


Testing Marginal and Conditional Coverage in Conformal Prediction for Non-Stationary Time Series via Value-at-Risk Backtesting

Konrad Retzlaff¹  RETZLAFF@STUDENT.MAASTRICHTUNIVERSITY.NL

Filip Schlembach¹ 

FILIP.SCHLEMBACH@MAASTRICHTUNIVERSITY.NL

Dennis Bams²

W.BAMS@MAASTRICHTUNIVERSITY.NL

Philippe Dreesen¹ 

PHILIPPE.DREESSEN@MAASTRICHTUNIVERSITY.NL

¹*Department of Advanced Computing Sciences (DACS), Maastricht University, Netherlands*

²*Finance Department, School of Business and Economics (SBE), Maastricht University, Netherlands*

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

Conformal Prediction (CP) constructs prediction intervals with marginal coverage guarantees under the assumption of exchangeability, yet it has also been widely applied to non-exchangeable settings such as time series, where temporal dependence and distribution shifts often violate this assumption. Despite this, CP methods are typically evaluated using descriptive metrics like empirical coverage and average interval width, without formal statistical testing. This lack of hypothesis-driven evaluation makes it unclear whether deviations are meaningful or due to random variation. We address this gap by establishing a formal equivalence between CP and Value at Risk (VaR), enabling the use of VaR-style backtesting methods to statistically assess both marginal and conditional coverage. Additionally, we incorporate Diebold-Mariano tests with interval scores to compare predictive performance. Applied to synthetic, electricity, and financial time series, our framework uncovers violation and adaptation issues overlooked by standard metrics. The Dynamic Binary Test and Geometric Conformal Backtesting, in particular, identify covariate- and drift-induced dependence and miscalibration, offering a sharper lens for evaluating CP methods in non-stationary settings.

Keywords: Conformal prediction, time series, marginal coverage, conditional coverage, coverage testing, Value-at-Risk, backtesting, nonstationarity, predictive intervals, distribution-free inference

1. Introduction

Conformal Prediction (CP) provides a flexible framework for constructing prediction intervals with finite-sample validity, meaning that the prediction interval contains the true outcome with a user-specified probability (e.g., 95%) on average, assuming the data is exchangeable (Vovk et al., 2022). This model-agnostic framework has gained traction in machine learning and statistics due to its ability to generate calibrated uncertainty estimates for any underlying predictive model. However, the core assumption of exchangeability, that the joint distribution of the data is invariant under permutation, is routinely violated in time series settings (Barber et al., 2023). Temporal dependencies such as autocorrelation and non-stationarity violate the assumption of exchangeability, thereby undermining the finite-sample validity guarantees of CP.

In the absence of exchangeability prediction intervals produced by CP can no longer be assumed to meet their nominal coverage levels. Consequently, empirical validation becomes essential. In particular, it is necessary to assess whether the prediction intervals still achieve the desired *marginal coverage* (i.e., whether the overall frequency of hits approximates the nominal level) and *conditional coverage* (typically defined as whether the nominal level holds when conditioned on covariates). Beyond assessing coverage validity, it is also crucial to determine which of two or more CP methods under consideration by a user yields better predictive performance, particularly in terms of efficiency and reliability of their intervals. Although many articles report empirical coverage, they rarely test whether deviations are statistically significant, whether violations are time-dependent, or which model offers statistically superior interval quality with respect to violation rate and interval width, thus providing a potentially misleading sense of reliability (Barber et al., 2023; Gibbs and Candès, 2021; Xu and Xie, 2021).

In this work, we argue first that the evaluation of CP methods in time series should be grounded in a formal hypothesis testing framework; and second, that the statistical tests developed for Value-at-Risk (VaR) backtesting in financial risk management provide a natural and rigorous foundation for such an evaluation. Both CP and VaR produce binary hit/miss sequences, which, under exact calibration, follow an i.i.d. Bernoulli distribution. This structural equivalence enables the use of VaR-style backtests as statistically principled tools to validate CP intervals in sequential settings. To compare the quality of predictions between different CP methods, we use a Diebold-Mariano (DM).

Our contributions are threefold:

1. We demonstrate the theoretical equivalence between CP and VaR under the lens of Bernoulli violation sequences.
2. We present a suite of statistical backtests from the VaR literature to evaluate both marginal and conditional coverage properties of CP. This includes the Diebold-Mariano test, which, when paired with the Interval Score as a proper scoring rule, enables statistically sound comparison of predictive performance across different CP methods.
3. We empirically validate our approach using both synthetic and real-world data, the latter in the form of electricity data and financial time series, highlighting systematic failures that are not detectable through standard empirical coverage metrics alone.

We start in Section 2 by reviewing recent developments in CP for time-series forecasting and highlighting limitations in current evaluation practices. In Section 3, we demonstrate the testing equivalence between Value at Risk (VaR) and CP in time series settings and introduce various categories of evaluation methods. Section 4 presents an example backtesting framework intended to be used by practitioners. Sections 5 describes the experiments, while Section 6 showcases and discusses the experiments conducted using our proposed framework. Finally, Section 7 concludes the paper.

2. Related Work

2.1. Conformal Prediction in Time Series Forecasting

Conformal Prediction (CP) provides predictive intervals with marginal coverage guarantees under two key assumptions. First, the data is assumed to be exchangeable, meaning that the joint distribution is invariant under permutations of the data points. This condition ensures that the rank of the test residual among the calibration residuals is uniformly distributed, which underpins the validity of the coverage guarantee (Vovk et al., 2022). Second, the model-fitting algorithm must be symmetric with respect to the data, which means that the output should be invariant under permutations of the input data. This ensures that exchangeability is preserved through the fitting process (Barber et al., 2023). When both conditions hold, CP guarantees that the predictive interval covers the true response with probability at least $1 - \alpha$, regardless of the underlying data distribution.

While marginal coverage ensures that coverage holds on average across the population, it does not guarantee uniform coverage across different subpopulations or regions of the input space. This shortcoming motivates the notion of *conditional coverage*, requiring that the predictive interval maintains nominal coverage even when conditioned on additional information.

We categorize conditional coverage broadly into two types. The first is *covariate-conditional coverage*, which refers to coverage that remains valid across different values of the model inputs, such as lagged observations. Exact covariate-conditional coverage is known to be unachievable in a distribution-free setting (Barber et al., 2020), but several methods aim for approximate solutions. For example, Colombo (2023) proposes a locally adaptive conformal method that adjusts intervals based on regions of the input space. The second type, which we define, is *conditional coverage on non-input variables*, which includes additional information that affects the predictive distribution but is not used by the model. Accounting for such variables strengthens the notion of conditional validity by requiring robustness beyond the modeled input space.

Marginal coverage is proven under the assumption that the data is exchangeable. Time series data violate this assumption in multiple ways: through autocorrelation, non-stationarity, seasonality, or gradual distributional drift over time (Barber et al., 2023).

To address the loss of exchangeability in time series, several update-based CP methods have been proposed. Adaptive online methods adjust the miscoverage rate α based on recent violations (Gibbs and Candès, 2021; Zaffran et al., 2022), while quantile-tracking approaches update the conformal quantile q via filtering or stochastic approximation (Angelopoulos et al., 2024). Weighted CP reassigns higher weight to recent or relevant observations to better reflect local distribution shifts (Barber et al., 2023). Ensemble methods combine predictions from bootstrapped models over sliding windows to capture temporal uncertainty (Xu and Xie, 2021). While these approaches aim to recover approximate validity under non-stationarity, their theoretical guarantees are limited, and robustness under complex dependence structures remains an open question.

