CONFORMAL COVERAGE FOR TIME SERIES: VALIDITY AND INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction provides marginal coverage guarantees, yet practitioners may wonder if the observed coverage is truly abnormal or consistent with sampling variation. Inference for realized coverage has received comparatively little attention, especially for time series. We study split conformal prediction with adjacent calibration and test sets of temporally dependent data. Using the functional dependence measure, we derive non-asymptotic bounds on marginal coverage error without mixing assumptions, which can be difficult to verify and may fail even for simple short-memory models. We establish a Bahadur representation to derive, to our knowledge, the first central limit theorem for realized coverage of split conformal prediction under temporal dependence. A consistent block-based estimator of the standard error yields an asymptotically justified test. We further study long-memory time series, which remain understudied in conformal prediction. For Gaussian linear processes, we show how very strong temporal dependence can lead to a non-Gaussian limiting law of realized coverage and establish block-sampling inference with an estimated normalization. The resulting theory explains how temporal dependence changes coverage uncertainty.

1 INTRODUCTION

Conformal prediction constructs prediction sets with a prescribed coverage level without specifying a parametric model for the response distribution (Vovk et al., 2005; Lei et al., 2018). For time series, the observations used to calibrate the prediction sets and the future observations are temporally dependent. This raises two questions. First, does the prediction set still attain the nominal marginal coverage? Second, how should we assess the fraction of observations covered over a future test window? The purpose of this paper is to answer these questions for split conformal prediction, with particular attention to inference for realized coverage under long memory.

Consider prediction intervals for electricity demand with nominal coverage of 90%, but realized coverage of 85% over the next evaluation window. Is this shortfall evidence of miscalibration, or can it be explained by sampling variation under persistent forecast errors? Answering this question requires the distribution of realized coverage, which under long memory may fluctuate on a slower scale and need not have a Gaussian limit.

Existing results establish conformal validity beyond exchangeability and provide quantitative guarantees under temporal dependence (Barber et al., 2023; Chernozhukov et al., 2018; Oliveira et al., 2024; Barber & Pananjady, 2026). Many time-series guarantees use mixing conditions (Xu & Xie, 2021). These conditions can be difficult to verify and may fail even for elementary short-memory models (Andrews, 1984; Han & Wu, 2023). General validity frameworks need not assume mixing, but their bounds still require suitable control of dependence. We use the functional dependence measure (FDM) introduced by Wu (2005b), which describes the effect of changing an innovation on future observations. This gives non-asymptotic bounds on marginal coverage error for processes outside mixing assumptions.

Marginal validity alone does not determine the distribution of realized coverage. Oliveira et al. (2024) also provide concentration bounds for the covered fraction over a test window. Such bounds control large deviations, whereas inference based on a limiting distribution provides a reference law for testing an observed departure from the nominal level. Gazin (2024) studies this distributional question under iid and exchangeable assumptions. For temporally dependent scores, we establish a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Bahadur representation of the calibration cutoff and then derive a central limit theorem (CLT) for realized coverage. The limiting variance is generally unknown; a block-based estimator makes the corresponding standard error available in practice.

Related coverage-inference targets should be distinguished. Ramos et al. (2026) study the coverage probability conditional on calibration, and Cotton (2026) studies variation in the marginal probability content of the estimated cutoff with an independent test draw. Neither is the realized fraction over an adjacent dependent test window. Retzlaff et al. (2025) develop conformal backtests for marginal and conditional calibration, including violation counts, but do not derive the split-conformal realized-coverage CLT studied here. Halkiewicz (2026) uses physical dependence and Bahadur expansions for rolling-origin conformal prediction; our claim is therefore not the first use of FDM in conformal inference. Online procedures that update their prediction rule to control long-run coverage address a different setting (Gibbs & Candès, 2024). We calibrate once and evaluate the resulting prediction sets over the next window.

Our general FDM conditions allow long memory in the observations in principle, but do not describe all degrees of persistence in the coverage indicators. We therefore study Gaussian linear processes separately. Long-memory time series remain understudied in conformal prediction. Earlier theory does permit some long-memory settings; for example, Ren (2024) allows long-memory stochastic volatility in a high-frequency event-study framework. This does not provide a distributional theory for realized coverage over growing, adjacent calibration and test windows. In our setting, absolute-error scores yield Gaussian or non-Gaussian coverage limits, depending on the strength of memory. Classical empirical-process and quantile theory supplies the tools for identifying these limits (Dehling & Taqqu, 1989; Coeurjolly, 2008).

The non-Gaussian regime requires more than estimating a standard error. We use the block-sampling method of Zhang et al. (2013) to estimate the coverage distribution. Each historical block repeats calibration followed by evaluation on an adjacent subwindow, with the same length ratio as the target problem. The resulting block distribution permits inference without a Gaussian approximation. Our proof combines the coverage expansion with innovation coupling, connecting this method to the functional-dependence approach used in the general results.

Our contributions can be summarized as follows:

1. **Marginal coverage.** We derive non-asymptotic marginal coverage bounds using FDM, broadening applicability beyond mixing-based assumptions, and allowing long-memory processes in principle.
2. **Inference for realized coverage.** We establish a Bahadur representation and, to our knowledge, the first central limit theorem for realized coverage of split conformal prediction under temporal dependence. We provide a consistent block-based standard-error estimator for asymptotically valid coverage testing.
3. **Long-memory inference.** We provide the first distributional theory and block-sampling inference for realized coverage of split conformal prediction under long memory that we are aware of. For long-memory Gaussian linear processes, we derive central and non-central limits. Building on Zhang et al. (2013), we estimate the memory parameter from calibration residuals and justify the resulting normalization, yielding asymptotically valid tests across all three long-memory regimes.

Section 2 introduces the setup and functional dependence. Section 3 presents the general marginal-coverage bound, Bahadur representation, CLT, and standard-error estimator. Section 4 studies the long-memory linear model and the block-sampling method. Proofs and supporting technical results are in the appendix.

2 SETUP

Let $Z_t = (X_t, Y_t)$ be a stationary time series, where $X_t \in \mathbb{R}^d$ and $Y_t \in \mathbb{R}$. A predictor $\widehat{\mu}$ trained on an independent dataset is treated as a fixed function on the calibration and test sets. We use the nonconformity score $V_t = s(X_t, Y_t) = |Y_t - \widehat{\mu}(X_t)|$; the process assumptions below apply with this predictor fixed.

The calibration set consists of $V_1, \ldots, V_n$. For a nominal coverage level $1-\alpha$, where $0 < \alpha < 1$, define

$$k_n = \lceil (n+1)(1-\alpha) \rceil, \qquad \widehat{q}_n = V_{(k_n)}, \qquad \widehat{C}_n(x) = [\widehat{\mu}(x) - \widehat{q}_n, \widehat{\mu}(x) + \widehat{q}_n], \tag{1}$$

where $V_{(k)}$ is the kth order statistic of the calibration set. We use $\widehat{q}_n = \infty$ if $k_n > n$. The estimated cutoff is unchanged over the adjacent test set, indexed by $n+1, \ldots, n+m$.

This paper concerns two quantities associated with this prediction interval. First, for a new data point, is the marginal coverage rate $\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\}$ close to $1-\alpha$? Second, what is the asymptotic distribution of realized coverage on the adjacent test set? Its *realized coverage* is the fraction of test responses covered:

$$\widehat{\mathrm{Cov}}_{n,m} = \frac{1}{m} \sum_{t=n+1}^{n+m} \mathbf{1}\{Y_t \in \widehat{C}_n(X_t)\} = \frac{1}{m} \sum_{t=n+1}^{n+m} \mathbf{1}\{V_t \le \widehat{q}_n\}. \tag{2}$$

We next describe the functional dependence measure (FDM), the key theoretical device of this paper (Wu, 2005b). Write $Z_t = G(\ldots, \varepsilon_{t-1}, \varepsilon_t)$ and $V_t = H(\ldots, \varepsilon_{t-1}, \varepsilon_t)$, where G and H are measurable functions and $(\varepsilon_t)_{t \in \mathbb{Z}}$ are iid random innovations, possibly vector-valued. These representations are assumed well defined; the same function is used at every time. Replace ε_0 by an independent copy ε_0', leaving all other innovations unchanged, and denote the resulting variables by Z_k^* and V_k^*. For $r \ge 1$, define

$$\delta_r^Z(k) = \|Z_k - Z_k^*\|_r, \qquad \delta_r^V(k) = \|V_k - V_k^*\|_r, \qquad k \ge 0, \tag{3}$$

where $\|U\|_r = (\mathbb{E}|U|^r)^{1/r}$, using the Euclidean norm for vectors. We use $r = 2$ unless stated otherwise. FDM measures how much changing one innovation affects an observation k steps later. Faster decay means that this effect disappears more quickly; geometric decay is typical of models satisfying geometric moment contraction. The coupling follows the data-generating mechanism and can often be bounded directly from a model's recursion or linear coefficients.

For a scalar process with finite variance, we call its memory short when its autocovariances are absolutely summable. The long-memory processes considered here have eventually positive, non-summable autocovariances. Summability of the L^2 FDM is a sufficient condition for short memory, although failure of that condition alone does not establish long memory. Long-memory models are used in financial and geophysical applications (Zhang et al., 2013). Such dependence can change both the normalization and the limiting distribution of sample statistics, making the usual root-n Gaussian approximation inappropriate (Wu, 2003).

In the conformal prediction setup, we are concerned with the FDM of score-threshold indicators:

$$\theta_k = \sup_{v \in \mathbb{R}} \|\mathbf{1}\{V_k \le v\} - \mathbf{1}\{V_k^* \le v\}\|_2, \qquad \Theta = \sum_{k=0}^{\infty} \theta_k. \tag{4}$$

Mixing conditions instead compare events in entire past and future sigma-fields. Verifying their decay can be difficult, and even elementary autoregressions with discrete innovations can fail to be strongly mixing (Andrews, 1984; Han & Wu, 2023; Wu, 2005b).

The temporal dependence of the observations, nonconformity scores, and threshold indicators need not be the same. Their FDMs are related in a model-specific way; Lemma 8 in Appendix A.1 gives useful sufficient bounds. Section 3 assumes summable indicator FDM, which ensures short memory of the threshold indicators while allowing long-memory observations.

3 General time series

We study marginal coverage and realized-coverage inference under functional dependence. Let $F(v) = \mathbb{P}(V_t \le v)$ be the marginal CDF of the nonconformity scores defined in Section 2.

Assumption 1 (Nonconformity scores and functional dependence)**.** *The marginal CDF F is continuous, and the threshold indicators have summable functional dependence: $\Theta < \infty$, with Θ defined in* (4)*.*

This summability condition implies short memory of each threshold-indicator process and can be checked directly even when mixing fails. Consider the stationary recursion $Y_t = (Y_{t-1} + \varepsilon_t)/2$, with iid $\varepsilon_t \sim \text{Bernoulli}(1/2)$, and the fixed predictor $\widehat{\mu} \equiv 1/2$. Then Y_t is uniform on $[0,1]$, the nonconformity scores $V_t = |Y_t - 1/2|$ are uniform on $[0, 1/2]$, and $\theta_k \leq \min\{1, 2^{(1-k)/2}\}$. Although the recursion has geometrically decaying autocovariances, $V_{t-1} = |2V_t - 1/2|$: an arbitrarily distant future score determines the present score, so the scores are not even α-mixing. This is the type of distinction between mixing and functional dependence discussed by Wu (2005b); Han & Wu (2023). Appendix A.2 verifies these claims and the regularity needed for the results below.

Assumption 1 also allows long-memory observations in principle. For example, let $X_t = M_t$ and $Y_t = M_t + \eta_t$, where M_t is a stationary causal long-memory process and (η_t) is iid, independent of M_t, with a continuous absolute-value distribution. The fixed predictor $\widehat{\mu}(x) = x$ removes the long-memory component and gives iid nonconformity scores $V_t = |\eta_t|$, satisfying Assumption 1. Section 4 studies Gaussian linear models beyond this assumption, including regimes where the nonconformity scores and their threshold indicators have long memory.

Let $q = F^{-1}(1-\alpha) = \inf\{v : F(v) \geq 1-\alpha\}$ be the population $(1-\alpha)$-quantile, and write f for the marginal density where it exists. The following theorem provides two non-asymptotic bounds on marginal coverage error.

Theorem 1 (Non-asymptotic marginal coverage bounds)**.** *Under Assumption 1, the following bounds hold.*

(i) *If* $2/n + (2\Theta^2/n)^{1/3} < \min\{\alpha, 1-\alpha\}$*, then the split-conformal prediction set satisfies*

$$\left|\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\} - (1-\alpha)\right| \leq \frac{2}{n} + 3\left(\frac{\Theta^2}{4n}\right)^{1/3}. \tag{5}$$

(ii) *Suppose additionally that, for some* $e, c_f, L > 0$*, F has a density $f(v) \geq c_f$ on $[q-e, q+e]$, and the conditional distribution of V_t given the innovation past $\mathcal{F}_{t-1} = \sigma(\varepsilon_j : j \leq t-1)$ has a density bounded by L on this interval almost surely. If $2/n \leq c_f e/2$, then*

$$\left|\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\} - (1-\alpha)\right| \leq \frac{2L}{c_f n} + \frac{2\sqrt{2}L\Theta}{c_f\sqrt{n}} + \frac{8\Theta^2}{nc_f^2e^2}. \tag{6}$$

Both bounds hold at every deterministic future index $t > n$.

Proof: Appendix A.3.

These bounds account for calibration error while allowing temporal dependence between the calibration set and immediately adjacent test points. Part (i) is density-free; part (ii) gives an $O(n^{-1/2})$ bound under an additional local conditional-density condition. That condition is used only in part (ii).

To study the asymptotic distribution of realized coverage, we add local regularity of the marginal distribution at the target quantile. Write $F_n(v) = n^{-1}\sum_{t=1}^{n} \mathbf{1}(V_t \leq v)$ for the empirical CDF of the calibration nonconformity scores.

Assumption 2 (Quantile regularity)**.** *There are $e > 0$ and $0 < c_f \leq C_f < \infty$ such that F has a density with $c_f \leq f(v) \leq C_f$ for $|v - q| \leq e$. The density is continuous at q. For distinct integer times s, t, $\mathbb{P}(V_s = V_t) = 0$.*

The Bahadur representation replaces the random calibration cutoff by an indicator average at the population quantile. This is the step needed to transfer a fixed-threshold limit to conformal coverage.

Theorem 2 (Bahadur representation)**.** *Under Assumptions 1–2,*

$$F(\widehat{q}_n) - (1-\alpha) = -\frac{1}{n}\sum_{t=1}^{n}\{\mathbf{1}(V_t \leq q) - (1-\alpha)\} + o_{\mathbb{P}}(n^{-1/2}), \tag{7}$$

$$\widehat{q}_n - q = -\frac{1}{nf(q)}\sum_{t=1}^{n}\{\mathbf{1}(V_t \leq q) - (1-\alpha)\} + o_{\mathbb{P}}(n^{-1/2}). \tag{8}$$

Proof: Appendix A.4.

Bahadur representations under dependence are classical (Wu, 2005a); here the proof uses the exact conformal rank and a local empirical-process bound under the stated indicator FDM. The following theorem is the main result of this section. The Bahadur representation allows us to establish asymptotic normality of realized coverage.

Theorem 3 (Asymptotic normality of realized coverage)**.** *Under Assumptions 1–2, with $n, m \to \infty$, the series $\sigma^2_{\mathrm{cov}} = \sum_{h\in\mathbb{Z}} \mathrm{Cov}\{\mathbf{1}(V_0 \le q), \mathbf{1}(V_h \le q)\}$ is absolutely convergent, and*

$$\sqrt{\frac{nm}{n+m}}\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\} \xrightarrow{d} N(0, \sigma^2_{\mathrm{cov}}). \tag{9}$$

The zero-variance Gaussian is allowed in this statement.

Proof: Appendix A.5.

Calibration uncertainty and test variation both contribute: the standard-error scale is $\sigma_{\mathrm{cov}}\sqrt{1/n + 1/m}$. Positive accumulated lagged indicator covariance increases this standard error relative to iid scores. For iid nonconformity scores, the classical first-order calibration and test variance contributions are $\alpha(1-\alpha)/n$ and $\alpha(1-\alpha)/m$, respectively, giving the standard error $\sqrt{\alpha(1-\alpha)(1/n + 1/m)}$. The limiting variance σ^2_{cov} is generally unknown, so feasible inference requires estimating it. We use nonoverlapping blocks of calibration indicators at the estimated cutoff. With a deterministic block length $\ell = \ell_n$ and number of blocks $B = \lfloor n/\ell \rfloor$, define

$$\overline{I}_j = \frac{1}{\ell}\sum_{t=(j-1)\ell+1}^{j\ell} \mathbf{1}(V_t \le \widehat{q}_n), \qquad \overline{I} = \frac{1}{B}\sum_{j=1}^{B}\overline{I}_j, \qquad \widehat{\sigma}^2_{\mathrm{cov}} = \frac{\ell}{B}\sum_{j=1}^{B}(\overline{I}_j - \overline{I})^2. \tag{10}$$

The estimator uses the B complete blocks and divisor B. The following proposition shows that this variance estimator is consistent.

Proposition 4 (Consistent block-based variance estimator)**.** *Under Assumptions 1–2, suppose that $\sum_{k\ge1}(\sum_{j\ge k}\theta_j)^{1/2} < \infty$, $\ell \to \infty$, and $\ell/n \to 0$. Then $\widehat{\sigma}^2_{\mathrm{cov}} \xrightarrow{\mathbb{P}} \sigma^2_{\mathrm{cov}} > 0$.*

Proof: Appendix A.6.

The additional tail condition holds for geometrically decaying θ_k and for $\theta_k = O(k^{-a})$ with $a > 3$. This condition is for variance estimation only. It also guarantees a positive limiting variance. Such block-based variance estimation is classical (Flegal & Jones, 2010).

