changing regimes [28]. Change-point detection through conformal test martingales could trigger recalibration when the data-generating process shifts [18], [50], and closer ties to adaptive normalization or differencing could steady the inputs further [7], [8], [10]. Evaluating intervals through an economic lens, with expected shortfall backtests and transaction-cost-aware objectives [42], would link calibration quality to decision value. In the end, this benchmark shows when a distribution-free calibration layer earns its modest width penalty, and it gives a reproducible foundation for that judgment in non-stationary equity markets.

Acknowledgments

All datasets used in this study are publicly available: the DAX, SMI, and CAC 40 index levels are from Yahoo Finance, and the FTSE 100 and S&P 500 index levels are from Investing.com. The merged daily data and the complete analysis run from a single reproducible notebook that is openly available at https://doi.org/10.5281/zenodo.20952837. During preparation, the author used a generative artificial intelligence assistant to help draft and edit the text and to scaffold the analysis code; the author verified the study design, all computations, the results, and their interpretation, and takes full responsibility for the content.

Author contributions. Following the CRediT taxonomy: Fathimah Al-Mashumah contributed to conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing of the original draft; Yasi Dani contributed to conceptualization, methodology, validation, supervision, and writing through review and editing. Both authors read and approved the final manuscript.

References

[1] N. Cakici, C. Fieberg, D. Metko, and A. Zaremba, "Predicting returns with machine learning across horizons, firm size, and time," *Journal of Financial Data Science*, vol. 5, no. 4, pp. 119–144, 2023, doi: 10.3905/jfds.2023.1.139.

[2] W. Drobetz and T. Otto, "Empirical asset pricing via machine learning: evidence from the European stock market," *Journal of Asset Management*, vol. 22, no. 7, pp. 507–538, 2021, doi: 10.1057/s41260-021-00237-x.

[3] A. Marsi, "Predicting European stock returns using machine learning," *SN Business & Economics*, vol. 3, no. 7, 2023, doi: 10.1007/s43546-023-00487-4.

[4] J. Shi and H. Yin, "Machine learning in multifactor modeling: a critical assessment," *Journal of Beta Investment Strategies*, vol. 17, no. 1, pp. 7–19, 2026, doi: 10.3905/jbis.2026.1.105.

[5] T. Liu, "A comparative study of transformer-based and classical models for financial time-series forecasting," *Journal of Risk and Financial Management*, vol. 19, no. 3, 2026, doi: 10.3390/jrfm19030203.

[6] T. Cipra, *Time Series in Economics and Finance*. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-46347-2.

[7] T. Kim, J. Kim, Y. Tae, C. Park, J. H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[8] S. Maitra, "Adaptive fractional differencing: theory and methodology," *IEEE Access*, vol. 13, pp. 209,312–209,327, 2025, doi: 10.1109/ACCESS.2025.3633728.

[9] S. Maitra, "Adaptive fractional differencing for learning-optimal long-memory preservation," *IEEE Access*, vol. 14, pp. 32,407–32,430, 2026, doi: 10.1109/ACCESS.2026.3666622.

[10] J. Wan, N. Xia, Y. Yin, X. Pan, J. Hu, and J. Yi, "TCDformer: a transformer framework for non-stationary time series forecasting based on trend and change-point detection," *Neural Networks*, vol. 173, 2024, doi: 10.1016/j.neunet.2024.106196.

[11] H. Fischer, Á. Blanco-Fernández, and P. Winker, "Predicting stock return volatility: can we benefit from regression models for return intervals?," *Journal of Forecasting*, vol. 35, no. 2, pp. 113–146, 2016, doi: 10.1002/for.2371.

[12] L. Hu and Y. Shen, "A predictive analytics approach for forecasting global stock index returns using deep learning techniques," *Decision Analytics Journal*, vol. 18, 2026, doi: 10.1016/j.dajour.2026.100685.

[13] B. Candelon, G. Colletaz, C. Hurlin, and S. Tokpavi, "Backtesting value-at-risk: a GMM duration-based test," *Journal of Financial Econometrics*, vol. 9, no. 2, pp. 314–343, 2011, doi: 10.1093/jjfinec/nbq025.

[14] K. Retzlaff, F. Schlembach, D. Bams, and P. Dreesen, "Testing marginal and conditional coverage in conformal prediction for non-stationary time series via value-at-risk backtesting," *Proceedings of Machine Learning Research*, vol. 266, pp. 725–747, 2025.