2.2. Evaluation Practices for CP

The evaluation of CP methods is typically based on descriptive performance metrics rather than formal statistical tests. The most common criterion is the *empirical coverage rate*, which measures the proportion of test points where the predicted interval contains the true response. In addition, the *average interval width* is often reported to capture the efficiency of the intervals in regression settings. Together, these metrics are used to describe the overall quality of the intervals (Smirnov, 2023), but they are generally presented in a purely descriptive way, without formal statistical interpretation.

While these descriptive metrics are informative, their interpretation can be sensitive to the size of the test set. For example, observing a 93% empirical coverage with a nominal level of 90% may be entirely reasonable on a test set of ten points, but could reflect a meaningful deviation when based on one million. Recent works, such as Gibbs et al. (2025), provide valuable context by reporting how coverage varies across repeated data splits. However, this focuses on variability across experiments rather than assessing the significance of a single observed result. To strengthen the evaluation of CP methods, we propose complementing descriptive metrics with formal statistical tests that take the test set size into account and help distinguish between meaningful and random deviations from the target coverage.

Another notable limitation is the lack of standard statistical procedures to assess *conditional coverage*. Although many CP methods aim to approximate coverage conditional on features or local regions of the input space, there are no formal hypothesis tests to determine whether such conditional guarantees hold. Instead, evaluations are often limited to global averages (Romano et al., 2019), informal stratification over hand-picked subgroups (Gibbs et al., 2025), or, as proposed by Barber et al. (2023), rolling window coverage plots that assess *dynamic coverage*, a type of the second form of conditional coverage, which reflects time-dependent rather than feature-conditional variation in coverage.

Another open issue is the absence of tools for a statistical comparison of the quality of different CP methods. In point forecasting, it is common to use paired statistical tests such as the Diebold-Mariano tests (Diebold and Mariano, 2002) to compare models based on expected loss. In the CP literature, however, no analogous standard exists for interval forecasts. Comparisons are typically made using informal metrics such as average interval width and empirical coverage, which give no indication about the statistical significance of measured differences.

In this work, we propose a framework that addresses these issues by introducing statistical evaluation tools inspired by the literature on Value-at-Risk (VaR) backtesting.

3. Statistical Equivalence of Conformal Prediction and Value at Risk

Both CP and VaR assess whether an observed outcome falls outside a predicted region, defined by a conformal set or a quantile threshold respectively, thus transforming continuous uncertainty estimates into binary events: a hit, when the observation lies within the region, and a violation (also referred to as a miss), when it falls outside. We use both terms interchangeably throughout this paper. Under exact calibration, each sequence behaves like independent draws from the same Bernoulli distribution with parameter α (Zhang and Nadarajah, 2017; Vovk, 2002). We begin by recalling why VaR violations are Bernoulli distributed, then show the identical structure in CP.

3.1. Value at Risk and its Bernoulli violations

Value at Risk (VaR) is a forecasted lower threshold in financial time series forecasting, such that future observations are not expected to fall below it more than a given α percent of the time. Let y_t denote the realized value at time t , and let $\text{VaR}_t(\alpha)$ denote the predicted threshold. We define the violation indicator as $I_t^{\text{VaR}}(\alpha) = \mathbf{1}\{y_t < \text{VaR}_t(\alpha)\}$, which equals 1 when the observation falls below the VaR threshold at time t , and 0 otherwise. Under exact model specification, this sequence of indicators should satisfy two key properties:

- **Unconditional Coverage (UC)** is expressed as

$$\Pr(I_t^{\text{VaR}}(\alpha) = 1) = \alpha,$$

meaning that the average frequency of violations equals the nominal level α .

- **Conditional Coverage (CC)** is expressed as

$$\Pr(I_t^{\text{VaR}}(\alpha) = 1 \mid \Omega_{t-1}) = \alpha,$$

which implies that the probability of a violation, even when conditioned on all past information up to time $t-1$ (denoted Ω_{t-1}), remains equal to the nominal level α .

If both UC and CC hold, then according to [Christoffersen \(1998\)](#), the violation indicators $I_t^{\text{VaR}}(\alpha)$ form an i.i.d. sequence with $I_t^{\text{VaR}}(\alpha) \sim \text{Bernoulli}(\alpha)$.

3.2. Conformal Prediction coverage as Bernoulli trials

CP constructs a prediction set $\Gamma_t(\alpha) \subset \mathbb{R}$ at each time t , based on a user-specified miscoverage rate $\alpha \in (0, 1)$. The goal is to ensure that the prediction set contains the future observation y_t with probability at least $1 - \alpha$, meaning that miscoverage occurs with probability at most α .

Let y_t denote the realized value at time t , and let $\Gamma_t(\alpha)$ denote the corresponding prediction set. We refer to the event $y_t \in \Gamma_t(\alpha)$ as a *coverage* event, and $y_t \notin \Gamma_t(\alpha)$ as a *miscoverage* event. To track miscoverage over time, we define the binary indicator

$$I_t^{\text{CP}}(\alpha) = \mathbf{1}\{y_t \notin \Gamma_t(\alpha)\},$$

which equals 1 when the prediction set fails to contain y_t , and 0 otherwise.

The sequence $\{I_t^{\text{CP}}(\alpha)\}$ is typically evaluated against two key properties:

- **Marginal Coverage (Unconditional Validity)**

For smoothed conformal prediction with randomized tie-breaking ([Vovk et al., 2022](#)):

$$\Pr(y_t \notin \Gamma_t(\alpha)) = \Pr(I_t^{\text{CP}}(\alpha) = 1) = \alpha.$$

- **Conditional Coverage (Conditional Validity)**

$$\Pr(I_t^{\text{CP}}(\alpha) = 1 \mid \Omega_{t-1}) = \alpha,$$

Conditional coverage by [Lei and Wasserman \(2014\)](#) requires that the coverage (or miscoverage) probability remains at the nominal level even when conditioning on covariates, which we extend to all past information Ω_{t-1} and thereby strengthening the definition.

If both marginal and conditional coverage hold exactly, the indicators $I_t^{\text{CP}}(\alpha)$ form an i.i.d. sequence with $I_t^{\text{CP}}(\alpha) \sim \text{Bernoulli}(\alpha)$.

Since CP reduces uncertainty quantification to a binary sequence of coverage events, recording a hit when $y_t \in \Gamma_t(\alpha)$ and a miss otherwise, most statistical tests designed for Bernoulli sequences, such as those used in Value at Risk evaluation, can be directly applied to assess the empirical calibration of conformal predictors.

4. An Example Backtesting Suite for Conformal Time-Series Predictions

4.1. Taxonomy of Backtesting Methods

We classify backtesting procedures into five principal categories, each targeting different aspects of CP evaluation. The categories one to three and five are adapted from [Zhang and Nadarajah \(2017\)](#), whose review paper provides the basis for most of the tests we include and where many examples for each category can be found. The fourth category is developed by us, drawing inspiration from [Nolde and Ziegel \(2017\)](#).