Combining Proposition 4 with Theorem 3 gives the feasible standard error $\widehat{\mathrm{SE}} = \widehat{\sigma}_{\mathrm{cov}}\sqrt{1/n + 1/m}$. After the test set is observed, the statistic $\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\}/\widehat{\mathrm{SE}}$ has an asymptotic standard normal reference under the stated nonconformity-score and calibration model. For $0 < \eta < 1$, a two-sided level-η diagnostic rejects when its absolute value exceeds $z_{1-\eta/2}$; a lower-tail diagnostic flags unusually low coverage, where z_u is the u-quantile of $N(0,1)$. Before observing the test set, $[(1-\alpha) \pm z_{1-\eta/2}\widehat{\mathrm{SE}}] \cap [0,1]$ is an asymptotic prediction interval for its random realized coverage.

For a small number $B \ge 2$ of blocks, a block-based Student-t approximation (Jones et al., 2006) replaces z_u by $\sqrt{B/(B-1)}\, t_{B-1,u}$, where $t_{\nu,u}$ is the u-quantile of Student's t distribution with ν degrees of freedom. Appendix A.6 derives this reference under fixed-B asymptotics; the scale factor accounts for divisor B.

4 A STUDY OF LONG-MEMORY LINEAR PROCESSES

The preceding theory assumes summability of the FDM of the score-indicator process. Although it allows long-memory observations, the class it covers remains restricted. We now study realized coverage beyond this assumption, using Gaussian linear processes to describe temporal dependence. Their coefficients allow arbitrarily strong memory within the stationary range.

Assumption 3 (Gaussian linear residuals)**.** *The fixed predictor has residuals $W_t = Y_t - \widehat{\mu}(X_t) = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$, where $\varepsilon_t \overset{\mathrm{iid}}{\sim} N(0,1)$, $a_0 \ne 0$, $a_j \sim c j^{-\beta}$, $c > 0$, and $1/2 < \beta < 1$. The nonconformity scores are $V_t = |W_t|$.*

Write $\sigma^2_W = \mathbb{E}W_0^2$ and $C_\gamma = c^2\int_0^\infty x^{-\beta}(1+x)^{-\beta}\,dx$. Then $\mathrm{Cov}(W_0, W_h) \sim C_\gamma h^{1-2\beta}$, so the residual process has long memory throughout $1/2 < \beta < 1$. Taking the absolute value changes the

covariance decay: both V_t and its coverage indicator have covariance of order $h^{2-4\beta}$. They have long memory when $\beta \leq 3/4$, and summable covariances when $\beta > 3/4$. As memory becomes stronger, the limiting law of realized coverage changes from Gaussian to non-Gaussian. A related phase transition appears in covariance inference for long-memory time series (Zhai et al., 2026). Define the corresponding rate by

$$r_n = \begin{cases} n^{-1/2}, & 3/4 < \beta < 1, \\ \sqrt{\log n / n}, & \beta = 3/4, \\ n^{1-2\beta}, & 1/2 < \beta < 3/4. \end{cases} \tag{11}$$

Theorem 5 (Bahadur representation under long memory)**.** *Under Assumption 3, the conformal cut-off satisfies*

$$\widehat{q}_n - q = -\frac{1}{nf(q)} \sum_{t=1}^{n} \left\{ \mathbf{1}(V_t \leq q) - (1-\alpha) \right\} + o_{\mathbb{P}}(r_n). \tag{12}$$

Proof: Appendix B.1.

The representation is proved for the exact conformal rank, using classical Gaussian empirical-process and quantile arguments (Dehling & Taqqu, 1989; Coeurjolly, 2008). The proof also provides an upper bound on marginal coverage error: $\sup_{t>n} |\mathbb{P}(V_t \leq \widehat{q}_n) - (1-\alpha)| = O(r_n)$ (Lemma 9); the supremum concerns deterministic future indices.

The following theorem provides the asymptotic distribution of realized coverage in the three memory regimes and identifies the transition from a Gaussian law to a law determined by a Rosenblatt process.

Theorem 6 (Limiting distributions of realized coverage)**.** *Under Assumption 3 and $m/n \to \kappa \in (0, \infty)$:*

(i) *If $3/4 < \beta < 1$, then*

$$\sqrt{n}\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\} \xrightarrow{d} N\big(0, \sigma^2_{\mathrm{cov}}(1+\kappa^{-1})\big). \tag{13}$$

Here σ^2_{cov} is the covariance sum defined in Theorem 3 and is strictly positive in this model.

(ii) *If $\beta = 3/4$, then*

$$\sqrt{\frac{n}{\log n}}\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\} \xrightarrow{d} N\left(0, \left\{\frac{qf(q)C_\gamma}{\sigma_W^2}\right\}^2 (1+\kappa^{-1})\right). \tag{14}$$

(iii) *If $1/2 < \beta < 3/4$, then*

$$\frac{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)}{n^{1-2\beta}} \xrightarrow{d} \frac{qf(q)C_\gamma}{\sigma_W^2\sqrt{(3-4\beta)(4-4\beta)}} \left\{ R_H(1) - \frac{R_H(1+\kappa) - R_H(1)}{\kappa} \right\}, \tag{15}$$

where $H = 2-2\beta$ and R_H is the variance-one Rosenblatt process, oriented as the limit of centered-square sums.

Part (i) has the same Gaussian form as Theorem 3, although the indicator FDM is not summable in this linear model. At $\beta = 3/4$, the limit remains Gaussian but requires the logarithmic normalization. In the ultra-long-memory regime $1/2 < \beta < 3/4$, it becomes non-Gaussian and retains dependence between adjacent Rosenblatt increments (Taqqu, 1975). Appendix B.2 proves the three laws. As before, the calibration set and test set are adjacent, with no independence assumption between them.

We use block sampling to estimate the limiting distribution of realized coverage, building on Zhang et al. (2013). Within the calibration set, we repeat calibration and adjacent evaluation on shorter windows with the same calibration–test length ratio. An estimated normalization makes the procedure applicable across all three regimes. Appendix C.1 gives the construction.

Theorem 7 (Block sampling distribution consistency)**.** *Suppose Assumption 3 holds, $m/n \to \kappa \in (0, \infty)$, $b = \lfloor n^{\zeta} \rfloor$ for a fixed $0 < \zeta < 1$, and $a_j = j^{-\beta}\{c + O(j^{-\varphi})\}$ for some $\varphi > 0$. Let $\widehat{G}_{n,b}$ and R_N be the block CDF and normalization defined in Appendix C.1. Let G be the CDF of the limit of $\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\}/R_n(\beta)$, a constant rescaling of the corresponding law in Theorem 6. Then*

$$\sup_{x \in \mathbb{R}} |\widehat{G}_{n,b}(x) - G(x)| \xrightarrow{\mathbb{P}} 0.$$

The estimated quantiles give an asymptotically valid test for realized coverage and a prediction interval for its future value; probabilities are over the joint calibration and test sets. Appendix C.4 gives the proof. Appendix D discusses the additional arguments needed for non-Gaussian innovations.

5 Experiments

We examine marginal coverage, the normal approximation for realized coverage, and block sampling under long memory. Throughout, we use nonconformity scores based on absolute errors and adjacent calibration and test sets. The conformal cutoff uses the exact rank $k_n = \lceil (n+1)(1-\alpha) \rceil$, with $\alpha = 0.1$ unless stated otherwise.

Study 1: marginal coverage. This study examines the marginal coverage guarantee in Theorem 1 using a nonlinear process that satisfies our FDM conditions but violates mixing assumptions. Consider $U_t = (U_{t-1} + B_t)/2$, with iid $B_t \sim \mathrm{Bernoulli}(1/2)$ and independent $U_0 \sim \mathrm{Unif}(0, 1)$. Set $Y_t = U_t^2$ and $V_t = |Y_t - 1/3|$. The score process is not strongly mixing, but its indicator FDM decays geometrically; the verification is given in Appendix E.1. Thus the conditions of Theorem 1(i) hold, and its bound applies for sufficiently large n. We apply standard split conformal calibration and estimate next-step marginal coverage at nominal levels 90% and 95%.

Table 1: For Study 1, marginal coverage (%). Columns give n; rows give the nominal coverage. At $n = 2^4$, the 95% procedure has an infinite cutoff.

Nominal	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
90%	89.61	88.80	89.48	89.84	90.30	90.42	89.94	90.07	90.19
95%	100.00	94.33	94.03	94.92	94.91	94.62	94.89	94.88	95.06

Table 1 illustrates Theorem 1(i): standard split conformal calibration can approach nominal coverage even for this nonmixing process. The 90% procedure gives only 88.80% at $n = 2^5$; at $n = 2^{12}$, coverage is 90.19% and 95.06% for the two nominal levels. The contribution is a guarantee under FDM conditions where mixing assumptions fail, without changing the prediction procedure.

Study 2: realized-coverage inference for general time series. This study evaluates realized-coverage inference for general time series under the FDM conditions of Theorem 3 and Proposition 4. We consider ARCH(1) and stochastic volatility (SV), with absolute-error scores. Both satisfy the CLT and variance-estimation conditions; their specifications and verification are given in Appendix E.2. We use $m = n$. The variance is estimated from the calibration set using nonoverlapping blocks of length $\ell = \lfloor n^{2/3} \rfloor$. We compare the empirical distribution of $\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\}/\widehat{\mathrm{SE}}$ with its standard normal limit, where $\widehat{\mathrm{SE}} = \widehat{\sigma}_{\mathrm{cov}}\sqrt{1/n + 1/m}$. We also report two-sided rejection rates using standard normal critical values. Rates near the 5% and 10% test levels indicate accurate test calibration; nominal prediction coverage remains 90%.

In Figure 1, the points approach the dashed line as n grows: the studentized realized-coverage error increasingly follows $N(0, 1)$ in both models. This supports Theorem 3 and Proposition 4, and the use of the estimated standard error for coverage testing. Tail discrepancies at small n lead to over-rejection with normal critical values. Appendix Table 3 quantifies this: the 5% rejection rates fall from 10.54% and 12.92% at $n = 2^9$ to 5.49% and 5.61% at $n = 2^{19}$ for ARCH and SV. The corresponding 10% rates at the largest n are 10.77% and 10.94%. The normal approximation therefore gives increasingly well-calibrated tests.

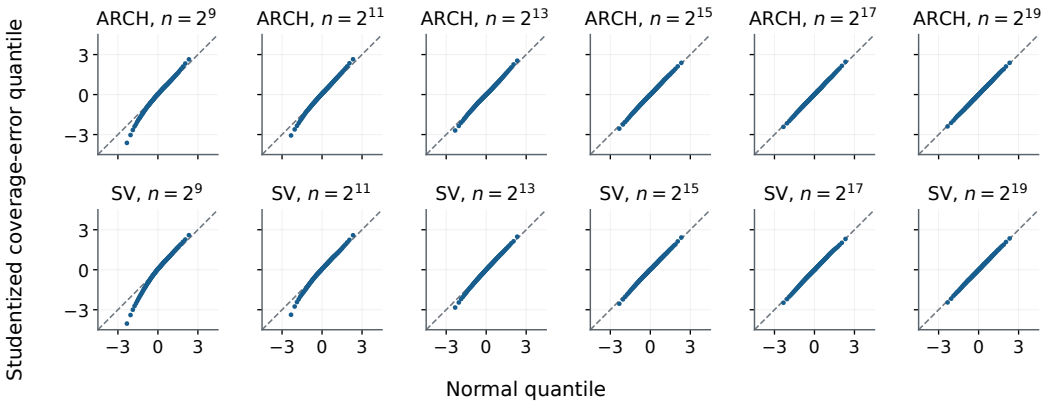

Figure 1: For Study 2, normal QQ plots of studentized realized-coverage errors. Points closer to the dashed equality line indicate closer agreement with the standard normal distribution $N(0, 1)$. No empirical recentering or rescaling is used.

We also compare with the exchangeable standard error $\sqrt{\alpha(1-\alpha)(1/n + 1/m)}$, which substantially understates coverage uncertainty in these models. At $n = 2^{19}$, the simulated standard deviations are 1.40 and 1.97 times that benchmark for ARCH and SV; nominal 5% tests using it reject 16.04% and 32.10% of the time. Ignoring temporal dependence would therefore lead to frequent false alarms about coverage. The corrected t reference reduces the rejection rates at $n = 2^9$ to 4.75% and 6.19%, using the same variance estimates. Appendix E.2, Table 3, gives the full comparison.

Study 3: block sampling method under long memory. This study evaluates the block sampling method for Gaussian linear processes across the three limiting regimes in Theorems 6 and 7. Consider the Gaussian linear model $Y_t = \sum_{j\geq 0}(1+j)^{-\beta}\varepsilon_{t-j}$ with iid $\varepsilon_t \sim N(0,1)$ and $V_t = |Y_t|$. We set $\beta \in \{0.9, 0.75, 0.6\}$, covering the central Gaussian, critical Gaussian, and non-Gaussian regimes, respectively. The $\beta = 0.6$ model has extremely strong long memory, so realized coverage is expected to converge very slowly to $1-\alpha$: its error scale is $n^{-0.2}$ rather than the usual $n^{-1/2}$. For each $n \in \{2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}\}$, we use $m = n$ and a separate independent reference sample. We use $h = \lfloor\sqrt{n}\rfloor$ for memory estimation and pairs of adjacent blocks, each of length $b = \lfloor n^{2/3}\rfloor$, within the calibration set. The memory regime is not supplied; simulation details are given in Appendix E.3.

Table 2: For Study 3, rejection rates (%) at the 5% test level. BS denotes block sampling; all methods use length $b = \lfloor n^{2/3}\rfloor$. Rates close to 5% indicate accurate two-sided rejection rates.

$n = m$	2^{12}			2^{16}			2^{20}		
β	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC
0.9	8.90	11.01	10.78	5.48	6.68	6.79	5.24	5.58	5.71
0.75	10.09	17.06	16.07	5.94	14.59	14.55	4.27	13.00	12.87
0.6	13.85	32.11	27.54	7.89	37.23	35.56	5.65	44.35	43.77

In Figure 2, agreement with the dashed line means that block quantiles approximate actual coverage-error quantiles from independent reference trajectories. Together with decreasing CDF distances (Appendix Table 7), this supports Theorem 7: block sampling estimates the distribution needed for coverage testing, including in the non-Gaussian regime. For $\beta = 0.6$, the 5% and 10% rejection rates decrease from 13.85% and 17.18% to 5.65% and 10.34%, respectively, across the sample sizes studied (Table 2 and Appendix Table 6). For reference, Appendix E.3.3 also reports $\beta = 0.7$, another ultra-long-memory case, with its QQ plots and rejection rates.

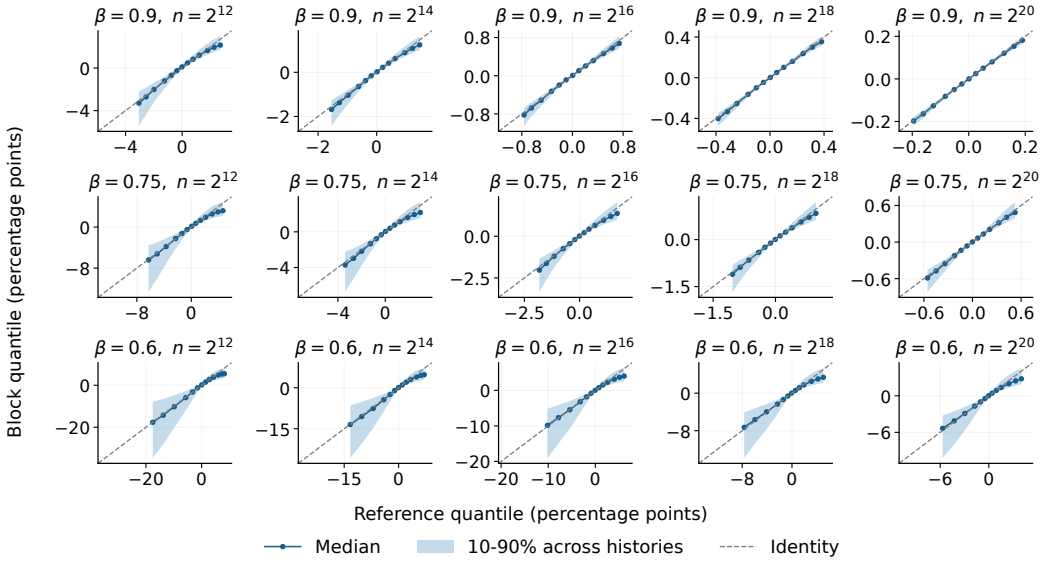

Figure 2: For Study 3, block versus independent finite-n reference error quantiles (percentage points). Curves are medians across histories; shading is the pointwise 10–90% range, not a confidence band. Closer agreement with the dashed equality line means more accurate estimation of coverage-error quantiles, not agreement with a Gaussian law.

We compare block sampling with the moving-block bootstrap (MBB) and Gaussian inference using a heteroskedasticity and autocorrelation consistent (HAC) variance estimator on the same trajectories over the full grid, using the same length $b = \lfloor n^{2/3} \rfloor$ for all three methods. Table 2 shows that the benchmarks substantially underestimate uncertainty under stronger memory, whereas block sampling improves test calibration. At $\beta = 0.6$ and $n = 2^{20}$, the 5% rejection rate is 5.65% for block sampling, compared with 44.35% and 43.77% for MBB and HAC. Appendix E.3.2 gives the procedures, 10% rejection rates, and distribution comparisons.

The strong memory also makes coverage uncertainty substantial: even at $n = 2^{20}$, the reference 2.5% and 97.5% quantiles of realized coverage for $\beta = 0.6$ are 84.29% and 94.04%, despite nominal prediction coverage of 90%.

Real Data Analysis. We analyze hourly electricity-demand forecasts in New England (Appendix F). The realized coverage is approximately 91.2% against a nominal level of 90%. The exchangeability-based test rejects nominal coverage, whereas our block sampling test does not. This example illustrates how accounting for temporal dependence can affect the interpretation of an observed coverage deviation.