[15] T. Kaiser and P. Herzog, "A tutorial on distribution-free uncertainty quantification using conformal prediction," *Advances in Methods and Practices in Psychological Science*, vol. 8, no. 4, 2025, doi: 10.1177/25152459251380452.

[16] N. Dewolf, B. De Baets, and W. Waegeman, "Valid prediction intervals for regression problems," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 577–613, 2023, doi: 10.1007/s10462-022-10178-5.

[17] U. Johansson, H. Boström, T. Löfström, and H. Linusson, "Regression conformal prediction with random forests," *Machine Learning*, vol. 97, no. 1–2, pp. 155–176, 2014, doi: 10.1007/s10994-014-5453-0.

[18] V. Vovk, I. Petej, I. Nouretdinov, V. Manokhin, and A. Gammerman, "Computationally efficient versions of conformal predictive distributions," *Neurocomputing*, vol. 397, pp. 292–308, 2020, doi: 10.1016/j.neucom.2019.10.110.

[19] I. I. Akpabio and S. A. Savari, "On uses of noise analysis for the uncertainty quantification of line edge roughness estimation," *IEEE Transactions on Semiconductor Manufacturing*, vol. 36, no. 3, pp. 319–326, 2023, doi: 10.1109/TSM.2023.3270230.

[20] M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: a unified review of theory and new challenges," *Bernoulli*, vol. 29, no. 1, pp. 1–23, 2023, doi: 10.3150/21-BEJ1447.

[21] P. Hajibabaee, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "Adaptive conformal prediction intervals using data-dependent weights with application to seismic response prediction," *IEEE Access*, vol. 12, pp. 53,579–53,597, 2024, doi: 10.1109/ACCESS.2024.3387858.

[22] I. Gibbs and E. Candès, "Conformal inference for online prediction with arbitrary distribution shifts," *Journal of Machine Learning Research*, vol. 25, 2024.

[23] M. Zaffran, O. Féron, Y. Goude, J. Josse, and A. Dieuleveut, "Adaptive conformal predictions for time series," *Proceedings of Machine Learning Research*, vol. 162, pp. 25,834–25,866, 2022.

[24] A. Makhin and M. Panov, "Automatic adaptive conformal inference for time series forecasting," *Lecture Notes in Computer Science*, vol. 15419, pp. 262–274, 2025, doi: 10.1007/978-3-031-88036-0_16.

[25] A. Farinhas, C. Zerva, D. Ulmer, and A. F. T. Martins, "Non-exchangeable conformal risk control," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[26] R. F. Barber and A. Pananjady, "Predictive inference for time series: why is split conformal effective despite temporal dependence?," *Proceedings of Machine Learning Research*, vol. 313, 2026.

[27] P. L. Seabe, C. R. B. Moutsinga, and M. Aphane, "Uncertainty-first forecasting of the South African equity market using deep learning and temporal conformal prediction," *Big Data and Cognitive Computing*, vol. 10, no. 3, 2026, doi: 10.3390/bdcc10030093.

[28] X. Zhou, B. Chen, Y. Gui, and L. Cheng, "Conformal prediction: a data perspective," *ACM Computing Surveys*, vol. 57, no. 7, 2025, doi: 10.1145/3714380.

[29] L. de Azevedo Takara, R. C. T. de Souza, V. C. Mariani, L. dos S. Coelho, and S. F. Stefenon, "Multi-scale convolutional Chebyshev Kolmogorov-Arnold networks using conformal prediction for photovoltaic power forecasting," *IEEE Access*, vol. 14, pp. 81,170–81,194, 2026, doi: 10.1109/ACCESS.2026.3694780.

[30] S. Ghimire, T. Nguyen-Huy, R. C. Deo, D. Casillas-Pérez, R. Prasad, and S. Salcedo-Sanz, "Point and confidence interval prediction of electricity price based on hybrid dual-decomposition and deep learning model," *Energy*, vol. 350, 2026, doi: 10.1016/j.energy.2026.140692.

[31] W. Sun, F. Tack, L. Clarisse, C. Clerbaux, and M. Van Roozendael, "Inferring surface NO_2 over Western Europe: a machine learning approach with uncertainty quantification," *Journal of Geophysical Research: Atmospheres*, vol. 129, no. 4, 2024, doi: 10.1029/2023JD039820.

[32] L. Li, H. Wang, Y. Liu, and F. Zhang, "Probabilistic interval prediction method based on shape-adaptive quantile regression," *Expert Systems*, vol. 41, no. 8, 2024, doi: 10.1111/exsy.13585.