- **Unconditional test methods** focus on whether the frequency of misses matches the marginal coverage rate. These methods evaluate the long-run average number of misses and are primarily concerned with verifying if the CP model is exactly calibrated.
- **Independence property test methods** exclusively target the violation of temporal independence of misses. These tests determine whether they are randomly scattered over time, without reference to their marginal frequency, and are useful for detecting time-dependent structures or clustering.
- **Conditional test methods** jointly assess both the frequency and the violation of independence. They examine whether violations occur with the exact probability and whether the occurrence of a violation is conditionally independent of past information, thus capturing potential clustering in violations.
- **Comparative performance tests** evaluate and contrast the forecasting ability of two or more CP methods by comparing their average predictive performance. This is typically done using scoring rules or loss functions that account for both coverage and interval width. Such tests are particularly relevant when selecting between competing methods that offer different validity and efficiency trade-offs.
- **Other test approaches** encompass more general or advanced methods that go beyond the binary violation framework. This includes tests that account for the magnitude of violation, evaluate the full predictive distribution, or use model residuals, durations, or other features to detect subtle misspecifications.

We now introduce four complementary diagnostics and one comparative test, each of which can be applied to the binary hit/miss sequence. Taken together, these diagnostics aim to explore how a CP method behaves in the presence of typical challenges in time-series applications, such as marginal miscoverage, temporal dependence, input dependence, data shifts, and relative differences between models, offering a compact starting point for empirical evaluation.

4.2. Binomial Tests for Marginal Coverage

To assess whether the empirical violation rate deviates from the nominal miscoverage level α , we define the **test statistic** as $K = \sum_{t=1}^n I_t(\alpha)$, where $I_t(\alpha) = \mathbf{1}\{y_t \notin \Gamma_t(\alpha)\}$ indicates whether the realized value falls outside the prediction set at time t .

The **null hypothesis** is that the prediction intervals achieve marginal coverage $H_0 : \mathbb{E}[I_t(\alpha)] = \alpha$. We consider three **alternative hypotheses**:

- (i) under-coverage, $H_A : \mathbb{E}[I_t(\alpha)] > \alpha$;
- (ii) over-coverage, $H_A : \mathbb{E}[I_t(\alpha)] < \alpha$;
- (iii) two-sided deviation, $H_A : \mathbb{E}[I_t(\alpha)] \neq \alpha$.

Under the null hypothesis, the **distribution of the test statistic** is binomial: $K \sim \text{Bin}(n, \alpha)$, where n is the number of test points. Let k_{obs} denote the observed number of violations. The corresponding p -values are computed as

$$p_{\text{under}} = \Pr(K \geq k_{\text{obs}}), \quad p_{\text{over}} = \Pr(K \leq k_{\text{obs}}),$$

$$p_{\text{two-sided}} = \sum_{\{k: \Pr(K=k) \leq \Pr(K=k_{\text{obs}})\}} \Pr(K = k).$$

At level β , the respective null hypothesis is rejected if the corresponding p -value falls below β . In a VaR context, binomial tests have been widely applied, including in [Kupiec \(1995\)](#) and the [Basel Committee on Banking Supervision \(1996\)](#). These are classified as **unconditional test methods**, as they assess whether the frequency of violations matches the nominal risk level, without considering temporal structure.

4.3. Geometric–Conformal Backtesting via Weibull Duration Models

To examine temporal dependence in CP violations, we follow the duration-hazard framework of [Pelletier and Wei \(2016\)](#), with the only modification being the use of a standard likelihood ratio test instead of their Monte Carlo-based p -value computation, which reduces computational cost but may lead to higher Type II error in finite samples.

Let D_i be the random variable denoting the number of steps between the i -th and $(i+1)$ -st violations, and let d_i be its observed realization. Durations are modeled with a discrete Weibull hazard

$$\Pr(I_{t_i+d} = 1 \mid \Omega_{t_i+d-1}) = a d^{b-1} \exp(c x_{t_i+d}), \quad d = 1, 2, \dots$$

where $a \in (0, 1)$ sets the unconditional error level, $b > 0$ captures duration dependence ($b < 1$: clustering, $b > 1$: dispersion), and $c \geq 0$ links the hazard to a covariate x_{t_i+d} . In [Pelletier and Wei \(2016\)](#), the covariate x_{t_i+d} is set to the predicted value at risk VaR_{t_i+d} , which is the α -quantile forecast made d time steps after the i -th violation. This scalar defines a clear threshold for a violation. In CP, violations occur when the observation falls outside a prediction interval $\Gamma_t(\alpha)$, so there is no single threshold value. We therefore set $c = 0$ and focus on the duration-based parameters a and b :

$$\Pr(I_{t_i+d} = 1 \mid \Omega_{t_i+d-1}) = a d^{b-1}.$$

Log-likelihood. The parameters a and b are estimated by maximizing the possibly censored log-likelihood of the observed run-lengths between misses. See Appendix A for the full derivation and explicit formulae.

We define three likelihood ratio tests to assess different aspects of CP validity. The first test, the Geometric Unconditional Coverage test (Geo-UC) evaluates whether the marginal coverage rate a matches the nominal level α , assuming independent violations ($b = 1$). The second test, Geo-Ind, assesses whether the run-lengths are geometrically distributed ($b = 1$), which corresponds to independence between violations, without testing the coverage level. The third test, Geo-Joint, simultaneously evaluates both marginal calibration and independence by jointly testing $a = \alpha$ and $b = 1$.

Test statistic.

$$\begin{aligned} \text{LR}_{\text{Geo-UC}} &= -2[\log L(a = \alpha, b = 1) - \log L(\hat{a}, b = 1)], \\ \text{LR}_{\text{Geo-Ind}} &= -2[\log L(\hat{a}, b = 1) - \log L(\hat{a}, \hat{b})], \\ \text{LR}_{\text{Geo-Joint}} &= -2[\log L(a = \alpha, b = 1) - \log L(\hat{a}, \hat{b})]. \end{aligned}$$

Null and alternative hypotheses.

$$\begin{aligned} \text{Geo-UC:} \quad & H_0: a = \alpha, b = 1 \quad H_A: a \neq \alpha, b = 1, \\ \text{Geo-Ind:} \quad & H_0: b = 1 \quad H_A: b \neq 1, \\ \text{Geo-Joint:} \quad & H_0: a = \alpha, b = 1 \quad H_A: a \neq \alpha, b \neq 1. \end{aligned}$$

Distribution of the test statistic under the null.

$$\text{LR}_{\text{Geo-UC}}, \text{LR}_{\text{Geo-Ind}} \sim \chi_1^2 \quad \text{and} \quad \text{LR}_{\text{Geo-Joint}} \sim \chi_2^2$$

4.4. Adapted Dynamic Binary Test (DBT) for Conformal Validity

A valid conformal predictor implies that no challenger model using historical data can systematically predict miscoverage events better than random guessing with probability α . To test this, we **adapt** Dumitrescu et al. (2012) slightly by using **more variables**, **dimensionality reduction** and by fitting a **two-stage logistic regression model** to the non-coverage indicators $I_t(\alpha) = \mathbf{1}\{y_t \notin \Gamma_t(\alpha)\}$, which under conformal validity satisfies $\Pr(I_t(\alpha) = 1 \mid \Omega_{t-1}) = \alpha$ for all t .