6 Discussion

We develop non-asymptotic marginal coverage bounds and asymptotic inference for realized coverage under functional dependence. The marginal bounds broaden applicability beyond mixing rather than improve the $O(\log n/n)$ coverage-loss rate available under geometric β-mixing (Barber & Pananjady, 2026). For Gaussian linear processes, block sampling extends this inference to long memory, including non-Gaussian limiting distributions. These results provide a reference for assessing whether an observed departure from nominal coverage is consistent with sampling variation. The simulations and electricity-demand example show that accounting for temporal dependence can substantially change coverage uncertainty and the conclusions of coverage tests. Our analysis assumes stationary data and a predictor trained independently of the adjacent calibration and test windows. Extending realized-coverage inference to nonstationary data and predictors trained or updated on the same series is beyond the scope of this paper and is left for future work.

AI USE STATEMENT

The research question, overall approach, and experimental design were conceived by the authors. We used generative AI tools to help check theoretical claims; provide assistance in mathematical proof; identify related papers for our literature review; and implement methods and experiments. AI tools also assisted with writing and polishing the manuscript. The authors are responsible for the paper's claims, citations, and results.

REPRODUCIBILITY STATEMENT

The assumptions underlying our theoretical results are stated explicitly, with proofs provided in Appendices A–C. Appendix C.1 specifies the block sampling procedure. Appendix E describes the simulation models, parameter settings, implementation details, and additional results. Appendix F describes the real-data study and its data sources.

REFERENCES

Donald W. K. Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934, 1984. doi: 10.2307/3213710.

Rina Foygel Barber and Ashwin Pananjady. Predictive inference for time series: why is split conformal effective despite temporal dependence? In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–24. PMLR, 2026.

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.

Annika Betken and Martin Wendler. Subsampling for general statistics under long range dependence with application to change point analysis. *Statistica Sinica*, 28:1199–1224, 2018. doi: 10.5705/ss.202015.0435.

Péter Breuer and Péter Major. Central limit theorems for non-linear functionals of Gaussian fields. *Journal of Multivariate Analysis*, 13(3):425–441, 1983. doi: 10.1016/0047-259X(83)90019-2.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 732–749. PMLR, 2018.

Jean-François Coeurjolly. Bahadur representation of sample quantiles for functional of Gaussian dependent sequences under a minimal assumption. *Statistics & Probability Letters*, 78:2485–2489, 2008. doi: 10.1016/j.spl.2008.02.037.

Peter Cotton. The width of the conformal fan: Dependence and the variance of realized coverage. SSRN preprint 6979679, June 2026.

Herold Dehling and Murad S. Taqqu. The empirical process of some long-range dependent sequences with an application to U-statistics. *The Annals of Statistics*, 17(4):1767–1783, 1989. doi: 10.1214/aos/1176347394.

R. L. Dobrushin and Péter Major. Non-central limit theorems for non-linear functionals of Gaussian fields. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 50(1):27–52, 1979. doi: 10.1007/BF00535673.

James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010. doi: 10.1214/09-AOS735.

Ulysse Gazin. Asymptotics for conformal inference. *arXiv preprint arXiv:2409.12019*, 2024. doi: 10.48550/arXiv.2409.12019. Version 3, revised February 18, 2026.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Stanisław M. S. Halkiewicz. Rolling-origin conformal prediction under local stationarity and weak dependence. *arXiv preprint arXiv:2605.08422*, 2026. doi: 10.48550/arXiv.2605.08422.

Peter Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980. ISBN 9780123193506.

Peter Hall, Bing-Yi Jing, and Soumendra Nath Lahiri. On the sampling window method for long-range dependent data. *Statistica Sinica*, 8(4):1189–1204, 1998.

Fang Han and Wei Biao Wu. Probability inequalities for high-dimensional time series under a triangular array framework. In Hoang Pham (ed.), *Springer Handbook of Engineering Statistics*, pp. 849–863. Springer, London, 2 edition, 2023. doi: 10.1007/978-1-4471-7503-2_41.

Galin L. Jones, Murali Haran, Brian S. Caffo, and Ronald Neath. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006. doi: 10.1198/016214506000000492.

Jan W. Kantelhardt, Eva Koscielny-Bunde, Henio H. A. Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3–4):441–454, 2001. doi: 10.1016/S0378-4371(01)00144-3.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38, 2024.

Thiago R. Ramos, Helton Graziadei, and Luben M. C. Cabezas. Conformal prediction via transported Beta laws. *arXiv preprint arXiv:2605.19024*, 2026. doi: 10.48550/arXiv.2605.19024.

Yuexuan Ren. Conformal prediction for high-frequency event studies. Working paper, November 2024. URL `https://yuexuanren.github.io/ConformalPI.pdf`. November 11, 2024.

Konrad Retzlaff, Filip Schlembach, Dennis Bams, and Philippe Dreesen. Testing marginal and conditional coverage in conformal prediction for non-stationary time series via value-at-risk backtesting. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 725–747. PMLR, 2025.

Murad S. Taqqu. Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(4):287–302, 1975. doi: 10.1007/BF00532868.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Wei Biao Wu. Empirical processes of long-memory sequences. *Bernoulli*, 9(5):809–831, 2003.

Wei Biao Wu. On the Bahadur representation of sample quantiles for dependent sequences. *The Annals of Statistics*, 33(4):1934–1963, 2005a. doi: 10.1214/009053605000000291.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005b. doi: 10.1073/pnas.0506715102.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 2021.

Percy S. Zhai, Mladen Kolar, and Wei Biao Wu. Simultaneous inference for covariance and precision matrices of long-range dependent time series. *IEEE Transactions on Information Theory*, 72(6): 4246–4296, 2026. doi: 10.1109/TIT.2026.3685246.

Ting Zhang, Hwai-Chung Ho, Martin Wendler, and Wei Biao Wu. Block sampling under strong dependence. *Stochastic Processes and their Applications*, 123(6):2323–2339, 2013. doi: 10.1016/j.spa.2013.02.006.

A PROOFS FOR GENERAL TIME SERIES

This appendix proves the general FDM results and explains how their assumptions can be checked. Appendix A.1 develops the empirical-CDF bounds and dependence transfer; Appendix A.2 verifies the examples. Appendices A.3, A.4, and A.5 then prove marginal validity, the Bahadur representation, and the realized-coverage CLT, respectively. Appendix A.6 proves variance-estimator consistency and justifies both the normal and corrected t references for inference.

Throughout this appendix let $\mathcal{F}_j = \sigma(\varepsilon_i : i \le j)$ and $P_j U = \mathbb{E}(U \mid \mathcal{F}_j) - \mathbb{E}(U \mid \mathcal{F}_{j-1})$. For a random variable constructed from the innovations, a superscript $*$ denotes replacement of the indicated innovation coordinates by an independent copy. All nonconformity-score maps are fixed as in the main text.

A.1 COUPLING, EMPIRICAL CONTROL, AND DEPENDENCE TRANSFER

For a causal $U \in L^r$, $r \ge 1$, with single-coordinate replacement $U^{(j)}$, conditional integration gives

$$P_j U = \mathbb{E}(U - U^{(j)} \mid \mathcal{F}_j), \qquad \|P_j U\|_r \le \|U - U^{(j)}\|_r. \tag{16}$$

The identity applies when U is measurable with respect to $\mathcal{F}_t$, $t \ge j$; innovations after j are integrated out. This is the standard projection argument associated with functional dependence (Wu, 2005b). The iid remote-past tail is trivial, so every centered causal L^2 variable has its martingale projection decomposition.

Let $F_N^{(a)}$ be the empirical CDF of $V_{a+1}, \dots, V_{a+N}$, so $F_n^{(0)} = F_n$ for the calibration set. For a deterministic threshold v and fixed lag k, the variables $P_{t-k}\{\mathbf{1}(V_t \le v) - F(v)\}$ are orthogonal across t. Consequently Minkowski's inequality and (16) give

$$\left\| \sum_{t=a+1}^{a+N} \{\mathbf{1}(V_t \le v) - F(v)\} \right\|_2 \le \sum_{k \ge 0} \sqrt{N}\, \theta_k = \sqrt{N}\, \Theta. \tag{17}$$

The finite sum in t and infinite lag sum can be interchanged because the displayed lag norms are summable. The bound is uniform in a deterministic threshold, not an L^2 bound on the supremum of the empirical process.

Lemma 8 (Dependence transfer for fixed absolute nonconformity scores)**.** *Let $Z_t = (X_t, Y_t)$ be a causal stationary process, use the Euclidean norm for Z_t, and let $r \ge 1$. If $\widehat{\mu}$ is fixed and L-Lipschitz, then*

$$\|V_k - V_k^*\|_r \le (1 + L)\|Z_k - Z_k^*\|_r.$$

If the marginal nonconformity-score density is globally bounded by M, then, with $\delta_r^V(k) = \|V_k - V_k^*\|_r$,

$$\theta_k \le \min\left\{1, \sqrt{1 + 2M}\, \delta_r^V(k)^{r/(2r+2)}\right\}. \tag{18}$$

In particular, geometric nonconformity-score FDM decay implies $\sum_k \theta_k < \infty$; the bound $\delta_r^V(k) = O(k^{-a})$ suffices when $a > 2 + 2/r$.

Proof. The reverse triangle inequality gives $|V_k - V_k^*| \le |Y_k - Y_k^*| + L\|X_k - X_k^*\| \le (1+L)\|Z_k - Z_k^*\|$. For any $u > 0$, a threshold-indicator flip implies either $|V_k - v| \le u$ or $|V_k - V_k^*| > u$. Thus its probability is at most $2Mu + \delta_r^V(k)^r / u^r$. If the coefficient is positive, take $u = \delta_r^V(k)^{r/(r+1)}$, then take square roots and the supremum over v. A zero coefficient is immediate. Summing the resulting bounds gives the last assertions. No conditional score density is used. This lemma is a sufficient bound, not an equality of the dependence of data, scores, and indicators. □

Barber & Pananjady (2026, Proposition 2) also transfer dependence from data to scores, but use total-variation data processing for finite-memory score maps. Here we control single-innovation perturbations and their threshold indicators.

A.2 Examples and their assumptions

For the Bernoulli recursion in Section 3, the stationary solution is $Y_t = \sum_{j\geq 0} 2^{-j-1}\varepsilon_{t-j}$. Its binary digits are independent and fair, so Y_t is uniform on $[0,1]$ and $V_t = |Y_t - 1/2|$ is uniform on $[0,1/2]$. The recursion also gives $\mathrm{Cov}(Y_0, Y_h) = 2^{-h}/12$ for $h \geq 0$. Single-innovation replacement changes Y_k, and therefore V_k, by at most 2^{-k-1}. Since the score density is two, the probability of a threshold flip is at most $4 \cdot 2^{-k-1} = 2^{1-k}$. This proves the stated geometric bound on θ_k and also the stronger tail condition in Proposition 4.

Let $T(v) = |2v - 1/2|$. Direct substitution in the recursion gives $V_{t-1} = T(V_t)$ almost surely. For every integer $h \geq 1$, the event $A = \{V_0 \leq 1/4\}$ equals $A_h = \{T^h(V_h) \leq 1/4\}$ up to a null set. The former is measurable in the score past and the latter in its future starting at h; both have probability $1/2$. Hence $|\mathbb{P}(A \cap A_h) - \mathbb{P}(A)\mathbb{P}(A_h)| = 1/4$, so the strong-mixing coefficient does not tend to zero. In particular the scores are not β-mixing.

There are nevertheless no inter-time ties. For $h \geq 1$, conditional on the intervening Bernoulli innovations, $Y_h = 2^{-h}Y_0 + d_h$ with a fixed d_h. The event $|Y_h - 1/2| = |Y_0 - 1/2|$ forces one of two fixed values of Y_0, because $2^{-h} \neq 1$. The variable Y_0 is continuous and independent of these future innovations, so this event has probability zero. The marginal score density is positive and constant near every interior quantile. Thus the conditions of Theorem 1(i), Theorems 2–3, and Proposition 4 hold for this nonmixing example. Given the innovation past, however, V_t takes the two values $Y_{t-1}/2$ and $(1 - Y_{t-1})/2$, each with probability $1/2$. These atoms enter any fixed neighborhood of an interior q with positive probability. The additional local conditional-density condition in Theorem 1(ii) therefore fails.

For the raw long-memory example, one concrete choice is $M_t = \sum_{j\geq 0}(j+1)^{-\beta}\zeta_{t-j}$ with $1/2 < \beta < 1$ and iid standard normal ζ_t, and independent iid normal η_t. The series is well defined in L^2 and almost surely, and its covariance is asymptotic to a positive constant times $h^{1-2\beta}$, which is nonsummable. The same holds for $Y_t = M_t + \eta_t$ at nonzero lags. Taking innovation pairs (ζ_t, η_t) and $\widehat{\mu}(x) = x$ produces $V_t = |\eta_t|$, so $\theta_k = 0$ for $k \geq 1$. The example verifies applicability to long-memory observations: the fixed predictor removes their common long-memory component, leaving iid nonconformity scores and indicators.

A.3 Proof of non-asymptotic marginal coverage bounds

Proof of Theorem 1. Part (i). The sample-size condition ensures $k_n \leq n$. Continuity of F implies that for each level in $(0,1)$ there is a finite threshold attaining it. Put $A = \Theta^2/n$ within this proof. This is positive: a threshold with $0 < F(v) < 1$ has positive indicator variance, whereas (17) with $N = 1$ bounds that variance by Θ^2.

For $2/n < u < \min(\alpha, 1-\alpha)$ choose v_-, v_+ with $F(v_\pm) = 1 - \alpha \pm u$. The conformal rank obeys $0 \leq k_n/n - (1-\alpha) < 2/n$. If $\widehat{q}_n > v_+$, then $F_n^{(0)}(v_+) < k_n/n$; if $\widehat{q}_n < v_-$, then $F_n^{(0)}(v_-) \geq k_n/n$. Therefore Chebyshev's inequality and (17) imply

$$\mathbb{P}(\widehat{q}_n > v_+) \leq \frac{A}{(u - 2/n)^2}, \qquad \mathbb{P}(\widehat{q}_n < v_-) \leq \frac{A}{u^2}.$$

These implications hold with ties; no identity for $F_n^{(0)}(\widehat{q}_n)$ has been used. For every deterministic $t > n$, the events at v_- and v_+ give

$$F(v_-) - \mathbb{P}(\widehat{q}_n < v_-) \leq \mathbb{P}(V_t \leq \widehat{q}_n) \leq F(v_+) + \mathbb{P}(\widehat{q}_n > v_+).$$

No independence is required. Both one-sided deviations from $1 - \alpha$ are consequently at most $u + A/(u - 2/n)^2$.

On $u > 2/n$, this expression is minimized at $u = 2/n + (2A)^{1/3}$. The sample-size condition makes this choice admissible, and its value is $2/n + 3(A/4)^{1/3}$. This proves part (i) and its uniformity over deterministic future indices. The tolerance u is only a proof device.

Part (ii). Again put $A = \Theta^2/n$, and let $\Delta_n = |\widehat{q}_n - q|$ and $d_n = k_n/n - (1-\alpha)$. Continuity of F gives $F(q) = 1 - \alpha$. The local density lower bound implies $c_f e \leq \min\{\alpha, 1-\alpha\}$, so the

sample-size condition ensures $k_n \le n$ and $0 \le d_n < 2/n \le c_f e/2$. For $d_n/c_f < u \le e$, the order-statistic inequalities and (17) give

$$\mathbb{P}(\widehat{q}_n > q + u) \le \frac{A}{(c_f u - d_n)^2}, \qquad \mathbb{P}(\widehat{q}_n < q - u) \le \frac{A}{(c_f u + d_n)^2}.$$

Consequently

$$\mathbb{P}(\Delta_n > u) \le \min\left\{1, \frac{2A}{(c_f u - d_n)^2}\right\}, \qquad d_n/c_f < u \le e.$$

Integrating this bound and using the trivial bound one below d_n/c_f yields

$$\mathbb{E}(\Delta_n \wedge e) \le \frac{d_n}{c_f} + \frac{1}{c_f}\int_0^\infty \min\{1, 2A/x^2\}\,dx = \frac{d_n + 2\sqrt{2A}}{c_f},$$
$$\mathbb{P}(\Delta_n > e) \le \frac{2A}{(c_f e - d_n)^2} \le \frac{8A}{c_f^2 e^2}.$$

Clip the cutoff to the density neighborhood: $\widetilde{q}_n = \max\{q - e, \min\{\widehat{q}_n, q + e\}\}$. For a deterministic $t > n$, both cutoffs are $\mathcal{F}_{t-1}$-measurable. Let $K_t(v)$ be a version of $\mathbb{P}(V_t \le v \mid \mathcal{F}_{t-1})$. The conditional-density assumption makes K_t L-Lipschitz on $[q - e, q + e]$ almost surely. Since $\mathbb{E}K_t(q) = 1 - \alpha$,

$$\begin{aligned}|\mathbb{P}(V_t \le \widehat{q}_n) - (1-\alpha)| &\le \mathbb{E}|K_t(\widehat{q}_n) - K_t(\widetilde{q}_n)| + \mathbb{E}|K_t(\widetilde{q}_n) - K_t(q)| \\ &\le \mathbb{P}(\Delta_n > e) + L\mathbb{E}(\Delta_n \wedge e) \\ &\le \frac{8\Theta^2}{n c_f^2 e^2} + \frac{L}{c_f}\left(\frac{2}{n} + \frac{2\sqrt{2}\Theta}{\sqrt{n}}\right).\end{aligned}$$

This is (6) for every such t. Only a truncated cutoff error was integrated; no score moment or untruncated L^1 quantile rate is required. □

A.4 Local oscillation and the Bahadur representation

Under Assumptions 1–2, continuity and local density positivity give $F(q) = 1 - \alpha$. Quantile inversion and (17), applied at $q \pm C/\sqrt{n}$, give $\widehat{q}_n - q = O_{\mathbb{P}}(n^{-1/2})$. More explicitly, for each fixed large C and all sufficiently large n, the probability of leaving this neighborhood is at most $C'\Theta^2/(c_f^2 C^2)$; the $O(n^{-1})$ rank correction is smaller than $c_f C/(2\sqrt{n})$. Taking $C \to \infty$ proves tightness without a score moment condition.