[33] P. Kumar, P. Das, and E. Ghaderpour, "Decomposed and hybridized deep learning models for an application to stock price prediction," in *Proceedings of the International Conference on Modeling, Simulation & Intelligent Computing (MoSICom)*, 2023, pp. 609–613, doi: 10.1109/MoSICom59118.2023.10458816.

[34] Z. Zhang, Q. Liu, Y. Hu, and H. Liu, "Multi-feature stock price prediction by LSTM networks based on VMD and TMFG," *Journal of Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01127-4.

[35] S. Kumar, A. Rao, and M. Dhochak, "Hybrid ML models for volatility prediction in financial risk management," *International Review of Economics & Finance*, vol. 98, 2025, doi: 10.1016/j.iref.2025.103915.

[36] A. S. Al-Jawarneh, A. R. M. Alsayed, and H. N. Ayyoub, "Optimizing modelling accuracy using variational mode decomposition and elastic net regression: evidence in stock market prediction," *Array*, vol. 28, 2025, doi: 10.1016/j.array.2025.100603.

[37] Y. Wang, X. Hao, and C. Wu, "Forecasting stock returns: a time-dependent weighted least squares approach," *Journal of Financial Markets*, vol. 53, 2021, doi: 10.1016/j.finmar.2020.100568.

[38] Y. Huang and C. Yang, "Enhancing stock price forecasting with a modular deep learning framework incorporating plug-and-play transformer variants," *Expert Systems with Applications*, vol. 315, 2026, doi: 10.1016/j.eswa.2026.131572.

[39] R. A. Andika and C. Dewi, "Importance of feature selection for multiple disease classification," *Jurnal Buana Informatika*, vol. 16, no. 1, pp. 34–45, 2025, doi: 10.24002/jbi.v16i01.11354.

[40] S. Alawiyah, D. Yuliati, and N. Ulinnuha, "Implementasi LightGBM dengan KNN imputation untuk deteksi dini risiko kehamilan," *Jurnal Buana Informatika*, vol. 17, no. 1, pp. 21–32, 2026, doi: 10.24002/jbi.v17i1.14454.

[41] M. Buczyński and M. Chlebus, "Old-fashioned parametric models are still the best: a comparison of value-at-risk approaches in several volatility states," *Journal of Risk Model Validation*, vol. 14, no. 2, pp. 1–20, 2020, doi: 10.21314/JRMV.2020.222.

[42] Z. Du, P. Pei, X. Wang, and T. Yang, "Powerful backtests for historical simulation expected shortfall models," *Journal of Business & Economic Statistics*, vol. 42, no. 3, pp. 864–874, 2024, doi: 10.1080/07350015.2023.2252881.

[43] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986, doi: 10.1016/0304-4076(86)90063-1.

[44] P. R. Hansen and A. Lunde, "A forecast comparison of volatility models: does anything beat a GARCH(1,1)?," *Journal of Applied Econometrics*, vol. 20, no. 7, pp. 873–889, 2005, doi: 10.1002/jae.800.

[45] A. Wilhelmsson, "GARCH forecasting performance under different distribution assumptions," *Journal of Forecasting*, vol. 25, no. 8, pp. 561–578, 2006, doi: 10.1002/for.1009.

[46] U. Johansson, H. Boström, and T. Löfström, "Investigating normalized conformal regressors," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, doi: 10.1109/SSCI50451.2021.9659853.

[47] H. Arian, D. Norouzi Mobarekeh, and L. Seco, "Backtest overfitting in the machine learning era: a comparison of out-of-sample testing methods in a synthetic controlled environment," *Knowledge-Based Systems*, vol. 305, 2024, doi: 10.1016/j.knosys.2024.112477.

[48] C. Spulbăr and C. C. Ene, "Rolling walk-forward LSTM for daily stock-return prediction in European defence markets," *Studies in Business and Economics*, vol. 20, no. 3, pp. 240–260, 2025, doi: 10.2478/sbe-2025-0053.

[49] A. Arasteh, "A data-driven prediction method for multi-period portfolio optimization using the real options approach," *Finance Research Letters*, vol. 80, 2025, doi: 10.1016/j.frl.2025.107403.

[50] I. Nouretdinov, "Conformal test martingales for hypergraphical models," *Annals of Mathematics and Artificial Intelligence*, 2024, doi: 10.1007/s10472-024-09951-9.