For estimating the conditional violation probability $\pi_t = \Pr(I_t = 1 \mid \Omega_{t-1})$ in our two-stage procedure, we first fit a base logistic regression model to lagged covariates to obtain predicted miss probabilities $\hat{\pi}_t$. These are then fed into a second-stage logistic regression model $1/(1 + e^{-z})$, whose linear predictor z is given by:

$$\begin{aligned} z = c &+ \sum_{i=1}^p \beta_i I_{t-i} + \sum_{j=1}^q \gamma_j^\ell \ell_{t-j} + \sum_{j=1}^q \gamma_j^u u_{t-j} + \sum_{j=1}^q \delta_j^\ell \ell_{t-j} I_{t-j} + \sum_{j=1}^q \delta_j^u u_{t-j} I_{t-j} \\ &+ \sum_{k=1}^s \psi_k y_{t-k} + \sum_{k=1}^s \phi_k y_{t-k} I_{t-k} + \sum_{r=1}^r \eta_r \hat{\pi}_{t-r}. \end{aligned}$$

This model includes lagged violations I_{t-i} , prediction interval bounds (ℓ_{t-j}, u_{t-j}) , their interactions with prior violations, lagged target values y_{t-k} , and lagged predicted probabilities $\hat{\pi}_{t-r}$ from stage one.

To reduce the effective degrees of freedom in the final test and thereby improve its statistical power, we use the common approach of applying principal component analysis (PCA) to the full matrix of predictor variables (Li et al., 2016), i.e., the covariates used to model miscoverage, in the second-stage regression. The number of retained components is chosen to preserve a fixed proportion (e.g., 95%) of the total variance. Logistic regression is then applied to these components to model the conditional non-coverage probability.

We use two complementary tests to assess the validity of conformal predictors. The independence test detects whether the violation process can be predicted based on past information, indicating dynamic dependence, while the conditional coverage test evaluates whether the violation probability remains equal to α when conditioning on past information. Together, they form a joint diagnostic for assessing temporal dependence and marginal calibration in conformal methods.

Independence test. This test evaluates whether violations are serially dependent:

$$\begin{aligned} H_0^{\text{ind}} : & \quad \pi_t = \text{const} \quad (\text{i.i.d. violations}), \\ H_1^{\text{ind}} : & \quad \pi_t \text{ depends on past information.} \end{aligned}$$

Conditional coverage test. This test evaluates whether the violation probability remains equal to α after conditioning:

$$\begin{aligned} H_0^{\text{cc}} : & \quad \pi_t = \alpha \quad \text{for all } t, \\ H_1^{\text{cc}} : & \quad \pi_t \text{ depends on past information.} \end{aligned}$$

Test statistic. Both tests are based on the likelihood ratio statistic

$$\text{LR} = -2(\ell_{\text{null}} - \ell_{\text{full}}),$$

where $\ell_{\text{full}} = \sum_t I_t \log(\pi_t) + (1 - I_t) \log(1 - \pi_t)$ is the log-likelihood of the full PCA-based model. The null model differs between the two tests: for the independence test, ℓ_{null} denotes the log-likelihood of an intercept-only logistic regression with a free intercept parameter c ; for the conditional coverage test, $\ell_{\text{null}} = \sum_t I_t \log(\alpha) + (1 - I_t) \log(1 - \alpha)$, corresponding to a model with fixed predicted probabilities $\pi_t = \alpha$.

Distribution under the null hypothesis.

- Under H_0^{ind} , we have $\text{LR}_{\text{ind}} \sim \chi_k^2$, where k is the number of retained principal components.
- Under H_0^{cc} , we have $\text{LR}_{\text{cc}} \sim \chi_{k+1}^2$, accounting for the fixed intercept under H_0^{cc} .

4.5. Comparative Interval-Score Tests for Conformal Prediction Quality

The coverage-based backtests from Sections 4.2, 4.3, and 4.4, even when extended to assess dynamic properties, do not allow us to differentiate among models that all meet the required coverage guarantees. Nolde and Ziegel (2017) argue that classical backtesting, rooted in null hypothesis testing, is not suited for comparative model evaluation. These tests are designed to reject poorly calibrated models but do not enable principled ranking among valid alternatives. For example, two conformal prediction (CP) methods may achieve similar empirical

coverage and average interval width, yet differ in how they adapt to local uncertainty. In such settings, violation-based tests alone offer no insight into which method yields *better* intervals.

To move beyond binary admissibility, we adopt the comparative backtesting framework of [Nolde and Ziegel \(2017\)](#), which leverages strictly proper scoring rules and Diebold-Mariano-type statistics ([Diebold and Mariano, 2002](#)) to compare predictive performance. For interval forecasts, we use the *interval score* ([Gneiting and Raftery, 2007](#)), defined as

$$S_t(\ell_t, u_t; y_t) = (u_t - \ell_t) + \frac{2}{\alpha}(y_t - u_t)\mathbf{1}\{y_t > u_t\} + \frac{2}{\alpha}(\ell_t - y_t)\mathbf{1}\{y_t < \ell_t\},$$

where (ℓ_t, u_t) is the prediction interval and y_t the observed value. The score penalizes wide intervals and miscoverage, and is minimized in expectation when the forecast interval is the true $(1 - \alpha)$ quantile interval.

The interval score is a *strictly proper scoring rule*, meaning it incentivizes honest reporting: the expected score is uniquely minimized when the forecaster reports their true belief. More formally, a scoring rule is *proper* if the forecaster cannot improve their expected score by reporting any distribution other than their true belief, and *strictly proper* if the minimum is uniquely achieved when the reported forecast matches the true distribution ([Gneiting and Raftery, 2007](#)). This ensures that score-based comparisons evaluate both calibration and sharpness in a principled manner.

To compare two methods with intervals $(\ell_t^{(1)}, u_t^{(1)})$ and $(\ell_t^{(2)}, u_t^{(2)})$, we compute the score difference

$$d_t = S_t(\ell_t^{(1)}, u_t^{(1)}; y_t) - S_t(\ell_t^{(2)}, u_t^{(2)}; y_t).$$

Null and alternative hypothesis.

$$H_0: \mathbb{E}[d_t] = 0 \quad \text{vs.} \quad H_1: \mathbb{E}[d_t] \neq 0.$$

The null states that both methods perform equally in terms of expected interval score; the alternative states that one method performs better (i.e., yields lower expected score).

Test statistic. We apply the Diebold-Mariano (DM) statistic:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{\sigma}^2/T}}, \quad \bar{d} = \frac{1}{T} \sum_{t=1}^T d_t,$$

where T is the number of test observations, and the long-run variance $\hat{\sigma}^2$ is estimated using a HAC estimator with Bartlett kernel:

$$\hat{\sigma}^2 = \hat{\gamma}_0 + 2 \sum_{k=1}^{h-1} w_k \hat{\gamma}_k, \quad \hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^T (d_t - \bar{d})(d_{t-k} - \bar{d}), \quad w_k = 1 - \frac{k}{h},$$

and $h = \lfloor T^{1/3} \rfloor$ is the truncation lag. This follows the Bartlett-weighted estimator proposed in [Andrews \(1991\)](#).

Distribution under the null hypothesis.

$$\text{DM} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } T \rightarrow \infty.$$

We use a two-sided test. A significantly negative (positive) value of DM indicates that Method 1 (Method 2) produces more informative intervals, as measured by the interval score.