The required random-threshold argument uses

$$D_{a,N}(v) = \{F_N^{(a)}(v) - F(v)\} - \{F_N^{(a)}(q) - F(q)\}, \qquad \psi(\delta) = 2\sum_{k \ge 0} \min\{\sqrt{C_f \delta}, \theta_k\}.$$

For $|v - q| \le \delta \le e$, the increment $G_t(v) = \mathbf{1}(V_t \le v) - \mathbf{1}(V_t \le q)$ has $\|G_t(v)\|_2 \le \sqrt{C_f \delta}$. Its coupled difference has norm bounded both by $2\sqrt{C_f \delta}$ and by $2\theta_k$ at lag k. The projection argument of (17) therefore gives $\|D_{a,N}(v)\|_2 \le \psi(\delta)/\sqrt{N}$. Dominated convergence yields $\psi(\delta) \to 0$ as $\delta \downarrow 0$.

Partition $[q - \delta, q + \delta]$ into J equal subintervals with endpoints $v_0, \dots, v_J$. Monotonicity of both CDFs gives

$$\sup_{|v-q| \le \delta} |D_{a,N}(v)| \le \max_{0 \le j \le J} |D_{a,N}(v_j)| + 2C_f \delta / J.$$

Hence, for every positive A_N,

$$\mathbb{E}\left[A_N^2 \sup_{|v-q| \le \delta} |D_{a,N}(v)|^2\right] \le 2(J+1)\frac{A_N^2}{N}\psi(\delta)^2 + \frac{8C_f^2 (A_N \delta)^2}{J^2}. \tag{19}$$

If $\delta \to 0$, A_N^2/N and $A_N \delta$ are bounded, take $J \to \infty$ slowly enough that $J\psi(\delta)^2 \to 0$. The left side then tends to zero, uniformly in a deterministic block start. All suprema here are measurable: the empirical CDF has finitely many jumps and F is continuous, so a countable dense grid suffices.

Proof of Theorem 2. Use (19) with $N = n$, $A_N = \sqrt{n}$, and $\delta = C/\sqrt{n}$. On $|\widehat{q}_n - q| \le C/\sqrt{n}$ it implies $D_{0,n}(\widehat{q}_n) = o_{\mathbb{P}}(n^{-1/2})$. First take $n \to \infty$, then $C \to \infty$ using the preceding tightness bound. No inter-time ties occur almost surely, so the exact rank identity is $F_n^{(0)}(\widehat{q}_n) = k_n/n$. Subtracting the empirical CDF at q gives

$$F(\widehat{q}_n) - (1-\alpha) = -\{F_n^{(0)}(q) - (1-\alpha)\} + \{k_n/n - (1-\alpha)\} - D_{0,n}(\widehat{q}_n).$$

The rank correction is $O(n^{-1})$. This proves (7); differentiability at q, $f(q) > 0$, and $\widehat{q}_n - q = O_{\mathbb{P}}(n^{-1/2})$ prove (8). □

A.5 Martingale approximation and realized coverage

Put $\xi_t = \mathbf{1}(V_t \le q) - (1-\alpha)$ within the proofs and define

$$\mathcal{D}_t = \sum_{k \ge 0} P_t \xi_{t+k}.$$

This series is absolutely convergent in L^2, with $\|\mathcal{D}_t\|_2 \le \Theta$. The resulting sequence is stationary and ergodic, since it is a function of the iid innovation shift, and is a martingale-difference sequence for $(\mathcal{F}_t)$. For a block of length N, reindexing the projection terms at lag k leaves two boundary sums, each of length $\min(k, N)$. Their terms are orthogonal, so

$$\left\| \sum_{t=a+1}^{a+N} (\xi_t - \mathcal{D}_t) \right\|_2 \le 2 \sum_{k \ge 1} \sqrt{\min(k, N)}\, \theta_k = o(\sqrt{N}). \tag{20}$$

For the last equality, divide by $\sqrt{N}$ and use dominated convergence with dominating summable sequence $2\theta_k$.

Orthogonality of martingale projections also gives, for $h \ge 0$,

$$|\operatorname{Cov}(\xi_0, \xi_h)| \le \sum_{j \ge 0} \theta_j \theta_{j+h}.$$

Summing over h proves absolute covariance summability. Thus $N^{-1}\operatorname{Var}(\sum_{t=1}^{N} \xi_t) \to \sigma_{\mathrm{cov}}^2$. Comparing this variance with the martingale sum using (20) shows $\sigma_{\mathrm{cov}}^2 = \mathbb{E}(\mathcal{D}_0)^2$.

Proof of Theorem 3. Set $A_{n,m} = \sqrt{nm/(n+m)}$. For arbitrary $n, m \to \infty$, $A_{n,m}^2/n \le 1$, $A_{n,m}^2/m \le 1$, and $A_{n,m}C/\sqrt{n} \le C$. Thus (19), with $N = n, m$, $A_N = A_{n,m}$, and $\delta = C/\sqrt{n}$, controls both random-cutoff terms. The required bounds hold for both block lengths, and the calibration cutoff lies in this neighborhood with probability arbitrarily close to one. Subtracting the calibration empirical CDF from the test empirical CDF at the same cutoff gives

$$\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha) = \frac{1}{m} \sum_{t=n+1}^{n+m} \xi_t - \frac{1}{n} \sum_{t=1}^{n} \xi_t + o_{\mathbb{P}}(A_{n,m}^{-1}). \tag{21}$$

The omitted rank correction is $O(n^{-1}) = o(A_{n,m}^{-1})$.

For the martingale version of the normalized contrast, use weights $w_t = -A_{n,m}/n$ on calibration and $w_t = A_{n,m}/m$ on test times. Their squared sum is one and $w_* := \max_t |w_t| \le \max\{n^{-1/2}, m^{-1/2}\} \to 0$. Let $Z_t = \mathbb{E}\{(\mathcal{D}_t)^2 \mid \mathcal{F}_{t-1}\}$. This is stationary ergodic and integrable. Its calibration average converges almost surely to σ_{cov}^2; its test-block average converges in probability to the same constant by stationarity and the ergodic theorem. The conditional variance is their convex combination:

$$\sum_t w_t^2 Z_t = \frac{m}{n+m} \frac{1}{n} \sum_{t=1}^{n} Z_t + \frac{n}{n+m} \frac{1}{m} \sum_{t=n+1}^{n+m} Z_t \xrightarrow{\mathbb{P}} \sigma_{\mathrm{cov}}^2.$$

The two deterministic weights need not converge. For every $u > 0$, the expectation of the conditional Lindeberg sum is at most

$$\mathbb{E}\left[(\mathcal{D}_0)^2 \mathbf{1}\{|\mathcal{D}_0| > u/w_*\}\right] \longrightarrow 0.$$

The martingale central limit theorem (Hall & Heyde, 1980) therefore gives the normal limit for $\sum_t w_t \mathcal{D}_t$. If $\sigma^2_{\mathrm{cov}} = 0$, this assertion follows directly in L^2. The two approximation errors in (20), multiplied by $A_{n,m}/n$ and $A_{n,m}/m$, are $o(1)$ in L^2. Indeed $A_{n,m}/\sqrt{n} \le 1$ and $A_{n,m}/\sqrt{m} \le 1$. Combining this fact with (21) proves (9). Finite calibration and test blocks have not been assumed independent. □

A.6 Block-based variance estimation and studentization

Proof of Proposition 4. Write $\Lambda(k) = \sum_{j \ge k} \theta_j$ and $\Theta_4 = \sum_{k \ge 0} \theta_k^{1/2}$. The tail condition in Proposition 4 implies $\Theta_4 < \infty$, since $\theta_k \le \Lambda(k)$ for $k \ge 1$. For the oracle indicators $I_t^o = \mathbf{1}(V_t \le q)$ let $S_j = \sum_{t=(j-1)\ell+1}^{j\ell} \{I_t^o - (1-\alpha)\}$ and $W_j = S_j^2/\ell$. Denote by $\widehat{\sigma}_o^2$ the estimator (10) with q in place of $\widehat{q}_n$. The identity for centering sums of squares gives

$$\widehat{\sigma}_o^2 = \frac{1}{B}\sum_{j=1}^{B} W_j - \ell \left\{ \frac{1}{B\ell} \sum_{t=1}^{B\ell} I_t^o - (1-\alpha) \right\}^2 .$$

The expectation of the last term is at most Θ^2/B by (17); it tends to zero. Also $\mathbb{E}W_j = \mathrm{Var}(\sum_{t=1}^{\ell} \xi_t)/\ell \to \sigma^2_{\mathrm{cov}}$.

We next control the variance of the first term. An indicator difference takes values in $\{-1, 0, 1\}$, so its L^4 norm is the square root of its L^2 norm. Consequently $\|P_0 \xi_k\|_4 \le \theta_k^{1/2}$. The projection moment inequality of Wu (2005b, Theorem 2(i)), at moment order four, yields

$$\|S_j\|_4 \le C\sqrt{\ell}\,\Theta_4, \qquad \|W_j\|_2 \le C\Theta_4^2. \tag{22}$$

All these moments are indicator moments; no moment of V_t is required.

For blocks numbered 1 and $1 + j$, $j \ge 2$, replace every innovation at times at most ℓ by an independent copy in the later block. Write S_{1+j}^* and W_{1+j}^* for the resulting quantities. They have their original marginal laws and are independent of W_1. Successive single-coordinate replacements give

$$\|I_t^o - (I_t^o)^*\|_2 \le \sum_{s \le \ell} \theta_{t-s} = \Lambda(t - \ell).$$

For rigor with the infinite past, first replace finitely many coordinates. The bounds are summable, so these coupled indicators converge in L^2; finite-coordinate approximation of their measurable causal function identifies the limit with whole-past replacement. Their L^4 distance is the square root of their L^2 distance. Thus, for the later block,

$$\|S_{1+j} - S_{1+j}^*\|_4 \le \ell\Lambda((j-1)\ell + 1)^{1/2}.$$

Using $S^2 - (S^*)^2 = (S - S^*)(S + S^*)$ and (22) gives

$$|\operatorname{Cov}(W_1, W_{1+j})| \le C\sqrt{\ell}\,\Lambda((j-1)\ell + 1)^{1/2}.$$

For $j = 0, 1$, the constant bound from (22) suffices. Since Λ is nonincreasing, disjoint groups of ℓ indices give $\sum_{j \ge 1} \Lambda(j\ell + 1)^{1/2} \le \ell^{-1} \sum_{k \ge 1} \Lambda(k)^{1/2}$. Therefore

$$\mathrm{Var}\left(\frac{1}{B}\sum_{j=1}^{B} W_j \right) \le \frac{C}{B} + \frac{C}{B\sqrt{\ell}} \sum_{k \ge 1} \Lambda(k)^{1/2} \longrightarrow 0.$$

This proves $\widehat{\sigma}_o^2 \to \sigma^2_{\mathrm{cov}}$ in probability.

It remains to handle the estimated cutoff. Let $D_j(v) = D_{(j-1)\ell,\ell}(v)$ and $\overline{D}(v) = B^{-1}\sum_{j=1}^{B} D_j(v)$. The change in the jth block average when replacing q by $\widehat{q}_n$ is $D_j(\widehat{q}_n) + F(\widehat{q}_n) - F(q)$. The last term cancels exactly after centering the block averages. The reverse triangle inequality for the Euclidean norm consequently gives

$$|\widehat{\sigma}_{\mathrm{cov}} - \widehat{\sigma}_o|^2 \le \frac{\ell}{B}\sum_{j=1}^{B} \{D_j(\widehat{q}_n) - \overline{D}(\widehat{q}_n)\}^2 \le \frac{\ell}{B}\sum_{j=1}^{B} D_j(\widehat{q}_n)^2. \tag{23}$$

On the event $|\widehat{q}_n - q| \le C_0/\sqrt{n}$, bound each summand by its supremum over this neighborhood. Apply (19) with $N = \ell$, $A_N = \sqrt{\ell}$, and $\delta = C_0/\sqrt{n}$. Here $A_N^2/N = 1$, $A_N\delta = C_0\sqrt{\ell/n} \to 0$, and $\psi(\delta) \to 0$. The expectation of the right side of (23), restricted to this event, tends to zero by stationarity and the local bound. Markov's inequality, followed by $C_0 \to \infty$ using root-n cutoff tightness, proves that it is $o_{\mathbb{P}}(1)$. Thus the plug-in estimator is consistent.

We finally verify the asserted strict positivity. The tail condition also implies $\sum_{k\ge1} \Lambda(k) < \infty$. Indeed, eventually $\Lambda(k) \le \Lambda(k)^{1/2}$. Since $\|\mathbb{E}(\xi_{t+k} \mid \mathcal{F}_t)\|_2 \le \Lambda(k)$, the series $H_t = \sum_{k\ge1} \mathbb{E}(\xi_{t+k} \mid \mathcal{F}_t)$ converges absolutely in L^2. Rearranging these convergent sums yields $\mathcal{D}_t = \xi_t + H_t - H_{t-1}$. If $\sigma_{\mathrm{cov}}^2 = \mathbb{E}(\mathcal{D}_0)^2 = 0$, then $\xi_t = H_{t-1} - H_t$ almost surely at every integer time. Put $Z_t = \exp(2\pi \mathrm{i} H_t)$, where $\mathrm{i}^2 = -1$. The integer-valued indicator in ξ_t implies

$$Z_t = \exp\{2\pi\mathrm{i}(1-\alpha)\} Z_{t-1}.$$

In particular Z_0 is $\mathcal{F}_{-r}$-measurable for every $r \ge 1$, since it is a deterministic multiple of Z_{-r}. Triviality of the iid remote-past tail makes Z_0 a constant of modulus one. Stationarity makes that same constant the value of Z_{-1}, contradicting the display because $0 < 1 - \alpha < 1$. Therefore $\sigma_{\mathrm{cov}}^2 > 0$. □

For completeness, Slutsky's theorem now gives the standard normal limit of the studentized statistic in Section 3. Positive fallback values can be assigned when the estimated variance is zero; consistency at a positive limit makes such events asymptotically negligible. Normal continuity at the relevant cutoffs proves the stated diagnostic size and prediction-interval probability. Intersecting the interval with $[0, 1]$ does not change whether it contains $\widehat{\mathrm{Cov}}_{n,m}$. Neither scale-estimator independence nor conditional-on-calibration inference is used.

A fixed-number-of-blocks reference. Assume Assumptions 1–2 and $\sigma_{\mathrm{cov}}^2 > 0$. Keep $B \ge 2$ fixed, with $\ell \to \infty$, $B = \lfloor n/\ell \rfloor$, $n = B\ell + o(n)$, and $m/n \to \kappa \in (0, \infty)$. The block conditions hold, for example, with $\ell = \lfloor n/B \rfloor$. For the estimator (10),

$$\frac{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)}{\widehat{\sigma}_{\mathrm{cov}}\sqrt{1/n + 1/m}} \xrightarrow{d} \sqrt{\frac{B}{B-1}}\, T_{B-1}, \qquad T_{B-1} \sim t_{B-1}. \tag{24}$$

The tail condition in Proposition 4 is one sufficient condition for the required positivity. Here $\ell/n \to 1/B$, giving a complementary asymptotic regime to that proposition's growing number of blocks. Equation (24) motivates the scaled-t approximation in the main text.

Proof. Write $\sigma = \sigma_{\mathrm{cov}}$ and $A_{n,m} = \sqrt{nm/(n+m)}$ within this proof, and define

$$U_{j,n} = \frac{1}{\sigma\sqrt{\ell}} \sum_{t=(j-1)\ell+1}^{j\ell} \xi_t \quad (1 \le j \le B), \qquad U_{0,n} = \frac{1}{\sigma\sqrt{m}} \sum_{t=n+1}^{n+m} \xi_t.$$

These variables converge jointly to independent standard normals $(Z_0, Z_1, \ldots, Z_B)$. To see this, apply the martingale approximation (20) to each of the finitely many disjoint blocks and use the Cramér–Wold device. For any fixed linear combination, its largest martingale weight tends to zero, and the conditional variance is a weighted sum of block averages of $\mathbb{E}\{\mathcal{D}_t^2 \mid \mathcal{F}_{t-1}\}/\sigma^2$. Each average converges to one by stationarity and ergodicity. The L^2 Lindeberg argument in the proof of Theorem 3 then gives variance equal to the sum of the squared combination coefficients. This proves the asserted joint normal limit and independence.

The incomplete calibration segment has length $n - B\ell = o(n)$. Its centered indicator sum is $o_{L^2}(\sqrt{n})$ by (17). Writing $\overline{Z} = B^{-1}\sum_{j=1}^{B} Z_j$, the coverage expansion (21) therefore gives, jointly with these block variables,

$$\frac{A_{n,m}}{\sigma}\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\} \xrightarrow{d} N_\kappa := \frac{Z_0 - \sqrt{\kappa B}\,\overline{Z}}{\sqrt{1+\kappa}}.$$

It remains to check the cutoff in the variance estimator. Write $D_j(v) = D_{(j-1)\ell,\ell}(v)$. The local bound (19), with $N = \ell$, $A_N = \sqrt{\ell}$, and $\delta = C/\sqrt{n}$, applies because $A_N^2/N = 1$ and

$A_N\delta = C\sqrt{\ell/n}$ is bounded. After localizing the cutoff as before, it gives $\sqrt{\ell}D_j(\widehat{q}_n) = o_{\mathbb{P}}(1)$ simultaneously for the fixed number of blocks. The common shift $F(\widehat{q}_n) - F(q)$ cancels after block centering. Hence, jointly with the displayed numerator,

$$\frac{\widehat{\sigma}^2_{\mathrm{cov}}}{\sigma^2} \xrightarrow{d} \frac{1}{B}\sum_{j=1}^{B}(Z_j - \overline{Z})^2 =: \frac{V}{B}.$$

Here $V \sim \chi^2_{B-1}$ is independent of $\overline{Z}$ and Z_0. Thus $N_\kappa \sim N(0,1)$ is independent of V. Since $V > 0$ almost surely, the continuous mapping theorem gives $N_\kappa/\sqrt{V/B} = \sqrt{B/(B-1)}\,T_{B-1}$, proving (24). This is an asymptotic reference; the finite-block statistic is not assumed Gaussian or exactly Student-t. □

B Proofs for the Gaussian linear model

This appendix proves Theorems 5 and 6 beyond the summable indicator-FDM conditions of Section 3. Lemma 9 first establishes the fluctuation scales and marginal coverage bound; the following calculation explains why the general FDM assumptions do not apply. Appendix B.1 derives the Bahadur representation and the calibration–test coverage contrast. Appendix B.2 uses this contrast to obtain the three limiting distributions.