Before applying the Diebold-Mariano test, it is important to ensure that the score difference series d_t satisfies weak stationarity. This means that the mean and variance of d_t are constant over time, and that the autocovariances depend only on the lag. Under these conditions, the long run variance estimator of Andrews (1991) is consistent, and the test statistic converges to the standard normal distribution as shown in Diebold and Mariano (2002). We recommend testing for weak stationarity using complementary procedures such as the Augmented Dickey-Fuller (ADF) test (Said and Dickey, 1984) and the KPSS test (Kwiatkowski et al., 1992). We present the Diebold-Mariano test here because it is a widely used and intuitive method for comparing forecast accuracy in time-series settings.

In general, the assumption of stationarity can still be reasonable even when the forecast target series is nonstationary. If both methods adapt similarly to changes in the data generating process, the score differences may remain stable. In such cases, the weak stationarity of d_t is preserved and the Diebold-Mariano comparison remains valid.

If weak stationarity is not plausible, we recommend using the Conditional Predictive Ability (CPA) test of Giacomini and White (2006). The CPA framework does not assume stationarity of the score difference series and is designed to assess whether two forecasting models have equal conditional predictive performance, that is, whether their expected forecast losses differ when conditioned on available information. Importantly, the CPA test does not rank models or identify which model performs better; rather, it tests whether there exists any systematic difference in performance, either unconditionally or conditional on past information. This makes it well suited for environments with structural instability or changing forecast accuracy across regimes.

It is important to note that the Diebold-Mariano test evaluates the overall quality of predictive intervals, a joint function of calibration and efficiency, rather than coverage adherence alone. As a result, it may favor a method with slightly lower coverage (e.g., 88% instead of 90%), if the intervals are substantially narrower. The interval score inherently balances coverage and width, and the DM test quantifies whether one method achieves a better trade-off on average.

Thus, it serves as an *absolute evaluation criterion* for interval forecasts, not a purely conditional one. This methodology constitutes a *comparative performance test*, enabling quality-based ranking among conformal predictors that satisfy approximate validity, even when one method marginally violates nominal coverage.

5. Experimental Setup¹

We now describe the experimental design used to evaluate the proposed backtesting framework. To highlight its ability to uncover calibration failures and structural weaknesses in conformal prediction methods, we conduct controlled experiments on synthetic data with known distributional regimes, and apply the framework to real-world electricity and financial time series characterized by non-stationarity and temporal dependence.

1. Code available at <https://github.com/KonradRtz/Coverage-Testing-in-Non-stationary-Time-Series>

5.1. Synthetic Data Sets

To evaluate whether our backtesting framework detects miscalibration under violations of exchangeability, we conduct controlled experiments on synthetic time-series data. Following Barber et al. (2023), we test CP methods under three regimes: i.i.d., abrupt changepoints, and smooth drift.

Data-generating process. We simulate $N = 2000$ observations from a linear model with Gaussian covariates and noise: $x_t \sim \mathcal{N}(0, I_4)$, $y_t = x_t^\top \beta^{(t)} + \varepsilon_t$, and $\varepsilon_t \sim \mathcal{N}(0, 1)$, where the regression coefficients $\beta^{(t)}$ vary over time to induce nonstationarity. The three scenarios are presented in Table 1:

Table 1: Time-varying regression coefficients.

Scenario	$\beta^{(t)}$
i.i.d.	constant $(2, 1, 0, 0)$
Changepoints	$(2, 1, 0, 0) \rightarrow (0, -2, -1, 0) \rightarrow (0, 0, 2, 1)$ at $t = 500, 1500$
Drift	linear from $(2, 1, 0, 0)$ to $(0, 0, 2, 1)$

Methods compared. We compare three variants of CP based on linear regression, following the framework of Barber et al. (2023). Each method uses a different combination of model fitting and residual weighting. Let $\hat{f}_t(x) = x^\top \hat{\beta}_t$ denote the linear predictor used to evaluate inputs x at time t .

- **CP+LS:** Ordinary Least Squares (OLS) regression with coefficients computed from $\hat{\beta}_n = \arg \min_{\beta} \sum_{t=1}^n (y_t - x_t^\top \beta)^2$, and residuals are computed via leave-one-out refitting and used unweighted.
- **NexCP+LS:** Same OLS model as CP+LS, but residuals are reweighted with exponential decay: $w_t = 0.99^{n-t}$, normalized via $\tilde{w}_t = w_t / \sum_{t'} w_{t'}$, and used to compute the conformal quantile.
- **NexCP+WLS:** Model is fitted via weighted least squares: $\hat{\beta}_n = \arg \min_{\beta} \sum_{t=1}^n l_t (y_t - x_t^\top \beta)^2$, with $l_t = 0.99^{n-t}$. Residuals are weighted as in NexCP+LS. Since the learner depends on the full tag vector $(l_1, \dots, l_n, l_{n+1})$, the inclusion of the test point breaks symmetry and thus exchangeability. To restore symmetry, we follow Barber et al. (2023) and randomly swap the tag l_{n+1} with the tag of a randomly selected training point before fitting.

Evaluation protocol. We apply full conformal prediction with leave-one-out residuals in a rolling-window setup from $n = 100$ to 1999. At each step:

1. For each $t = 1, \dots, n$, fit the regression model $\hat{f}_t(x_t)$ using all data points except (x_t, y_t) , and compute the absolute residual $r_t = |y_t - \hat{f}_t(x_t)|$.
2. Fit \hat{f}_n on all n points and evaluate $\hat{f}_n(x_{n+1})$.

3. Construct the prediction interval $\Gamma_n(x_{n+1}) = [\hat{f}_n(x_{n+1}) \pm q_{1-\alpha}]$, where $q_{1-\alpha}$ is the (possibly weighted) $(1 - \alpha)$ -quantile of the residuals.
4. Record whether $y_{n+1} \in \Gamma_n(x_{n+1})$; use the resulting binary sequence for statistical testing.

5.2. Real-World Data Sets

Real-world data provides a critical setting for evaluating the practical performance of CP methods under complex, uncontrolled conditions. In contrast to synthetic setups with fully specified mechanisms, real-world covariates often contain nontrivial structure and statistically exploitable dependencies. This creates favorable conditions for methods that target conditional coverage, which may construct tighter intervals by exploiting meaningful feature–response relationships. We assess both classical and conditionally valid conformal approaches on two domains: electricity demand forecasting and financial time series.

5.2.1. ELECTRICITY DATA SET

We adopt the first part of the real-world data setup from [Barber et al. \(2023\)](#) to evaluate our backtesting framework on the ELEC2 dataset ([Harries, 1999](#)), which records electricity prices and demand in New South Wales and Victoria at 30-minute intervals from 1996 to 1999. Following their preprocessing steps, we focus on the 9:00am–12:00pm window and discard an initial segment with constant response values, yielding $N = 3444$ time points.

The prediction target is the variable **transfer**, which measures the quantity of electricity transferred between the two states. As covariates, we use four input variables: **nswprice** and **vicprice** (electricity prices in New South Wales and Victoria, respectively), and **nswdemand** and **vicdemand** (electricity demand in the two states).

We use the first 3000 observations to construct one-step-ahead 90% prediction intervals in a rolling-window fashion, as described previously. The final 444 observations are held out for evaluating coverage and applying our statistical backtests.

As in the synthetic setup, we use these 4 covariates and compare three conformal methods: CP+LS, NexCP+LS, and NexCP+WLS. Prior work by [Vovk et al. \(2021\)](#) has shown that this dataset exhibits distributional drift that violates exchangeability, making it a natural test case for nonexchangeable CP methods and our diagnostic framework.