Constants in this appendix may depend on the fixed model and on α, but not on the length or deterministic starting point of a block. The notation r_N means the rate in (11) evaluated at a generic block length N. Set $\gamma(h) = \mathrm{Cov}(W_0, W_h)$ and $\rho(h) = \gamma(h)/\sigma^2_W$. For a block starting after a, write $F_N^{(a)}(v) = N^{-1}\sum_{t=a+1}^{a+N}\mathbf{1}\{V_t \le v\}$. All assertions involving the conformal rank are for sufficiently large sample sizes, so $k_n \le n$ and $0 \le k_n/n - (1-\alpha) < 2/n$.

Lemma 9 (Square-sum scales and marginal coverage)**.** *Under Assumption 3, define, for this appendix,* $s^2_{2,N} = \mathrm{Var}\{\sum_{t=1}^{N}(W_t^2 - \sigma^2_W)\}$. *Then*

$$s^2_{2,N} = 2\left\{N\sigma_W^4 + 2\sum_{h=1}^{N-1}(N-h)\gamma(h)^2\right\},$$

$$s^2_{2,N} \sim \begin{cases} \dfrac{4C_\gamma^2}{(3-4\beta)(4-4\beta)}N^{4-4\beta}, & \beta < 3/4, \\ 4C_\gamma^2 N\log N, & \beta = 3/4, \\ 2N\displaystyle\sum_{h\in\mathbb{Z}}\gamma(h)^2, & \beta > 3/4. \end{cases} \tag{25}$$

In particular, $s_{2,N}/N \asymp r_N$. *Furthermore* $\widehat{q}_n - q = O_{\mathbb{P}}(r_n)$, $\mathbb{E}|\widehat{q}_n - q| = O(r_n)$, *and*

$$\sup_{t>n}|\mathbb{P}(V_t \le \widehat{q}_n) - (1-\alpha)| = O(r_n). \tag{26}$$

Proof. Square summability of (a_j) gives a stationary Gaussian linear process. The coefficient bound $|a_j| \le C(1+j)^{-\beta}$ implies

$$\gamma(h) \sim C_\gamma h^{1-2\beta}, \qquad \sum_{j\ge 0}|a_j a_{j+h}| \le C(1+h)^{1-2\beta}. \tag{27}$$

For the first equivalence, rescale the convolution by $h^{1-2\beta}$. On any compact subinterval of $(0,\infty)$ the resulting Riemann sum converges to $c^2x^{-\beta}(1+x)^{-\beta}$; the coefficient bound controls the discarded neighborhoods of zero and infinity by integrable power functions. Initial coefficients contribute only $O(h^{-\beta})$. The Gaussian identity $\mathrm{Cov}(W_0^2, W_h^2) = 2\gamma(h)^2$ now gives (25) by summing the covariances. The critical sum is harmonic; the central sum converges absolutely.

We record a useful covariance bound. If A is a centered even square-integrable function of a standard normal variable, its Hermite expansion gives

$$\begin{aligned} &\mathrm{Cov}\{A(W_0/\sigma_W), A(W_h/\sigma_W)\} = \sum_{j\ge 2,\ j\ \text{even}}\frac{\{\mathbb{E}A(Z)H_j(Z)\}^2}{j!}\rho(h)^j, \\ &0 \le \mathrm{Cov}\{A(W_0/\sigma_W), A(W_h/\sigma_W)\} \le \|A(Z)\|_2^2\rho(h)^2. \end{aligned} \tag{28}$$

Here $Z \sim N(0,1)$ and $\mathbb{E}H_j(Z)^2 = j!$. The formula follows first for finite Hermite expansions and then by L^2 approximation. It remains valid for negative $\rho(h)$ because only even orders occur. Both the centered absolute score and the centered indicator at q have a nonzero second Hermite coefficient. In fact, with $z = q/\sigma_W$ and ϕ the standard normal density, the leading covariances are

$$\mathrm{Cov}(V_0, V_h) \sim \frac{\sigma_W^2}{\pi}\rho(h)^2, \qquad \mathrm{Cov}\{\mathbf{1}(V_0 \le q), \mathbf{1}(V_h \le q)\} \sim 2z^2\phi(z)^2\rho(h)^2.$$

This proves the covariance-memory distinction stated in the main text. Applied to threshold indicators uniformly in their threshold, (28) also proves

$$\sup_v \mathrm{Var}\{F_n^{(0)}(v)\} \le Cr_n^2. \tag{29}$$

The absolute Gaussian marginal density is smooth and positive near $q > 0$. Write $\Delta_n = \widehat{q}_n - q$ and fix a sufficiently small $e_0 > 0$. Order-statistic inequalities, (29), and the rank correction give $\mathbb{P}(|\Delta_n| > u) \le Cr_n^2/u^2$ for $Kr_n \le u \le e_0$, with a fixed K. This proves tightness at scale r_n and contributes $O(r_n)$ to the integral of the tail over $[0, e_0]$. For $u > e_0$, monotonicity bounds the tail by Cr_n^2. Moreover, the event $\Delta_n > u$ requires at least $\alpha n/2$ scores above $q + u$ for large n, and $\Delta_n < -u$ requires at least $(1-\alpha)n$ scores below $q - u$. Markov's inequality on these counts gives

$$\mathbb{P}(|\Delta_n| > u) \le C_\alpha \mathbb{P}(|V_0 - q| \ge u) \le C_\alpha \mathbb{E}(V_0 - q)^2/u^2.$$

Integrating the minimum of this bound and Cr_n^2 over (e_0, ∞) gives $O(r_n)$, hence the asserted L^1 rate.

Conditionally on innovations through time $t-1$, W_t is normal with variance a_0^2. Thus the conditional density of $|W_t|$ is bounded by $2/(|a_0|\sqrt{2\pi})$. For $t > n$, the cutoff is measurable with respect to this past, so

$$|\mathbb{P}(V_t \le \widehat{q}_n) - F(q)| \le \frac{2}{|a_0|\sqrt{2\pi}}\mathbb{E}|\widehat{q}_n - q|.$$

This proves (26), uniformly over deterministic future indices. The same conditional density shows that for $s < t$, $\mathbb{P}(|W_t| = |W_s|) = 0$; consequently there are no inter-time score ties. □

For clarity, summable covariances in the central regime do not imply the indicator-FDM assumption of Section 3. To verify this for the present model, write $W_k = U_k + a_k\varepsilon_0$ and $W_k^* = U_k + a_k\varepsilon_0'$, where U_k is independent of both innovations and has variance $\sigma_W^2 - a_k^2 \ge a_0^2$ for $k \ge 1$. Conditioning on the two innovations and bounding the density of U_k gives, uniformly over thresholds v,

$$\mathbb{P}(\mathbf{1}\{|W_k| \le v\} \ne \mathbf{1}\{|W_k^*| \le v\}) \le C|a_k|.$$

For a matching lower bound at $v = q$, restrict $\varepsilon_0 \in [1, 2]$ and $\varepsilon_0' \in [-2, -1]$. For large k, $a_k > 0$ is small and the two translated intervals have a symmetric difference of length at least $2a_k$ near q. The Gaussian density of U_k is uniformly bounded below there, so the disagreement probability is at least $c_0 a_k$ for a fixed $c_0 > 0$. Thus $\theta_k \asymp |a_k|^{1/2} \asymp k^{-\beta/2}$ and $\Theta = \infty$ throughout $1/2 < \beta < 1$. The Gaussian result in Theorem 6(i) therefore uses the predictive-projection argument below, rather than the assumptions of Theorem 3.

B.1 Random-cutoff substitution

Proof of Theorem 5. Let

$$D_{a,N}(v) = F_N^{(a)}(v) - F(v) - F_N^{(a)}(q) + F(q).$$

For $|v - q| \le \delta$ in a fixed neighborhood of q, the summand is a centered even function with variance at most $C\delta$. Equation (28) therefore gives $\mathbb{E}D_{a,N}(v)^2 \le C\delta r_N^2$. Put a grid of J equal subintervals on $[q - \delta, q + \delta]$. Monotonicity of the empirical CDF and local boundedness of f bound the supremum by the maximum over grid endpoints plus $C\delta/J$. Bounding a squared maximum by the sum of squares gives

$$\mathbb{E}\sup_{|v-q|\le\delta} |D_{a,N}(v)|^2 \le C(J+1)\delta r_N^2 + C\delta^2/J^2. \tag{30}$$

Take $\delta = C_0 r_n$ for fixed C_0, and $J = \lceil r_n^{-1/2} \rceil$. For calibration $(a, N) = (0, n)$ and for the adjacent test block $(a, N) = (n, m)$ with $m/n \to \kappa$, r_N/r_n is bounded. The right side is $o(r_n^2)$. By Lemma 9, localizing to $|\widehat{q}_n - q| \le C_0 r_n$ and then letting $C_0 \to \infty$ yields

$$D_{0,n}(\widehat{q}_n) = o_{\mathbb{P}}(r_n), \qquad D_{n,m}(\widehat{q}_n) = o_{\mathbb{P}}(r_n). \tag{31}$$

For proof-local notation put $\xi_t = \mathbf{1}(V_t \le q) - (1 - \alpha)$. The exact identity $F_n^{(0)}(\widehat{q}_n) = k_n/n$, absence of ties, and $n^{-1} = o(r_n)$ now give

$$F(\widehat{q}_n) - (1 - \alpha) = -\frac{1}{n} \sum_{t=1}^{n} \xi_t + o_{\mathbb{P}}(r_n). \tag{32}$$

Taylor expansion at q, with $f(q) > 0$, proves (12). The same calculation on the test block gives the additional representation used below:

$$\widehat{\mathrm{Cov}}_{n,m} - (1 - \alpha) = \frac{1}{m} \sum_{t=n+1}^{n+m} \xi_t - \frac{1}{n} \sum_{t=1}^{n} \xi_t + o_{\mathbb{P}}(r_n). \tag{33}$$

Only marginal and local empirical bounds were used in (31); the random cutoff and test block need not be independent. □

B.2 The three limiting distributions

Proof of Theorem 6. Write $z = q/\sigma_W > 0$. For the centered indicator ξ_t, integration by parts gives the second Hermite coefficient

$$\mathbb{E}\{\mathbf{1}(|Z| \le z) H_2(Z)\} = -2z\phi(z).$$

It is nonzero and the odd coefficients vanish. Consequently

$$\xi_t = -\frac{q f(q)}{2\sigma_W^2}(W_t^2 - \sigma_W^2) + A_t, \tag{34}$$

where A_t has only even Hermite orders at least four. Orthogonality and (27) imply

$$\left\| \sum_{t=1}^{N} A_t \right\|_2^2 \le C \left\{ N + 2 \sum_{h=1}^{N-1} (N - h) |\rho(h)|^4 \right\} = o(s_{2,N}^2), \qquad \beta \le 3/4. \tag{35}$$

Indeed the bracket is $O(N^{6-8\beta})$ for $\beta < 5/8$, $O(N \log N)$ at $\beta = 5/8$, and $O(N)$ above it. Each is negligible relative to (25) in the indicated range. Combining (34) with (33) gives

$$\widehat{\mathrm{Cov}}_{n,m} - (1 - \alpha) = \frac{q f(q)}{2\sigma_W^2} \left\{ \frac{1}{n} \sum_{t=1}^{n} (W_t^2 - \sigma_W^2) - \frac{1}{m} \sum_{t=n+1}^{n+m} (W_t^2 - \sigma_W^2) \right\} + o_{\mathbb{P}}(r_n), \quad \beta \le 3/4. \tag{36}$$

If $1/2 < \beta < 3/4$, the noncentral limit belongs to the classical Gaussian Hermite theory of Dobrushin & Major (1979); Taqqu (1975). The Gaussian rank-two functional limit in Dehling & Taqqu (1989, Eq. (1.6) and Theorem 1.1), applied to the absolute-score indicator, gives the Rosenblatt limit. Its covariance hypothesis holds with $D = 2\beta - 1 < 1/2$ and slowly varying $L(h) = h^{2\beta-1}\rho(h) \to C_\gamma/\sigma_W^2 > 0$; the source only requires L to be positive eventually. Using (34)–(35) and the source's variance-one normalization, the equivalent adjacent square-sum limit is

$$\frac{1}{s_{2,n}} \left(\sum_{t=1}^{n} (W_t^2 - \sigma_W^2), \sum_{t=n+1}^{n+m} (W_t^2 - \sigma_W^2) \right) \xrightarrow{d} \big(R_H(1), R_H(1 + \kappa) - R_H(1) \big).$$

To allow $m/n \to \kappa$, apply the functional theorem on a fixed compact time interval containing the endpoints; continuity of the limit permits the varying endpoint. Equations (25) and (36) give (15).

At $\beta = 3/4$ the strict-regime source theorem does not apply. Write $T_N = \sum_{t=1}^{N} (W_t^2 - \sigma_W^2)$. Stationarity and (25) give

$$2\,\mathrm{Cov}(T_n, T_{n+m} - T_n) = s_{2,n+m}^2 - s_{2,n}^2 - s_{2,m}^2 = o(n \log n).$$

The two normalized component variances tend to $4C_\gamma^2$ and $4C_\gamma^2\kappa$. For joint normality, every Cramér–Wold combination is a quadratic form $\sum_{t=1}^{n+m} w_{n,t}(W_t^2 - \sigma_W^2)$ with $\max_t |w_{n,t}| = O\{(n\log n)^{-1/2}\}$. Let $\Gamma_N = (\gamma(i-j))_{i,j\le N}$, $N = n+m$. The row-sum bound gives $\|\Gamma_N\|_{\mathrm{op}} = O(\sqrt{N})$. Hence the eigenvalues $\lambda_{n,j}$ of $\Gamma_N^{1/2}\,\mathrm{diag}(w_{n,t})\Gamma_N^{1/2}$ satisfy $\max_j |\lambda_{n,j}| = O\{(\log n)^{-1/2}\}$. Diagonalization expresses the form as $\sum_j \lambda_{n,j}(Z_j^2 - 1)$ with independent standard normal Z_j. Its variance converges, while $\sum_j \lambda_{n,j}^4 \le (\max_j \lambda_{n,j}^2)\sum_j \lambda_{n,j}^2 \to 0$. Lyapunov's theorem proves convergence to a normal variable for positive limiting variance; a zero-variance combination converges in L^2 to zero. This proves the required joint Gaussian limit and, through (36), (14). Only a two-window joint limit, not boundary functional tightness, is asserted or needed.

Finally suppose $\beta > 3/4$. The rank-two Gaussian limit is consistent with the classical theorem of Breuer & Major (1983); we give a predictive-projection argument that also supplies the coupling used for block sampling. The full covariance sum of ξ_t is absolutely convergent by (28), and all its terms are nonnegative. Thus $\sigma_{\mathrm{cov}}^2 \ge \alpha(1-\alpha) > 0$. Lemma 11 proves $\sum_{k\ge 0} d_k < \infty$, where $d_k = \|\mathcal{P}_0\xi_k\|_2$. This is a predictive-projection condition, distinct from summable indicator FDM. To make the martingale approximation explicit, put $D_i = \sum_{j\ge 0}\mathcal{P}_i\xi_{i+j}$, converging in L^2. Then (D_i) is a stationary ergodic square-integrable martingale difference sequence. Comparing the two boundary terms at each lag and using orthogonality gives

$$\frac{1}{\sqrt{N}}\left\|\sum_{t=1}^{N}\xi_t - \sum_{i=1}^{N} D_i\right\|_2 \le 2\sum_{j\ge 0} d_j \min\{\sqrt{j/N}, 1\} \longrightarrow 0.$$

Dominated convergence proves the limit. The ergodic martingale central limit theorem gives joint Brownian increments for the two adjacent sums: ergodicity supplies convergence of the conditional variance averages, and $\mathbb{E}[D_0^2\mathbf{1}(|D_0| > \epsilon\sqrt{n})] \to 0$ supplies Lindeberg's condition. This is the standard predictive-dependence route; see Wu (2005b, Theorem 3(i)). The L^2 approximation and the convergent covariance sum identify $\mathbb{E}D_0^2 = \sigma_{\mathrm{cov}}^2$. Applying (33) proves (13). Higher even orders are not discarded. These covariance and projection arguments allow arbitrary finitely many initial coefficients in Assumption 3; their signs need not be restricted to apply the source results. □

The random-cutoff remainders above are $o_{\mathbb{P}}(r_n)$. We do not infer convergence of the actual normalized coverage variance from these distributional statements.

C Block sampling: procedure and proof

This appendix specifies the block sampling method and proves its validity with an estimated memory parameter under Theorem 7's conditions. Appendix C.1 gives the paired-block procedure and its use for inference, building on Hall et al. (1998); Zhang et al. (2013); Betken & Wendler (2018). Appendix C.2 justifies the estimated normalization; Appendix C.3 separately controls dependence between coverage blocks. Appendix C.4 combines these ingredients to prove Theorem 7 and justify inference using the estimated quantiles.

C.1 Block-sampling procedure

Paired calibration and test blocks. Choose $b = \lfloor n^\zeta \rfloor$, $0 < \zeta < 1$, a memory-estimation block length $h = \lfloor\sqrt{n}\rfloor$, and $\ell_b = \lfloor bm/n \rfloor$. For $k_b = \lceil (b+1)(1-\alpha)\rceil$, take n large enough that $k_b \le b$, $\ell_b \ge 1$, and $\max\{b+\ell_b, 2h\} \le n$. At each start $i = 0, \dots, n-b-\ell_b$, calibrate on $V_{i+1}, \dots, V_{i+b}$ using rank k_b, then evaluate the cutoff $\widehat{q}_{i,b}$ on the next ℓ_b scores. The block's realized coverage is $\widehat{\mathrm{Cov}}_{i,b} = \ell_b^{-1}\sum_{t=i+b+1}^{i+b+\ell_b}\mathbf{1}\{V_t \le \widehat{q}_{i,b}\}$. All pairs lie within the calibration set and match the target calibration–test length ratio.

Estimated normalization. We estimate memory from the signed residuals W_t, whose mean is zero under Assumption 3, while using $V_t = |W_t|$ for calibration and evaluation. The two-scale block-variance method of Zhang et al. (2013), applied at lengths $l = h, 2h$, gives

$$\widehat{Q}_l = \frac{1}{n-l+1}\sum_{i=0}^{n-l}\left(\sum_{t=i+1}^{i+l} W_t\right)^2, \quad \widehat{\beta} = \frac{3}{2} - \frac{\log(\widehat{Q}_{2h}/\widehat{Q}_h)}{2\log 2}. \tag{37}$$

Use the same normalization formula throughout $1/2 < \beta < 1$:

$$R_N(u) = \left\{ \frac{1 + \int_1^N x^{2-4u}\, dx}{N} \right\}^{1/2}, \qquad \widehat{r}_N = R_N(\widehat{\beta}), \quad N = b, n. \tag{38}$$

If either estimated block variance is zero, set $\widehat{\beta} = 3/4$. This convention is asymptotically irrelevant. Lemma 10 proves that $\widehat{r}_N / R_N(\beta) \to 1$ in probability and that $R_N(\beta)$ is asymptotically proportional to r_N.

Distribution estimation and inference. The normalized block CDF is

$$\widehat{G}_{n,b}(x) = \frac{1}{n - b - \ell_b + 1} \sum_{i=0}^{n-b-\ell_b} \mathbf{1}\left\{ \frac{\widehat{\mathrm{Cov}}_{i,b} - (1-\alpha)}{\widehat{r}_b} \le x \right\}. \tag{39}$$

It includes the unknown constants in Theorem 6, so these need not be estimated separately. Let $\widehat{c}_u$ be its u-quantile. An asymptotic two-sided level-η test rejects when $\{\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha)\}/\widehat{r}_n$ lies outside $[\widehat{c}_{\eta/2}, \widehat{c}_{1-\eta/2}]$. Multiplying these endpoints by $\widehat{r}_n$ and adding $1 - \alpha$ gives a prediction interval for future realized coverage. The probabilities are over the joint calibration and test sets; Appendix C.4 verifies these claims.

C.2 Accuracy of the memory estimate and normalization

We establish the stronger accuracy $\widehat{\beta} - \beta = o_{\mathbb{P}}(1/\log n)$, which justifies extrapolating the normalization from blocks to the full window in all three regimes.

Lemma 10 (Memory estimation and normalization)**.** *Under Assumption 3 and the refinement $a_j = j^{-\beta}\{c + O(j^{-\varphi})\}$ for some $\varphi > 0$, the estimator* (37) *with $h = \lfloor \sqrt{n} \rfloor$ satisfies, for any $0 < \delta < \min(\varphi, 1-\beta)$,*

$$\widehat{\beta} - \beta = O_{\mathbb{P}}(h^{-\delta} + v_{n,h}), \qquad v_{n,h} = \begin{cases} (h/n)^{2\beta-1}, & \beta < 3/4, \\ \sqrt{(h/n)\{1 + \log(n/h)\}}, & \beta = 3/4, \\ \sqrt{h/n}, & \beta > 3/4. \end{cases} \tag{40}$$

In particular $(\widehat{\beta} - \beta)\log n = o_{\mathbb{P}}(1)$ and $\widehat{r}_N / R_N(\beta) \to 1$ in probability for $N = b, n$. For each fixed $\beta \in (1/2, 1)$, $R_N(\beta)/r_N$ converges to a strictly positive finite constant.

Proof. This applies the two-scale method of Zhang et al. (2013, Section 2.5) to the rank-one transform $K(w) = w$, whose self-similarity index is $3/2 - \beta$ throughout $1/2 < \beta < 1$. We prove the rate directly for the observed blocks in (37).

First the coefficient refinement yields

$$\gamma(k) = C_\gamma k^{1-2\beta}\{1 + O(k^{-\delta})\}, \qquad 0 < \delta < \min(\varphi, 1-\beta). \tag{41}$$

For the leading coefficients, the sum–integral error in $\sum_{j \ge 1} j^{-\beta}(j+k)^{-\beta}$ is $O(k^{-\beta})$. Writing the coefficient error as $O(j^{-\beta-\varphi})$, and splitting its convolution at $j = k$, bounds the remaining error by $O(k^{1-2\beta-\varphi} + k^{-\beta})$, with an additional logarithm if $\beta + \varphi = 1$. The strict choice of δ absorbs this logarithm and all fixed initial coefficients. Let $S_{i,l} = \sum_{t=i+1}^{i+l} W_t$ and $s_{1,l}^2 = \mathrm{Var}(S_{0,l})$. Summing the covariances gives

$$s_{1,l}^2 = K_\beta l^{3-2\beta}\{1 + O(l^{-\delta})\}, \qquad K_\beta = \frac{2C_\gamma}{(2-2\beta)(3-2\beta)}. \tag{42}$$

Indeed the leading sum–integral error is $O(l)$, and the covariance remainder contributes $O(l^{3-2\beta-\delta})$ because $\delta < 1 - \beta < 2 - 2\beta$.

For a lag $k > 2l$, the covariance envelope in (27) gives $|\mathrm{Cov}(S_{0,l}, S_{k,l})| \le C l^2 k^{1-2\beta}$. The exact Gaussian covariance-of-squares identity therefore yields

$$\frac{\mathrm{Cov}(S_{0,l}^2, S_{k,l}^2)}{s_{1,l}^4} = 2\left\{ \frac{\mathrm{Cov}(S_{0,l}, S_{k,l})}{s_{1,l}^2} \right\}^2 \le C(l/k)^{4\beta-2}.$$

For smaller lags the normalized covariance is at most two by Cauchy–Schwarz. Hence $A_{n,l} = (n-l+1)^{-1}\sum_{i=0}^{n-l} S_{i,l}^2/s_{1,l}^2$ has expectation one and, for $l = h, 2h$,

$$\mathrm{Var}(A_{n,l}) \le C\begin{cases}(l/n)^{4\beta-2}, & \beta < 3/4,\\ (l/n)\{1+\log(n/l)\}, & \beta = 3/4,\\ l/n, & \beta > 3/4.\end{cases} \tag{43}$$

All blocks in this calculation are observed; no pre-sample blocks are introduced. Since $\widehat{Q}_l/s_{1,l}^2 = A_{n,l}$, Chebyshev's inequality gives $\widehat{Q}_l/s_{1,l}^2 = 1 + O_{\mathbb{P}}(v_{n,h})$ at both lengths. Combining with (42) gives $\widehat{Q}_{2h}/\widehat{Q}_h = 2^{3-2\beta}\{1+O_{\mathbb{P}}(h^{-\delta}+v_{n,h})\}$. Taking logarithms proves (40). Both estimated block variances are positive with probability tending to one, so the fallback does not affect the conclusion. For $h = \lfloor\sqrt{n}\rfloor$, both $h^{-\delta}\log n$ and $v_{n,h}\log n$ tend to zero.

For the normalization in (38), write $J_N(u) = NR_N(u)^2$. The integral has the closed form

$$J_N(u) = \begin{cases}1 + \{N^{3-4u}-1\}/(3-4u), & u \ne 3/4,\\ 1+\log N, & u = 3/4.\end{cases}$$

Near $u = 3/4$, evaluate the numerator as $\mathrm{expm1}\{(3-4u)\log N\}$ to avoid cancellation, using the continuous limit when $3-4u = 0$. For every real u, differentiation under the finite integral gives

$$|\partial_u \log R_N(u)| = \frac{2\int_1^N (\log x)x^{2-4u}\,dx}{1+\int_1^N x^{2-4u}\,dx} \le 2\log N.$$

The mean-value theorem and $(\widehat{\beta}-\beta)\log n = o_{\mathbb{P}}(1)$ imply $\widehat{r}_N/R_N(\beta) \to 1$ for $N = b, n$. Finally, direct evaluation gives

$$\frac{R_N(\beta)}{r_N} \longrightarrow d_\beta := \begin{cases}\sqrt{1+(4\beta-3)^{-1}}, & \beta > 3/4,\\ 1, & \beta = 3/4,\\ (3-4\beta)^{-1/2}, & \beta < 3/4.\end{cases} \tag{44}$$

In particular the unknown constant cancels from the block-to-full scale ratio: $\{\widehat{r}_n/\widehat{r}_b\}/(r_n/r_b) \to 1$ in probability. These conclusions are pointwise in the fixed model, not uniform over sequences of memory parameters approaching $3/4$. □

C.3 Dependence between block statistics

The memory-estimation lemma controls the normalization, but the overlapping coverage blocks are themselves dependent. We show that replacing distant past innovations by independent copies has a vanishing effect on their normalized statistics. These bounds will make the variance of the block CDF vanish in Appendix C.4. They concern coverage statistics, not estimation of β.

Quadratic sums for $\beta \le 3/4$. In these regimes, the leading term of the coverage indicator is quadratic, so it suffices to control centered-square sums.

Replace all innovations at times at most zero by independent copies, keeping the innovations at positive times unchanged. A star denotes the resulting process. For a separation lag $d \ge b$ and $d < t \le d+b$, write $W_t = P_t + Q_t$, where P_t contains innovations at times at most zero and Q_t contains the remaining innovations. Then $W_t^* = P_t^* + Q_t$, and the Gaussian vectors P, P^*, Q are independent. Equation (27) and the squared coefficient tail give

$$|\mathrm{Cov}(P_t, P_s)| \le Cd^{1-2\beta}, \qquad |\mathrm{Cov}(Q_t, Q_s)| \le C(1+|t-s|)^{1-2\beta}.$$

For $B_{d,b} = \sum_{t=d+1}^{d+b}(W_t^2 - \sigma_W^2)$, the copied difference is

$$B_{d,b} - B_{d,b}^* = 2\sum_{t=d+1}^{d+b} Q_t(P_t - P_t^*) + \sum_{t=d+1}^{d+b}\{P_t^2 - (P_t^*)^2\}.$$

Independence and the Gaussian square-covariance identity bound the squared norms of these two terms by $Cd^{1-2\beta}b^{3-2\beta}$ and $Cb^2d^{2-4\beta}$, respectively. Thus

$$\|B_{d,b} - B^*_{d,b}\|_2 \le C\{d^{1/2-\beta}b^{3/2-\beta} + bd^{1-2\beta}\}. \tag{45}$$

In the noncentral regime, division by $s_{2,b} \asymp b^{2-2\beta}$ gives $C(b/d)^{\beta-1/2}$. At the critical boundary, division by $\sqrt{b\log b}$ gives

$$C\frac{(b/d)^{1/4} + (b/d)^{1/2}}{\sqrt{\log b}}. \tag{46}$$

Indicator sums for $\beta > 3/4$. In this regime, higher Hermite terms also contribute to the limit, so the full coverage indicator must be controlled.

Lemma 11 (Predictive projections and copied indicator blocks)**.** *Let $\xi_t = \mathbf{1}(|W_t| \le q) - (1-\alpha)$ and $\mathcal{F}_j = \sigma(\varepsilon_s : s \le j)$. For $\mathcal{P}_j = \mathbb{E}(\cdot \mid \mathcal{F}_j) - \mathbb{E}(\cdot \mid \mathcal{F}_{j-1})$,*

$$d_k := \|\mathcal{P}_0\xi_k\|_2 \le C|a_k|\left(\sum_{j\ge k} a_j^2\right)^{1/2}, \qquad k \ge 1. \tag{47}$$

If $\beta > 3/4$, then $\sum_{k\ge 0} d_k < \infty$. In this central regime, for whole-past replacement at time zero, there is a deterministic nonincreasing $e(d) \to 0$ such that

$$\left\|\sum_{t=d+1}^{d+b}(\xi_t - \xi_t^*)\right\|_2 \le \sqrt{b}\, e(d), \qquad d, b \ge 1. \tag{48}$$

Proof. For $k \ge 1$, conditioning on $\mathcal{F}_0$ integrates out a Gaussian variable of variance $\sum_{j<k} a_j^2 \ge a_0^2$. The resulting function $g_k(x) = \mathbb{P}(|x + N(0, \sum_{j<k} a_j^2)| \le q) - (1-\alpha)$ is even with $g_k'(0) = 0$ and uniformly bounded second derivative. Put $R_k = \sum_{j>k} a_j\varepsilon_{k-j}$. Conditional centering in ε_0 and Jensen's inequality bound d_k by $\|g_k(R_k + a_k\varepsilon_0) - g_k(R_k + a_k\varepsilon_0')\|_2$. The mean-value theorem bounds this by

$$C|a_k|\,\||\varepsilon_0 - \varepsilon_0'|\{|R_k| + |a_k|(|\varepsilon_0| + |\varepsilon_0'|)\}\|_2 \le C|a_k|\left(\sum_{j\ge k} a_j^2\right)^{1/2}.$$

Thus $d_k \le Ck^{1/2-2\beta}$ eventually, proving summability in the central regime. Finitely many zero or negative coefficients do not affect the bound.

Let $D_t = \xi_t - \xi_t^*$. In the decomposition $W_t = P_t + Q_t$, Q_t is independent of P_t, P_t^* and has variance at least a_0^2. The symmetric difference of the two intervals for Q_t has total length at most $2|P_t - P_t^*|$. Its bounded density therefore gives

$$\|D_t\|_2^2 \le C\mathbb{E}|P_t - P_t^*| \le C\left(\sum_{j\ge t} a_j^2\right)^{1/2}, \qquad \sup_{t\ge d}\|D_t\|_2 \le Cd^{(1-2\beta)/4}. \tag{49}$$

Use the enlarged independent-coordinate filtration with coordinate $(\varepsilon_j, \varepsilon_j')$ at each $j \le 0$ and coordinate ε_j at each $j > 0$; both past copies enter at the same time index. Denote its projections by $\mathcal{P}_j'$. Each component of D_t has projection norm d_k at lag k, because unused innovations are independent. Orthogonal projection is also a contraction, so for $t \ge d$,

$$\|\mathcal{P}_{t-k}'D_t\|_2 \le \min\{2d_k, Cd^{(1-2\beta)/4}\}.$$

The remote past of this independent-coordinate filtration is trivial and $\mathbb{E}D_t = 0$, hence $D_t = \sum_{k\ge 0}\mathcal{P}_{t-k}'D_t$ in L^2. For fixed k, the projections indexed by t are orthogonal. Minkowski's inequality consequently gives

$$\frac{1}{\sqrt{b}}\left\|\sum_{t=d+1}^{d+b} D_t\right\|_2 \le \sum_{k\ge 0}\min\{2d_k, Cd^{(1-2\beta)/4}\} =: e(d).$$

Dominated convergence using $\sum_k d_k < \infty$ proves $e(d) \to 0$. This uses predictive projections, not summability of the single-innovation indicator FDM from the general theory. □

C.4 PROOF OF DISTRIBUTION CONSISTENCY

Proof of Theorem 7. By (44), the limit defining G is the corresponding random variable in Theorem 6 divided by d_β. The CDF G is continuous and nondegenerate in all three regimes. This is immediate for the two Gaussian limits and their positive variances. In the noncentral regime, the covariance formula $\mathrm{Cov}\{R_H(s), R_H(t)\} = (s^{2H} + t^{2H} - |t-s|^{2H})/2$ shows that the adjacent Rosenblatt contrast has variance

$$v_H(\kappa) = \frac{1+\kappa}{\kappa}\{1 + \kappa^{2H-1} - (1+\kappa)^{2H-1}\} > 0,$$

by strict concavity of x^{2H-1}. It is a nonzero second-chaos random variable. Diagonalizing its Hilbert–Schmidt kernel writes it as an L^2-convergent sum $\sum_j \lambda_j(Z_j^2 - 1)$. Separate a nonzero term from the independent remaining sum: the first has no atoms, so their convolution has no atoms either. This proves continuity of G.

Let $L_b = b + \ell_b$, $N_b = n - L_b + 1 \asymp n$ and, for this proof,

$$T_{i,b} = \frac{1}{R_b(\beta)}\left\{\frac{1}{\ell_b}\sum_{t=i+b+1}^{i+L_b}\xi_t - \frac{1}{b}\sum_{t=i+1}^{i+b}\xi_t\right\}, \qquad Y_{i,b} = \frac{\widehat{\mathrm{Cov}}_{i,b} - (1-\alpha)}{R_b(\beta)}.$$

The joint fixed-threshold limits established in Appendix B.2, applied at lengths b, ℓ_b, imply $T_{0,b} \Rightarrow G$, since $R_b(\beta)/r_b \to d_\beta$. Equation (33) at these same lengths gives

$$Y_{0,b} - T_{0,b} = o_{\mathbb{P}}(1). \tag{50}$$

Stationarity gives the same statements at every deterministic start. The exact conformal rank k_b is used in each block.

Consider two paired windows with starting points separated by $d > 2L_b$. In the later pair replace every innovation through the end of the earlier pair. The resulting $T^*_{d,b}$ is independent of $T_{0,b}$ and has the same marginal law as $T_{d,b}$. In the noncentral and critical regimes, (34)–(35) reduce each indicator sum to its quadratic component with an $o_{L^2}(br_b)$ remainder. Both subwindow lengths are comparable to b. Applying (45) or (46) to the two components therefore gives

$$\|T_{d,b} - T^*_{d,b}\|_2 \le e_b + C(b/d)^{\beta-1/2}, \qquad e_b \to 0, \quad \beta \le 3/4. \tag{51}$$

At the boundary the sharper quadratic term has the additional factor $1/\sqrt{\log b}$; the displayed weaker bound suffices. The copied reduction remainder has its original marginal norm. In the central regime Lemma 11, applied separately to the two subwindows and shifted to the copying time, instead gives

$$\|T_{d,b} - T^*_{d,b}\|_2 \le C_\kappa e(b), \qquad e(b) \to 0, \quad \beta > 3/4. \tag{52}$$

Thus all three cases have vanishing average copied discrepancy over well-separated paired windows.