5.2.2. FINANCIAL TIME SERIES

We furthermore conduct experiments on financial time series data characterized by non-stationarity, volatility clustering, and temporal dependence, features that systematically violate standard CP assumptions such as exchangeability.

We consider daily return series for 101 stocks from the S&P 100 index, spanning from 2000 to 2025. Each time series is divided into three contiguous periods: a training set from 2000 to 2012 for model fitting, a calibration set from 2013 to 2019 for interval calibration, and a test set from 2019 to 2025 for evaluation.

A single global **LightGBM** ([Ke et al., 2017](#)) model is trained using 60 lags of the daily percentage change on the aggregated training data across all assets. To ensure a uniform amount of calibration data per time series, we retain only those stocks with continuous

price data available from 2013 to 2025, resulting in a final set of 98 time series out of 101. To ensure model reliability, we perform a random hyperparameter search and 5-fold cross-validation. Prediction intervals for the test period are then constructed using three CP methods.

- **Split Conformal Prediction (SCP)** by [Papadopoulos et al. \(2002\)](#): a standard baseline applying conformal correction to residuals from the calibration set,
- **Conformalized Quantile Regression (CQR)** by [Romano et al. \(2019\)](#): designed to achieve conditional coverage by applying conformal correction to a quantile regression model,
- **Adaptive Conformal Inference (ACI)** by [Gibbs and Candès \(2021\)](#): an adaptive procedure that updates prediction intervals to maintain marginal coverage over time, without explicitly targeting local conditional validity.

6. Experimental Results & Discussion

We now report and interpret the results of applying our backtesting framework to the synthetic and real life experiments described in Section 5. For each method, we apply the full suite of backtests and record whether the corresponding null hypotheses were rejected.

6.1. Synthetic Data Results

In the i.i.d. setting (Table 2), all methods passed every test, confirming theoretical validity and showing no false rejections under exchangeability.

Under changepoints (Table 2), **CP+LS** failed nearly all tests due to lack of adaptation. **NexCP+LS** passed more marginal coverage tests but still failed the Binomial under-coverage test, suggesting insufficient responsiveness. **NexCP+WLS** passed all the marginal coverage test but failed the independence and conditional coverage tests completely, which was due to an overreaction after a changepoint, where **NexCP+WLS** became too conservative ([Barber et al., 2023](#)).

In the distribution drift setting (Table 2), only **NexCP+WLS** passed all marginal, independence and conditional coverage tests; **CP+LS** failed entirely, and **NexCP+LS** did not pass the unconditional coverage tests. This underscores the benefit of jointly adapting both model and residuals under gradual distribution shifts.

Predictive performance (Table 3) was identical across methods under i.i.d. conditions. Under changepoints, both adaptive methods outperformed **CP+LS**, with **NexCP+WLS** performing best. Under distribution drift, **NexCP+WLS** demonstrated superior predictive ability, while **CP+LS** and **NexCP+LS** were statistically indistinguishable.

Table 2: Test outcomes across settings. “Yes” indicates the test did not reject the null hypothesis at the 5% significance level (i.e., the model passed). “No” indicates rejection of the null.

Setting	Method	Binomial (under.)	Binomial (over.)	Binomial (two-s.)	Geo- UC	Geo- Ind	Geo- Joint	DBT Ind	DBT CC
i.i.d.	CP+LS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	NexCP+LS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	NexCP+WLS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Changepoint	CP+LS	No	Yes	No	No	No	No	No	No
	NexCP+LS	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	NexCP+WLS	Yes	Yes	Yes	Yes	No	No	No	No
Drift	CP+LS	No	Yes	No	No	No	No	No	No
	NexCP+LS	No	Yes	No	No	Yes	Yes	Yes	Yes
	NexCP+WLS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 3: Pairwise Diebold-Mariano comparisons across settings. “Winner” indicates the method with lower avg. interval score. “Tie” indicates no significant difference at the 5% level.

Setting	Comparison	Winner
i.i.d.	CP+LS vs. NexCP+LS	Tie
	CP+LS vs. NexCP+WLS	Tie
	NexCP+LS vs. NexCP+WLS	Tie
Changepoints	CP+LS vs. NexCP+LS	NexCP+LS
	CP+LS vs. NexCP+WLS	NexCP+WLS
	NexCP+LS vs. NexCP+WLS	NexCP+WLS
Distribution drift	CP+LS vs. NexCP+LS	Tie
	CP+LS vs. NexCP+WLS	NexCP+WLS
	NexCP+LS vs. NexCP+WLS	NexCP+WLS

6.2. Real-World Data Results

6.2.1. ELECTRICITY DATA SET RESULTS

In the ELEC2 dataset (Table 4), only **NexCP+WLS** passed the unconditional coverage tests. All methods failed the independence and conditional coverage tests, indicating that the residual violation sequence may not be fully independent or conditionally calibrated. This could reflect persistent temporal structure, covariate-conditioned miscoverage, or adaptation-induced instability. In the comparative backtest (Table 5), **NexCP+LS** and **NexCP+WLS** demonstrated higher predictive quality than **CP+LS**, with no significant difference between **NexCP+LS** and **NexCP+WLS**.

Table 4: Test outcomes for the electricity data set. “Yes” indicates that the test did not reject the null hypothesis at the 5% significance level (i.e., the model passed the test), while “No” indicates rejection of the null.

Method	Binomial (undercov.)	Binomial (overcov.)	Binomial (two-sided)	Geo-UC	Geo-Ind	Geo-Joint	DBT Ind	DBT CC
CP+LS	No	Yes	No	No	No	No	No	No
NexCP+LS	No	Yes	No	No	No	No	No	No
NexCP+WLS	Yes	Yes	Yes	Yes	No	No	No	No

Table 5: Pairwise Diebold-Mariano comparisons for the electricity data set. “Winner” indicates the method with lower avg. interval score. “Tie” indicates no significant difference at the 5% significance level.

Setting	Comparison	Winner
Electricity	CP+LS vs. NexCP+LS	NexCP+LS
	CP+LS vs. NexCP+WLS	NexCP+WLS
	NexCP+LS vs. NexCP+WLS	Tie

6.2.2. FINANCIAL TIME SERIES RESULTS

In the financial dataset (Table 6), **SCP** failed nearly all tests, confirming its breakdown in volatile, nonstationary environments. **CQR**, which leverages quantile regression to approximate conditional validity, passed the independence tests in at least 80 out of 98 stocks and the conditional coverage tests in at least 45, the strongest performance among all methods. **ACI** passed nearly all marginal and geometric tests but failed the independence tests for up to 88 stocks and the conditional coverage tests for up to 76, reflecting a focus on long-run coverage rather than local structure. In terms of predictive quality, the comparative

backtest based on the interval score (Table 7) showed that both **ACI** and **CQR** significantly outperformed **SCP**, with no significant difference between them.

Table 6: Number of time series (out of 98) for which each test *did not reject* the null hypothesis at the 5% level. Higher values indicate stronger validity.

Method	Binomial (under)	Binomial (over)	Binomial (two-s.)	Geo-UC	Geo-Ind	Geo-Joint	DBT Ind	DBT CC
SCP	8	94	4	5	1	1	8	4
CQR	31	97	39	41	90	45	80	53
ACI	96	98	97	97	10	22	28	30

Table 7: Pairwise Diebold-Mariano comparisons on interval score. “Better” indicates that the first-mentioned method performed significantly better at the 5% significance level for a given stock.