For $I_{i,b}(x) = \mathbf{1}\{T_{i,b} \le x\}$ and fixed $\epsilon > 0$, copying, independence, and Markov's inequality give

$$|\,\mathrm{Cov}(I_{0,b}(x), I_{d,b}(x))| \le \mathbb{P}(|T_{0,b} - x| \le \epsilon) + \epsilon^{-1}\|T_{d,b} - T^*_{d,b}\|_2.$$

The $O(b)$ close lags have covariance bounded by one. For $\beta \le 3/4$, averaging (51) over the other lags gives $e_b + C(b/n)^{\beta-1/2} \to 0$; for $\beta > 3/4$ use (52). Consequently the variance of $\widetilde{G}_{n,b}(x) = N_b^{-1}\sum_i I_{i,b}(x)$ has limit superior at most $2\{G(x+\epsilon) - G(x-\epsilon)\}$. Letting $\epsilon \downarrow 0$ makes it zero, and its expectation converges to $G(x)$. Pointwise convergence in probability follows. A finite grid with two tail points and monotonicity of CDFs upgrades this to uniform convergence over $x \in \mathbb{R}$.

It remains to replace the oracle contrasts by actual recalibrated coverages. For each fixed $\epsilon > 0$, stationarity and (50) imply

$$\mathbb{E}\left[\frac{1}{N_b}\sum_i \mathbf{1}\{|Y_{i,b} - T_{i,b}| > \epsilon\}\right] = \mathbb{P}(|Y_{0,b} - T_{0,b}| > \epsilon) \longrightarrow 0.$$

Markov's inequality makes the fraction of such blocks negligible. Sandwiching their empirical CDF between $\widetilde{G}_{n,b}(x-\epsilon)$ and $\widetilde{G}_{n,b}(x+\epsilon)$, up to this fraction, proves uniform convergence for the $Y_{i,b}$. This does not require a maximum-over-blocks remainder bound or an L^2 random-cutoff remainder. Lemma 10 gives $\widehat{r}_b/R_b(\beta) \to 1$ in all three regimes. A common random rescaling preserves uniform CDF convergence by compact-set uniform continuity and tail control. No independence of the memory estimate and the blocks is needed. This proves the displayed conclusion of Theorem 7. □

For the test and interval in Appendix C.1, $(\widehat{\mathrm{Cov}}_{n,m} - (1-\alpha))/\widehat{r}_n \Rightarrow G$, by Theorem 6 and Lemma 10. Uniform CDF convergence places an estimated u-quantile between $G^{-1}(u-\epsilon)$ and $G^{-1}(u+\epsilon)$ with probability tending to one. Since G is continuous, these deterministic brackets give asymptotic rejection probability η for the two-sided diagnostic, and coverage probability $1-\eta$ for the future-coverage interval, by letting $\epsilon \downarrow 0$. This argument also permits flat portions of G and dependent estimated endpoints. Its probability is over the whole calibration–test trajectory. The stated null includes the residual-process assumptions and calibration procedure; a restriction on the mean coverage alone does not imply the reference law.

D Scope of a non-Gaussian extension

This appendix examines what would be needed to extend the Gaussian results to non-Gaussian innovations; it does not establish an extension. We first examine the indicator's power rank and a symmetric example, then discuss empirical-process tools, covariance calculations, and memory estimation, and finally identify the missing random-cutoff and critical-boundary arguments. The block results of Zhang et al. (2013) allow non-Gaussian innovations, but their fixed-transform conclusions alone do not cover the recalibrated statistics studied here.

For $W_t = \sum_{j\geq 0} a_j \varepsilon_{t-j}$ with a smooth marginal density f_W, the first two derivatives of the smoothed absolute indicator are

$$\frac{d}{dh}\mathbb{P}(|W_0+h|\leq q)\bigg|_{h=0} = f_W(-q) - f_W(q),$$
$$\frac{d^2}{dh^2}\mathbb{P}(|W_0+h|\leq q)\bigg|_{h=0} = f'_W(q) - f'_W(-q).$$

Symmetric innovations make W_t symmetric and cancel the first derivative for every nearby cutoff. If $f'_W(q) \neq 0$, the absolute indicator has power rank two. Symmetry alone only implies rank at least two. Without symmetry the first derivative is generally nonzero, and a rank-one normalization replaces the quadratic scale used in Section 4.

A concrete class with suitable fixed-transform regularity uses iid innovations $\varepsilon_t = S_t Z_t / \sqrt{\mathbb{E}S_t^2}$, where Z_t are standard normal, S_t takes two distinct positive values, and the two sequences are independent. All moments exist and the innovation density and its derivatives are smooth and bounded. Conditional on the scale variables, W_t is centered Gaussian with variance bounded above and away from zero. Its marginal density is consequently symmetric and satisfies $f'_W(q) < 0$ for every $q > 0$. This verifies the rank-two requirement locally as well as the moment and smoothing conditions for the fixed-transform block results.

The empirical-process expansion in Wu (2003, Eq. (4), Theorems 2–3) suggests a route to the two strict regimes. It requires moment and smoothness assumptions on an appropriate innovation convolution, including weighted integrability of its density derivatives. Subtracting the expansions at the positive and negative endpoints cancels the first-order term under symmetry. A local reduced-process bound is also available in Wu (2005a, Lemma 15(i), Eq. (57)) under its moment, density, and neighborhood-size conditions. These tools must be applied to both endpoints and checked uniformly near the cutoff; a theorem about quantiles of W_t cannot simply be applied to the nonlinear score $|W_t|$.

The Gaussian covariance identity also changes. For centered iid, variance-one innovations with finite fourth moment,

$$\mathrm{Cov}(W_0^2, W_h^2) = 2\,\mathrm{Cov}(W_0, W_h)^2 + \mathrm{cum}_4(\varepsilon_0)\sum_{j\geq 0} a_j^2 a_{j+h}^2.$$

The additional covariance sequence is absolutely summable, since $\sum_j a_j^2 < \infty$, and contributes $O(n)$ to a square partial-sum variance. It is negligible relative to $n^{4-4\beta}$ for $\beta < 3/4$, but may change central variances. In the noncentral rank-two reduction the indicator coefficient is $f'_W(q)$ rather than the Gaussian-specific $-qf(q)/(2\sigma_W^2)$. Actual recalibrated coverage blocks would include that coefficient automatically; a separate density-derivative estimator is not intrinsic to the method.

For the all-moment mixture class above, Condition 1 with $\nu = 4$ of Zhang et al. (2013) holds for the linear transform $K(w) = w$. The coefficient refinement $a_j = j^{-\beta}\{c + O(j^{-\varphi})\}$, $\varphi > 0$,

supplies their Condition 2; it is not a consequence of the innovation distribution. Their Theorem 3 and Corollary 1 then support estimation of the memory parameter from signed residuals throughout $1/2 < \beta < 1$. This step concerns the linear residuals, not the rank-two coverage statistic. The Gaussian block-moment calculation in Lemma 10 cannot simply be imported unchanged; the cited non-Gaussian moment conditions provide the relevant alternative. Neither memory estimation nor a consistent scale alone establishes a limiting law for actual recalibrated block coverages.

Completing a non-Gaussian extension still requires a local random-cutoff argument and its transfer to actual block coverages. The Gaussian even-Hermite covariance bound used in (30) is not available unchanged. The non-Gaussian critical boundary also needs a separate proof: the cited strict-regime theorems do not include it, and the Gaussian quadratic-form argument above does not transfer directly. Thus this discussion identifies a concrete extension route, not an established extension of Theorems 5–7.

E Additional simulation details

This appendix provides the assumption checks, implementation details, and additional comparisons for Section 5. Appendix E.1 verifies the nonmixing example and its FDM bound. Appendix E.2 specifies the nonlinear models and the exchangeable and corrected t comparisons. Appendix E.3 documents the long-memory experiment: Appendix E.3.1 gives its setup and memory-estimation results, Appendix E.3.2 compares block sampling with moving-block bootstrap and HAC inference, Appendix E.3.3 gives the additional $\beta = 0.7$ results.

E.1 Study 1: marginal coverage

We use 20,000 independent replications for each sample size and nominal coverage level. The stationary state U_t is uniform on $[0, 1]$. A score-threshold event $|U_t^2 - 1/3| \le v$ is an interval in U_t. Replacing one Bernoulli innovation changes U_k by at most 2^{-k-1}; crossing either endpoint therefore gives $\theta_k \le \min\{1, 2^{(1-k)/2}\}$. The score CDF is continuous, so the assumptions of Theorem 1(i) hold. In contrast, whenever $V_j > 1/3$, the state is recovered as $U_j = \sqrt{V_j + 1/3}$. Such events occur infinitely often, and $U_{t-1} = 2U_t \bmod 1$ recovers U_0 from every future score tail. A probability-$1/2$ event in V_0 belongs to both the score past and every future tail, so the strong-mixing coefficients are at least $1/4$. In particular, the model is not β-mixing.

The implementation uses a 128-bit binary state and exact integer score comparisons. It approximates the infinite-past model; the score error under the binary coupling is less than 2^{1-128}. Nonmixing is a property of the mathematical model, not a conclusion from finite-precision simulation. If $k_n > n$, we use an infinite conformal cutoff.

E.2 Study 2: realized-coverage inference for general time series

We use 10,000 independent trajectories for each model and sample size. We consider two nonlinear models: ARCH(1), $Y_t = \sqrt{0.5 + 0.5Y_{t-1}^2}\,\varepsilon_t$, and stochastic volatility (SV), $Y_t = e^{H_t/2 - 1/4}\varepsilon_t$ with $H_t = 0.9H_{t-1} + \sqrt{0.19}\,\eta_t$. The innovations are independent standard normal variables, and the score is $V_t = |Y_t|$. ARCH trajectories start from zero and discard 2,000 observations; SV trajectories start with an independent $H_0 \sim N(0, 1)$. The respective indicator FDM bounds are $\theta_k \lesssim 0.5^{k/6}$ and $\theta_k \lesssim 0.9^{k/3}$, so both models meet the stronger dependence condition in Proposition 4. The score densities are positive and continuous near the target quantiles, and inter-time ties have probability zero. Only complete calibration blocks enter the variance estimator; the full calibration set of length n determines the cutoff. The QQ plots use quantile levels $0.01, 0.02, \ldots, 0.99$. Both test levels use the same saved studentized statistics and predictive miscoverage $\alpha = 0.1$.

Both designs also meet the local conditional-density condition in Theorem 1(ii). For ARCH the conditional scale is at least $1/\sqrt{2}$, giving the folded-normal density bound $2/\sqrt{\pi}$. For SV, for every $v > 0$, $\sup_{s>0} \sqrt{2/\pi}s^{-1}\exp\{-v^2/(2s^2)\} = \sqrt{2/(\pi\mathrm{e})}/v$. Conditional mixing over scales preserves this bound on a neighborhood of the positive target quantile. The marginal densities have a positive local lower bound there.

Comparison with exchangeable and corrected t references. Table 3 compares references on the same trajectories. The exchangeable reference divides the coverage error by $\sqrt{\alpha(1-\alpha)(1/n+1/m)}$ and uses normal critical values. With $B = \lfloor n/\ell \rfloor$ complete calibration blocks, the corrected t reference uses the same studentized statistic but replaces the normal cutoff by $\sqrt{B/(B-1)}\, t_{B-1,1-\eta/2}$. This allows for variance-estimation uncertainty with few blocks; the scale factor accounts for divisor B rather than $B-1$. It is motivated by the fixed-B asymptotic reference in Appendix A.6. No statistic is recentered or empirically rescaled. At $n = 2^{19}$, the exchangeable nominal-95% reference intervals contain realized coverage in only 83.96% and 67.90% of replications for ARCH and SV, respectively.

Table 3: For Study 2, rejection rates (%) under three references, using the same saved trajectories. Columns specify the nominal test level.

		Normal		Exchangeable SE		Corrected t	
Model	n	5%	10%	5%	10%	5%	10%
ARCH	2^9	10.54	16.27	14.38	23.19	4.75	9.60
ARCH	2^{11}	9.12	14.67	15.86	24.20	5.39	10.37
ARCH	2^{13}	7.48	12.79	15.77	23.39	5.31	10.34
ARCH	2^{15}	6.40	11.81	16.15	24.21	5.08	9.95
ARCH	2^{17}	6.02	11.50	16.14	23.69	5.29	10.45
ARCH	2^{19}	5.49	10.77	16.04	24.26	5.03	10.11
SV	2^9	12.92	19.65	30.57	40.59	6.19	11.89
SV	2^{11}	8.98	14.87	31.63	40.42	5.37	10.28
SV	2^{13}	7.60	13.20	32.58	41.45	5.36	10.74
SV	2^{15}	6.03	11.25	31.45	39.71	4.89	9.86
SV	2^{17}	5.75	11.04	32.10	40.56	4.79	9.94
SV	2^{19}	5.61	10.94	32.10	39.96	5.15	10.24

E.3 Study 3: block sampling method under long memory

E.3.1 Simulation setup

For each β and sample size, we use 10,000 independent trajectories to evaluate the procedure and a separate sample of 10,000 independent trajectories to estimate the reference distribution. Stationary Gaussian trajectories are generated by FFT circulant embedding using numerically evaluated covariances $\gamma_\beta(k) = \sum_{j\geq 0}(1+j)^{-\beta}(1+j+k)^{-\beta}$. Thus the simulation approximates the infinite linear model through its covariance function, rather than truncating each trajectory's past. We use the procedure in Appendix C.1, with signed residuals $W_t = Y_t$, absolute scores $V_t = |Y_t|$, pilot length $h = \lfloor \sqrt{n} \rfloor$, and coverage-block length $b = \lfloor n^{2/3} \rfloor$. The $n^{2/3}$ choice also appears in long-memory covariance inference (Zhai et al., 2026); longer blocks capture dependence over longer spans while $b/n \to 0$. These choices of h and b are simple and satisfy our asymptotic conditions. Since the residual model has mean zero, we use uncentered block second moments to avoid the additional estimation error associated with the slowly converging sample mean under long memory. The estimate $\widehat{\beta}$ uses only the calibration set and is not clipped. It determines the same continuous normalization at $N = b, n$; the true memory parameter is not supplied to the procedure. We use equal adjacent calibration and test windows, $m = n$. Every historical pair has lengths b, b and uses its own exact conformal rank; all $n - 2b + 1$ pairs lie within the calibration set.

The independent reference sample approximates the finite-n distribution, not an imposed limiting Gaussian distribution. For each calibration history we multiply its raw block errors by $R_n(\widehat{\beta})/R_b(\widehat{\beta})$ and compare the resulting CDF with the independent CDF of full-window realized-coverage errors. No estimated normalization enters the reference CDF. Table 7 reports the median of this Kolmogorov distance over calibration histories. The QQ plots use levels $0.025, 0.05, 0.10, 0.20, \ldots, 0.90, 0.95, 0.975$. The two-sided level-$\eta$ diagnostic rejects when the observed full-window error lies strictly outside the scaled block quantiles at $\eta/2$ and $1-\eta/2$. Nominal

prediction coverage remains $1-\alpha = 0.9$ at both test levels. Table 4 summarizes the bias and RMSE of $\widehat{\beta}$. No trajectories were discarded and no normalization failures occurred.

Table 4: For Study 3, bias and root mean squared error (RMSE) of $\widehat{\beta}$. All entries are dimensionless.

$n = m$	2^{12}		2^{14}		2^{16}		2^{18}		2^{20}	
β	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.9	-0.1283	0.1452	-0.1177	0.1275	-0.1060	0.1116	-0.0943	0.0975	-0.0840	0.0859
0.75	-0.0542	0.0849	-0.0497	0.0695	-0.0436	0.0569	-0.0381	0.0472	-0.0325	0.0384
0.7	-0.0304	0.0710	-0.0296	0.0572	-0.0271	0.0466	-0.0236	0.0377	-0.0200	0.0307
0.6	0.0099	0.0607	0.0070	0.0490	0.0045	0.0398	0.0031	0.0338	0.0024	0.0291

E.3.2 Comparison with moving-block bootstrap and HAC inference

We compare block sampling with moving-block bootstrap (MBB) (Künsch, 1989) and Gaussian inference using a heteroskedasticity and autocorrelation consistent (HAC) variance estimator (Newey & West, 1987). The comparison uses the same trajectories in every setting with $\beta \in \{0.9, 0.75, 0.6\}$ and $n \in \{2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}\}$, with $m = n$. All methods estimate coverage uncertainty from the calibration set only, and use the same independent reference sample. Our block sampling procedure is unchanged and estimates β; neither benchmark uses the true memory parameter or its regime. The MBB block length and HAC bandwidth both equal the coverage-block length, $L = b = \lfloor n^{2/3} \rfloor$.

MBB samples, independently with replacement, from the $n - L + 1$ overlapping length-L blocks of calibration scores. Concatenating sampled blocks and truncating at $2n$ gives one pseudo trajectory. We recompute the exact conformal cutoff from its first n scores and evaluate coverage on the next n, allowing a sampled block to cross this boundary. Repeating this 999 times estimates the distribution of bootstrap realized coverage minus $1 - \alpha$. Bootstrap ties use the same non-strict coverage inequality as the original procedure, and empirical quantiles use the left inverse CDF.

For HAC, let $I_t = \mathbf{1}\{V_t \le \widehat{q}_n\}$ and $\bar{I} = n^{-1}\sum_{t=1}^{n} I_t$. We compute $\widehat{\gamma}_k = n^{-1}\sum_{t=1}^{n-k}(I_t - \bar{I})(I_{t+k} - \bar{I})$ and the Bartlett estimator $\widehat{v}_L = \widehat{\gamma}_0 + 2\sum_{k=1}^{L}\{1 - k/(L+1)\}\widehat{\gamma}_k$. The reference distribution is $N(0, \widehat{v}_L(1/n + 1/m))$, which includes calibration uncertainty.

Table 2 reports 5% rejection rates at $n = 2^{12}, 2^{16}, 2^{20}$; Table 5 gives the remaining settings. Table 6 gives 10% rates over the full grid. The additional $\beta = 0.7$ results use block sampling only; dashes mark benchmarks not run in this setting. The Monte Carlo unit is an independent trajectory, not an overlapping block or bootstrap draw. No trajectory failed and no HAC variance was zero.

Table 5: For Study 3, supplementary 5% rejection rates (%). The upper panel completes Table 2; the lower panel gives the additional $\beta = 0.7$ block sampling results.

$n = m$	2^{14}			2^{18}		
β	BS	MBB	HAC	BS	MBB	HAC
0.9	6.55	8.26	8.14	5.29	6.07	6.20
0.75	7.24	15.14	14.68	4.93	14.31	14.35
0.6	9.94	34.38	31.45	6.88	40.80	39.88

$n = m$	2^{12}	2^{14}	2^{16}	2^{18}	2^{20}
$\beta = 0.7$, BS	10.68	7.83	5.79	5.20	4.57

Figure 3 compares estimated coverage-error quantiles with the independent reference. Under stronger memory, the benchmark quantiles are too concentrated near zero, whereas block sampling better captures their spread. Tables 7 and 8 quantify this comparison through Kolmogorov distances and interval widths across all five sample sizes. At $\beta = 0.6$ and $n = 2^{20}$, the 10% rejection rate is 10.34% for block sampling, compared with 51.78% for MBB and 51.16% for HAC. Thus the

Table 6: For Study 3, rejection rates (%) at the 10% test level. BS denotes block sampling; all methods use length $b = \lfloor n^{2/3} \rfloor$.

$n = m$	2^{12}			2^{14}			2^{16}			2^{18}			2^{20}		
β	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC
0.9	13.40	17.13	16.82	11.16	14.26	14.16	10.14	12.07	12.14	10.19	11.28	11.36	10.12	10.88	10.64
0.75	14.22	24.45	22.81	11.29	22.27	21.63	10.60	21.91	21.64	9.78	21.37	21.37	8.56	19.80	19.82
0.7	14.44	—	—	11.99	—	—	10.26	—	—	9.46	—	—	8.37	—	—
0.6	17.18	39.81	36.17	14.35	42.17	39.67	12.48	45.26	43.96	11.32	48.22	47.72	10.34	51.78	51.16

benchmark methods substantially underestimate coverage uncertainty under strong memory, while block sampling substantially improves test calibration.

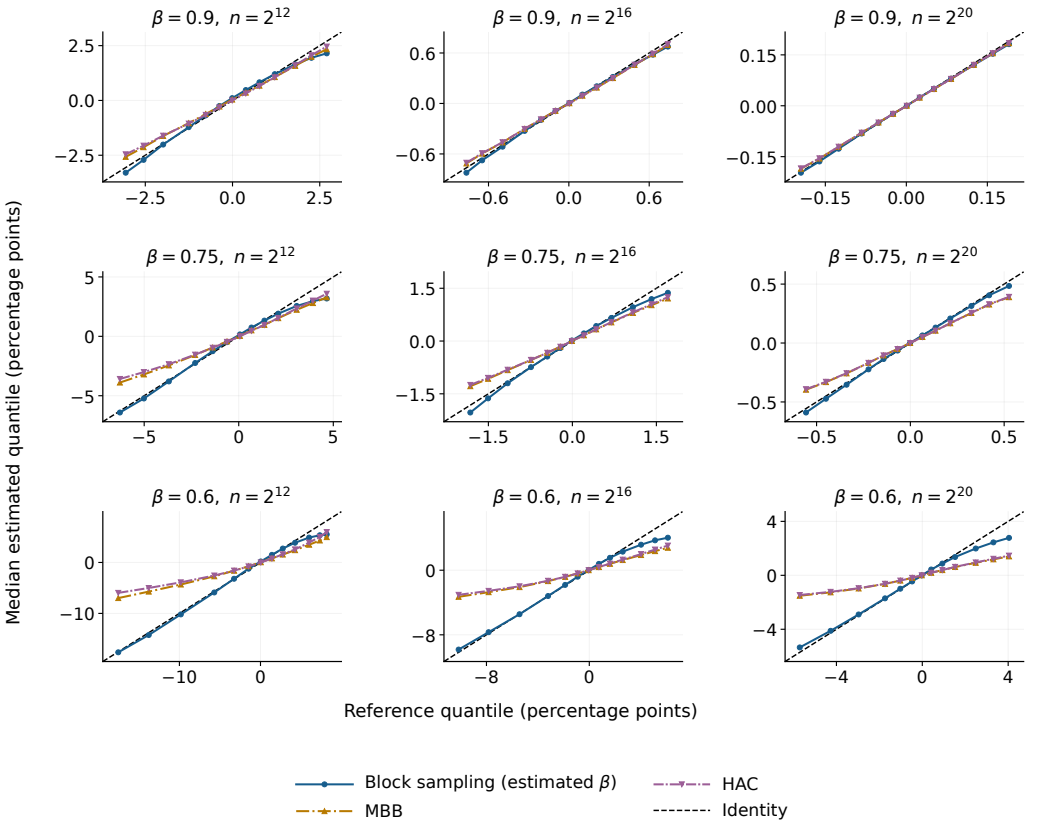

Figure 3: For Study 3, comparative QQ plots of realized-coverage errors (percentage points). Each curve gives the median estimated quantile across paired histories, against the corresponding quantile of the independent reference sample. Closeness to the dashed equality line indicates agreement with the finite-n coverage-error distribution. Rows give β and columns give n.

E.3.3 Additional ultra-long-memory results

We also study $\beta = 0.7$, another ultra-long-memory model with a non-Gaussian limit. Its coverage-error scale is $n^{-0.4}$, so its memory is less persistent than that of the $\beta = 0.6$ example in the main text. The procedure and sample sizes are unchanged. Figure 4 and Tables 5 and 6 give the full sample-size comparison. The median Kolmogorov distance decreases from 0.1467 to 0.0572; at $n = 2^{20}$, the 5% and 10% rejection rates are 4.57% and 8.37%, respectively, both below nominal. Distributional approximation improves with n for both ultra-long-memory examples, but more slowly for $\beta = 0.6$.

Table 7: For Study 3, median Kolmogorov distance from the independent finite-n reference CDF over simulation histories. BS denotes block sampling. Smaller values indicate closer distributional agreement.

$n = m$	2^{12}			2^{14}			2^{16}			2^{18}			2^{20}		
β	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC	BS	MBB	HAC
0.9	0.108	0.060	0.054	0.074	0.048	0.036	0.052	0.043	0.027	0.036	0.036	0.021	0.027	0.034	0.015
0.75	0.133	0.098	0.100	0.103	0.088	0.076	0.077	0.085	0.079	0.060	0.082	0.074	0.045	0.078	0.069
0.7	0.147	—	—	0.114	—	—	0.088	—	—	0.075	—	—	0.057	—	—
0.6	0.168	0.164	0.178	0.139	0.170	0.175	0.119	0.190	0.195	0.099	0.203	0.207	0.093	0.220	0.220

Table 8: For Study 3, median estimated 95% coverage-error interval widths over paired histories, in percentage points. BS denotes block sampling; Ref. is the 97.5th minus 2.5th quantile of the independent reference sample. Widths below Ref. indicate underestimation of coverage uncertainty.

$n = m$	$\beta = 0.9$				$\beta = 0.75$				$\beta = 0.6$			
	BS	MBB	HAC	Ref.	BS	MBB	HAC	Ref.	BS	MBB	HAC	Ref.
2^{12}	5.453	4.907	4.915	5.762	9.663	7.178	7.175	10.938	23.254	11.987	11.941	25.708
2^{14}	2.928	2.667	2.664	3.009	5.904	4.297	4.303	6.354	18.289	8.557	8.597	20.398
2^{16}	1.500	1.411	1.409	1.505	3.399	2.496	2.494	3.528	13.868	6.082	6.099	16.280
2^{18}	0.755	0.727	0.724	0.766	1.946	1.418	1.417	2.010	10.693	4.273	4.262	12.744
2^{20}	0.379	0.370	0.369	0.385	1.073	0.785	0.784	1.082	8.135	2.920	2.920	9.748

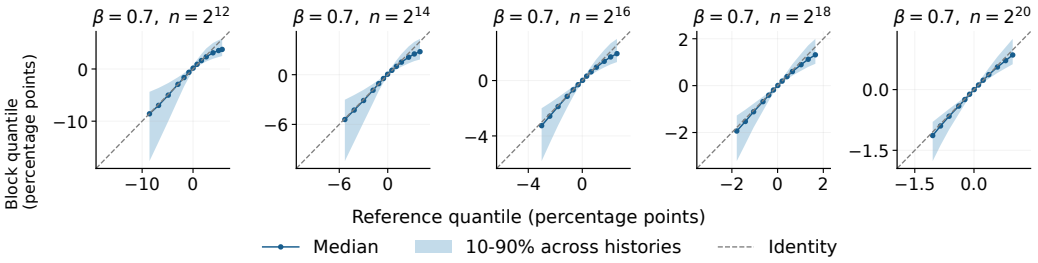

Figure 4: For Study 3, block versus independent finite-n reference error quantiles at $\beta = 0.7$ (percentage points). Curves and shading have the same interpretation as in Figure 2.

F REALIZED COVERAGE OF ELECTRICITY-DEMAND FORECAST INTERVALS

We use real-world electricity-demand data to illustrate why temporal dependence matters when deciding whether an observed departure from nominal coverage is unusual.

F.1 SETTING

This study examines realized-coverage inference for hourly electricity demand in ISO New England. Day-ahead demand forecasts inform market decisions and next-day operating plans.[1] For prediction intervals targeting 90% coverage, a practitioner may observe coverage above or below this level over the next year. The question is whether this departure indicates miscalibration or is consistent with sampling variation. We compare the exchangeable reference with our block sampling method, using the same prediction intervals throughout.

We use hourly demand and day-ahead demand forecasts from EIA's Hourly Electric Grid Monitor, selecting the reported demand and forecast fields for ISO New England.[2] The data are divided into three consecutive windows of 8,760 hours for training, calibration, and testing. Training starts on January 1, 2023, at 01:00 UTC. Calibration starts on January 1, 2024, at 01:00 UTC, and testing starts on December 31, 2024, at 01:00 UTC, ending on December 31, 2025, at 00:00 UTC. Thus $n = m = 8{,}760$, with adjacent calibration and test windows.

To account for calendar effects, we regress the training forecast errors on hour-of-day by weekday/weekend indicators and two annual sine/cosine pairs. The fitted correction is added to the published point forecast and is fixed over calibration and testing. We use the absolute corrected forecast error as the nonconformity score and calibrate at $1 - \alpha = 0.9$. The resulting cutoff is $\widehat{q}_n = 548.66$ MW; each prediction interval extends this distance on either side of the corrected forecast. The cutoff is not updated during testing.

For realized-coverage inference, we use the block sampling method in Appendix C.1, with $h = \lfloor \sqrt{n} \rfloor = 93$ for memory estimation and $b = \lfloor n^{2/3} \rfloor = 424$. All 7,913 pairs of adjacent calibration and test blocks, each of length b, lie within the calibration window. The block distribution is scaled to the full-window distribution using $\widehat{\beta} = 0.632$, estimated by the procedure in Appendix C.1. We assess realized coverage using its 2.5% and 97.5% quantiles and compare the result with the exchangeable Gaussian reference, whose standard error is $\sqrt{\alpha(1-\alpha)(1/n + 1/m)}$. Both comparisons use a two-sided 5% test level.

F.2 EVIDENCE OF LONG MEMORY

The calendar correction reduces calibration RMSE by 28.55%, but persistent dependence remains in the corrected residuals. We examine it using detrended fluctuation analysis (DFA). For a stationary process with autocorrelation proportional to k^{2H-2} at large lag k, a Hurst exponent $1/2 < H < 1$ corresponds to long memory: the autocorrelations are not summable.

We apply first-order DFA to the corrected calibration residuals (Kantelhardt et al., 2001). We use 280 overlapping windows of 2,048 hours, with starts one day apart. Within each window, we form the cumulative centered sum, remove a linear trend in each segment, and calculate the root-mean-square fluctuation $F(s)$ at 20 logarithmically spaced segment lengths from 16 to 512 hours, partitioning from both ends. The slope of the log–log fit $F(s) \propto s^H$ gives one Hurst estimate per window.

In Figure 5, all 280 window estimates exceed 0.5, with median 0.631 and range 0.518–0.683, providing diagnostic evidence consistent with long memory after calendar correction. These findings motivate accounting for persistent dependence when assessing coverage. We use the long-memory model as a working approximation and apply block sampling, which estimates coverage fluctuations across the three regimes in Theorem 7, rather than imposing the exchangeable Gaussian reference.

[1] ISO New England, *2023 Annual Markets Report*, Section 3.4.3: `https://www.iso-ne.com/static-assets/documents/100011/2023-annual-markets-report.pdf`.

[2] Data: `https://www.eia.gov/electricity/gridmonitor/`; UTC hour-ending convention: `https://www.eia.gov/survey/form/eia_930/instructions.pdf`.

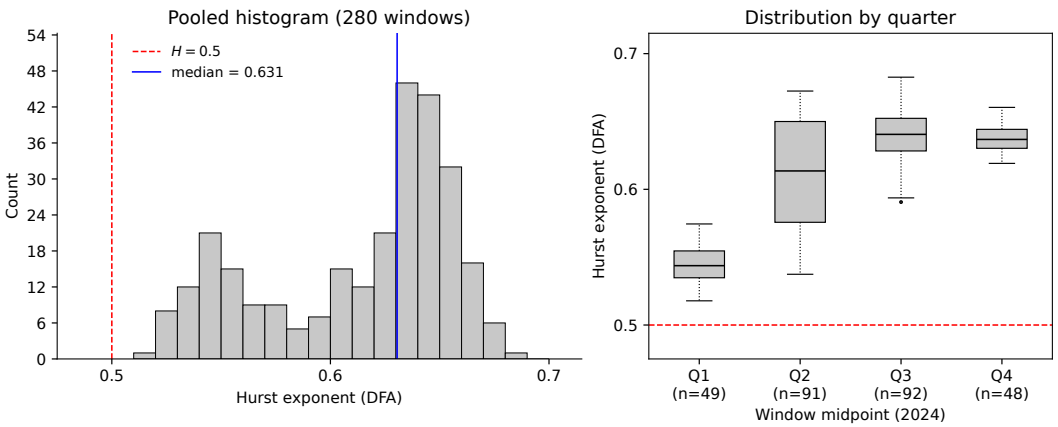

Figure 5: For the electricity-demand study, Hurst estimates from corrected calibration residuals. Left: histogram of all 280 window estimates, with the median in blue. Right: boxplots grouped by the quarter of each window's midpoint. Red dashed lines mark $H = 0.5$. Boxes show the interquartile range, with 1.5-IQR whiskers and outlying points. These are summaries of overlapping windows, not confidence intervals.

F.3 Results and interpretation

Realized coverage over the test window is 91.18%–91.20%, approximately 1.2 percentage points above nominal. The small range accounts for two missing demand observations: 7,987 of the 8,758 observed responses are covered, so the full-window fraction lies between $7987/8760$ and $7989/8760$ without imputing or deleting any observations.

Table 9: For the electricity-demand study, reference ranges for realized coverage and two-sided testing against the 90% nominal level. Decisions are unchanged for every possible outcome of the two missing responses. The reported p-values are approximate.

Reference	95% coverage range	p-value	5% test decision
Exchangeable Gaussian	[89.11%, 90.89%]	0.0082–0.0095	Reject
Block sampling	[84.36%, 93.47%]	0.558	Do not reject

Table 9 shows that the exchangeable reference rejects at the 5% level, whereas block sampling does not. Figure 6 explains the difference: block sampling assigns substantial probability to coverage values farther from 90%, giving a much wider reference range. The observed coverage is therefore compatible with the fluctuations estimated by block sampling, although it lies outside the exchangeable range.

This comparison illustrates the practical role of realized-coverage inference. The forecasts, prediction intervals, and observed coverage are identical in the two analyses; only the reference distribution changes. Accounting for temporal dependence can therefore change the conclusion of a coverage test. Here, block sampling does not flag the approximately 1.2-percentage-point overcoverage. Since our theoretical and simulation results support better-calibrated tests when temporal dependence is taken into account, we tend to believe that the block sampling test is more credible in this setting and that the observed 91.2% coverage does not provide statistically significant evidence of a departure from the 90% nominal level. The result illustrates the importance of estimating coverage uncertainty, rather than assessing a forecasting system from its realized coverage alone.

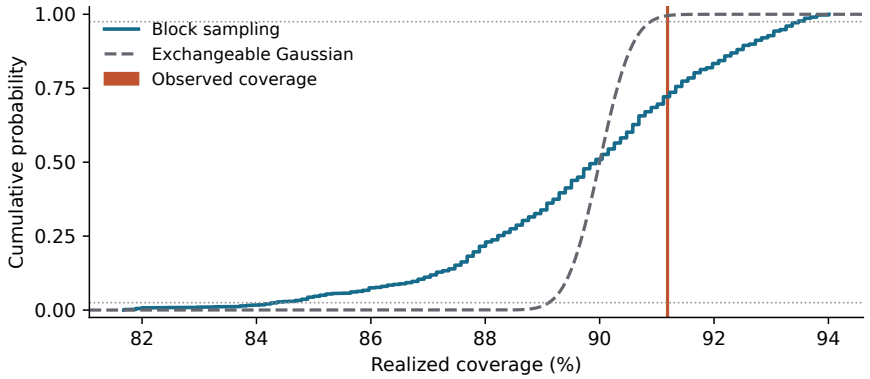

Figure 6: For the electricity-demand study, the estimated cumulative distribution of realized coverage from block sampling and the exchangeable Gaussian reference. Dotted horizontal lines mark probabilities 0.025 and 0.975. The observed coverage (orange) lies between these quantiles of the block distribution but above the exchangeable 97.5% quantile.