Comparison	Better	Worse	Tie
SCP vs. CQR	0	27	71
SCP vs. ACI	0	12	86
CQR vs. ACI	0	0	98

6.3. General Implications

The tests behaved as expected across all settings. The independence tests (Geometric Ind and Dynamic Binary Ind) and the conditional coverage tests (Geometric Joint and Dynamic Binary CC) proved especially sensitive, uncovering dependence and conditioned violations that were missed by other diagnostics, such as the coverage rate in [Barber et al. \(2023\)](#), where some CP methods were considered acceptable, that we rejected. The Geometric-Conformal and Dynamic Binary Tests complement each other in how they assess dependence in the violation sequence. Dynamic Binary Tests rely on information from a short recent window but incorporate multiple covariates, which allows them to detect very subtle local dynamics. This is in line with [Dumitrescu et al. \(2012\)](#), who argue that dynamic binary models offer a more appropriate structure for modeling the violation process than i.e. linear Dynamic Quantile tests ([Engle and Manganelli, 2004](#)), as they directly accommodate the binary nature of violations and allow for state-dependent behavior driven by past violations and forecasts. In contrast, the Geometric independence tests aggregate information over all days since the last violation, making them less sensitive to fine-grained changes but more effective at identifying broader temporal structures such as clustering or overdispersion. [Pelletier and Wei \(2016\)](#) show that geometric duration models outperform their continuous counterparts in detecting clustering, and that their discrete hazard struc-

ture aligns well with the expected distribution of violation gaps under a well-calibrated risk model.

Importantly, we observed that joint tests can occasionally pass even when one of their underlying conditions fails. For example, in the changepoint setting, **NexCP+LS** failed the Geometric Unconditional Coverage (UC) test but still passed the Geometric Joint test, which jointly evaluates both UC and independence. This behavior arises because such tests operate under a single null hypothesis, and may absorb moderate violations in one component when the other holds. As highlighted by [Zhang and Nadarajah \(2017\)](#), joint coverage tests tend to have reduced sensitivity when misspecification affects only one dimension, potentially masking isolated calibration issues. Nevertheless, their strength lies in providing a concise overall summary of calibration, which is particularly valuable in high-frequency or large-scale model evaluations.

The results from the Diebold–Mariano test were also largely consistent with expectations, except when **CQR** and **ACI** tied which is explained by the different coverage to interval tradeoff these two methods have. However, since we do not explicitly test for stationarity of the loss differential series d_t , these results should be interpreted with caution. The comparison of **CP+LS** and **nexCP+WLS** shows that models that adapt differently to changepoints and linear drifts can still be flagged by the Diebold-Mariano test. Future work should explore alternative testing procedures like [Giacomini and White \(2006\)](#) that are robust to nonstationary forecast error dynamics.

Regarding the power of the tests, the Binomial Coverage Test ([Kupiec, 1995](#)) offers a straightforward framework with well-controlled Type I error by construction. However, the test suffers from limited power (i.e., high Type II error), especially when the empirical violation rate is close to the nominal level. As Kupiec’s original power analysis shows, detecting moderate misspecification requires a substantial number of observations, making the test relatively weak in small samples.

The Geometric Duration Tests introduced by [Pelletier and Wei \(2016\)](#) improve upon classic duration-based independence tests by explicitly modeling discrete inter-violation times. Nonetheless, these tests rely only on the subset of violation points (roughly $n \cdot \alpha$), effectively reducing sample size. As shown in their simulation studies, this leads to low power (high Type II error) in finite samples, particularly for the Geometric Independence and Joint Conditional Coverage tests, which under-reject under the alternative.

In contrast, the Dynamic Binary Tests (DBTs) proposed by [Dumitrescu et al. \(2012\)](#) exhibit much stronger power characteristics. By modeling the violation process as a binary outcome, conditioned on past violations and covariates, DBTs detect subtle patterns of conditional dependence that other methods may miss. Simulation evidence suggests they achieve significantly lower Type II error rates than Geometric Duration Tests while maintaining acceptable Type I error levels under the null, assuming correct model specification. This makes DBTs especially suitable for detecting localized or time-varying miscalibration in small to moderate sample sizes.

The Diebold–Mariano Test ([Diebold and Mariano, 2002](#)), while asymptotically valid, performs poorly in small samples. As shown in comparative studies such as [Giacomini and White \(2006\)](#), its power is low when differences in predictive accuracy are subtle or when the loss differential series d_t exhibits nonstationary behavior. Without corrections for

these issues, the test may fail to reject even when predictive accuracy diverges meaningfully, leading to inflated Type II error in practice.

Taken together, the results confirm that each method responds to our coverage diagnostics in line with its design. CP+LS fails under nonstationarity. NexCP+LS offers partial robustness. NexCP+WLS combines validity and adaptability most effectively. CQR adjusts best to local structure, while ACI secures stable coverage but misses local structures. Our test suite offers a principled foundation for selecting and evaluating conformal methods in time series forecasting, grounded in formal hypothesis testing and proper scoring rules.

7. Conclusion and Future Work

We established a formal connection between CP evaluation and Value-at-Risk (VaR) backtesting in time series. Based on this equivalence, we proposed a backtesting framework that evaluates unconditional coverage, violation independence, and conditional coverage using statistically principled methods. We also introduced a Diebold-Mariano test based on interval score to compare the forecasting abilities of CP methods.

Applied to synthetic, electricity, and financial time series data, the framework revealed coverage violations and structural weaknesses not visible through marginal metrics alone. In particular, the independence and conditional coverage tests proved sensitive to both input-conditioned and adaptation-induced miscoverage, especially in non-stationary environments. However, future work should do more extensive reviewing of CP coverage tests and report their behaviour for varying sample size or synthetic environments to look at the Type I and II errors.

We identify three main applications of our testing framework. First, it can be used to select the most appropriate CP method for a given underlying model. Second, it enables the selection of the best underlying model for a fixed CP method, recognizing that the underlying model influences not only the marginal distribution of residuals but also their conditional properties. Third, the framework facilitates joint optimization of both the underlying model and the CP method, as demonstrated in [Barber et al. \(2023\)](#).

To improve interpretability, we recommend reporting p-values at multiple conventional significance levels (e.g., 1%, 5%, and 10%). This approach facilitates the assessment of statistical evidence strength and highlights cases, where p-values lie close to, but above, typical rejection thresholds.

Future work should account for sampling bias when interpreting statistical tests for CP validity. In particular, Hoeffding’s inequality ([Vovk et al., 2022](#)) offers a finite-sample bound on the deviation between the observed coverage and its expectation. This provides a principled way to quantify the maximum sampling error under the i.i.d. assumption, and can serve as a conservative correction when interpreting empirical coverage results or computing p-values. Research in this direction could for example define a minimum calibration set size to make the sampling bias negligible.

In addition, the effect of repeated model refitting remains an open question. [Escanciano and Olmo \(2011\)](#) showed that refitting introduces additional variance in predictive systems. It is still unclear whether this affects only the underlying model, the CP method, or both. Understanding how this refitting bias propagates in CP pipelines will be crucial for ensuring validity in dynamically updated systems.

Acknowledgements

The authors would like to thank Christof Seiler for providing feedback on an earlier version of the paper.

References

- Donald W. K. Andrews. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858, 1991. doi: 10.2307/2938229.
- Anastasios Nikolas Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online Conformal Prediction with Decaying Step Sizes. In *Proc. 41st International Conference on Machine Learning*, volume 235, pages 1616–1630, 2024.
- Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The Limits of Distribution-Free Conditional Predictive Inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2020. doi: 10.1093/imaiai/iaaa017.
- Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal Prediction Beyond Exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.
- Basel Committee on Banking Supervision. Supervisory Framework for the Use of “Back-testing” in Conjunction with the Internal Models Approach to Market Risk Capital Requirements. Technical Report 22, Bank for International Settlements, 1996.
- Peter F. Christoffersen. Evaluating Interval Forecasts. *International Economic Review*, 39(4):841–862, 1998. doi: 10.2307/2527341.
- Nicolo Colombo. On Training Locally Adaptive CP. In *Proc. 12th Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204, pages 384–398, 2023.
- Francis X. Diebold and Roberto S. Mariano. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144, 2002. doi: 10.1198/073500102753410444.
- Elena-Ivona Dumitrescu, Christophe Hurlin, and Vinson Pham. Backtesting Value-at-Risk: From Dynamic Quantile to Dynamic Binary Tests. *Finance*, 33(1):79–112, 2012. doi: 10.3917/fina.331.0079.
- Robert F. Engle and Simone Manganelli. Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381, 2004.
- Juan Carlos Escanciano and Jose Olmo. Robust Backtesting Tests for Value-at-Risk Models. *Journal of Financial Econometrics*, 9(1):132–161, 2011. doi: 10.1093/jjfneec/nbq021.
- Raffaella Giacomini and Halbert White. Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578, 2006.
- Isaac Gibbs and Emmanuel J. Candès. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.
- Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025. doi: 10.1093/jrsssb/qkaf008.

- Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.
- Michael Harries. Splice-2 Comparative Evaluation: Electricity Pricing. Technical report, University of New South Wales, School of Computer Science and Engineering, 1999.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Paul H. Kupiec. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives*, 3(2):73–84, 1995.
- Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3):159–178, 1992. doi: 10.1016/0304-4076(92)90104-Y.
- Jing Lei and Larry Wasserman. Distribution-Free Prediction Bands for Non-Parametric Regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014. doi: 10.1111/rssb.12021.
- Zhengbang Li, Wei Zhang, Dongdong Pan, and Qizhai Li. Power Calculation of Multi-Step Combined Principal Components with Applications to Genetic Association Studies. *Scientific Reports*, 6:26243, 2016. doi: 10.1038/srep26243.
- Natalia Nolde and Johanna F. Ziegel. Elicitability and Backtesting: Perspectives for Banking Regulation. *The Annals of Applied Statistics*, 11(4):1833–1874, 2017. doi: 10.1214/17-AOAS1041.
- Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*, pages 345–356, 2002.
- Denis Pelletier and Wei Wei. The Geometric-VaR Backtesting Method. *Journal of Financial Econometrics*, 14(4):725–745, 2016. doi: 10.1093/jjfinec/nbv015.
- Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Said E. Said and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984. doi: 10.1093/biomet/71.3.599.
- Evgeni Smirnov. Coverage vs Acceptance-Error Curves for Conformal Classification Models. In *Proc. 12th Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204, pages 534–545, 2023.
- Vladimir Vovk. On-line confidence machines are well-calibrated. In *Proc. 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 187–196, 2002. doi: 10.1109/SFCS.2002.1181895.
- Vladimir Vovk, Ivan Petej, and Alex Gammerman. Protected Probabilistic Classification. In *Proc. 10th Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152, pages 297–299, 2021.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2022. doi: 10.1007/978-3-031-06649-8.

- Chen Xu and Yao Xie. Conformal Prediction Interval for Dynamic Time-Series. In *Proc. 38th International Conference on Machine Learning*, volume 139, pages 11559–11569, 2021.
- Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive Conformal Predictions for Time Series. In *Proc. 39th International Conference on Machine Learning*, pages 25834–25866, 2022.
- Yuanyuan Zhang and Saralees Nadarajah. A Review of Backtesting for Value at Risk. *Communications in Statistics - Theory and Methods*, 47(15):3616–3639, 2017. doi: 10.1080/03610926.2017.1361984.

Appendix A. Log-Likelihood for Discrete Weibull Duration Models in Conformal Backtesting

This appendix details the likelihood formulation used to assess temporal dependence in Conformal Prediction (CP) misses. A *miss* refers to the event that an observation lies outside the predicted interval. To examine whether such misses are independent over time, we model the number of time steps between consecutive misses as a discrete random variable and estimate the likelihood under a duration-dependent hazard model.

A.1. Discrete Hazard Framework

Let $D_i \in \{1, 2, \dots\}$ denote the number of steps between the i -th and $(i + 1)$ -st misses. Following the discrete duration framework of [Pelletier and Wei \(2016\)](#), the conditional hazard function is given by:

$$\Pr(I_{t_i+d} = 1 \mid I_{t_i+1} = 0, \dots, I_{t_i+d-1} = 0) = a d^{b-1}, \quad d = 1, 2, \dots$$

where:

- $a \in (0, 1)$ is the baseline miss rate,
- $b > 0$ controls the duration dependence:
 - $b = 1$: constant hazard (geometric),
 - $b < 1$: decreasing hazard (miss clustering),
 - $b > 1$: increasing hazard (miss dispersion).

This hazard specification induces a discrete distribution over the inter-miss durations (run-lengths).

A.2. Duration Probabilities

Define the discrete hazard at duration step j as $h(j) = a j^{b-1}$. The probability that the duration between two misses is exactly d steps is:

$$\Pr(D = d) = h(d) \cdot \prod_{j=1}^{d-1} (1 - h(j))$$

The corresponding survival function, which gives the probability that a miss occurs after at least d steps, is:

$$\Pr(D \geq d) = \prod_{j=1}^{d-1} (1 - h(j))$$

A.3. Censored Log-Likelihood

Let $D = (D_1, \dots, D_N)$ denote the observed durations between misses. In practice, the first or last duration may be only partially observed, a situation referred to as censoring:

- If the time series **begins** without a miss, then D_1 is **left-censored**: the start of this run is unknown; you only observe the duration from the beginning of the sample to the first miss.
- If the time series **ends** without a subsequent miss, then D_N is **right-censored**: the end of this run is unknown; you only observe the duration from the last miss to the end of the sample.

Let d_i denote the observed value (minimum possible duration) for the i -th run, and $C_1, C_N \in \{0, 1\}$ be censoring indicators:

- $C_1 = 1$ if D_1 is left-censored, else $C_1 = 0$
- $C_N = 1$ if D_N is right-censored, else $C_N = 0$

The log-likelihood function accounting for censoring is:

$$\begin{aligned} \log L(D \mid a, b) &= C_1 \log \Pr(D_1 \geq d_1) + (1 - C_1) \log \Pr(D_1 = d_1) \\ &\quad + \sum_{i=2}^{N-1} \log \Pr(D_i = d_i) \\ &\quad + C_N \log \Pr(D_N \geq d_N) + (1 - C_N) \log \Pr(D_N = d_N) \end{aligned}$$

The parameters a and b are estimated via numerical maximization of this likelihood, typically using constrained optimization (e.g., L-BFGS-B) with bounds $a \in (0, 1)$, $b > 0$. The resulting estimates \hat{a} and \hat{b} are used in the likelihood ratio tests presented in Section 4.3.