UNCERTAINTY QUANTIFICATION OF NEXT GENERATION RESERVOIR COMPUTING WITH APPLICATIONS TO MEMORY-DRIVEN DYNAMICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonlinear dynamical systems with memory arise across science and engineering, yet uncertainty quantification for efficient forecasting methods such as Next Generation Reservoir Computing (NGRC) remains underdeveloped. We study Bayesian ridge and conformal prediction intervals for NGRC and characterize when their uncertainty estimates agree or differ. In low dimensions, their asymptotic widths are governed by different summaries of the residual distribution, so agreement depends on residual shape rather than dimensionality alone. In high dimensions, regularization introduces a further tradeoff between estimation variance, shrinkage bias, and posterior uncertainty, leading to an explicit transition between regimes where Bayesian intervals are wider or narrower than conformal intervals. We extend these results to quadratic NGRC feature maps and give sufficient conditions for transferring the analysis to temporally dependent forecast windows. Simulations and real-data experiments support the theoretical predictions and illustrate how residual distribution, regularization, dimensionality, and distribution shift affect interval calibration and efficiency. These results provide a principled framework for choosing and interpreting uncertainty quantification methods in reservoir-based forecasting.

1 INTRODUCTION

Forecasting uncertainty for nonlinear systems with memory is important across scientific domains, including neuroscience (Aram et al., 2015), epidemiology (Pakkanen et al., 2023), materials engineering (Patlashenko et al., 2001), and finance (Han & Wong, 2021). In these settings, future behavior can depend on nonlinear interactions among past states and accumulated history. Reliable predictive uncertainty is therefore needed alongside point forecasts Gneiting et al. (2007).

Next generation reservoir computing (NGRC) provides an efficient framework for forecasting such systems. NGRC constructs delay-coordinate features from the observed process, augments them with nonlinear interactions, and estimates only a linear readout Gauthier et al. (2021); Grigoryeva et al. (2025). Its fixed nonlinear representation and linear readout make NGRC convenient for uncertainty quantification, but the resulting feature construction also creates statistical challenges. As the number of delays and nonlinear interactions increases, the feature dimension can become comparable to the available sample size, while polynomial features induce strong dependence among coordinates. Consequently, uncertainty estimates based on classical low-dimensional intuition need not remain valid.

Two natural approaches are Bayesian regression, which provides model-based predictive uncertainty, and conformal prediction (CP), which calibrates prediction intervals from observed residuals with substantially weaker distributional assumptions. Recent work has developed conformal methods for reservoir-based time-series forecasting, including Reservoir Conformal Prediction (ResCP) Neglia et al. (2026). Yet the relationship between Bayesian and conformal uncertainty quantification for NGRC remains poorly understood. In particular, it is unclear when the two methods should produce comparable intervals, what determines which interval is wider, and how regularization and increasing feature dimension change their calibration and efficiency.

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.

We study Bayesian ridge and split-conformal prediction intervals constructed from a common NGRC readout. Our analysis connects uncertainty quantification for NGRC with high-dimensional ridge regression and random-matrix theory Dobriban & Wager (2018); Marčenko & Pastur (1967), while extending the comparison to dependent quadratic features and temporally dependent forecasting settings.

Our main contributions are summarized as follows.

- **Low-dimensional characterization.** We show that Bayesian and split-conformal interval widths converge to different residual-distribution functionals, so asymptotic agreement depends on residual shape rather than low dimensionality alone.
- **High-dimensional phase characterization.** In the proportional regime, we derive a bias–variance–posterior decomposition and an explicit phase boundary determining when Bayesian intervals are wider or narrower than conformal intervals, while also characterizing Bayesian marginal coverage.
- **Nonlinear NGRC features and temporal dependence.** We extend the high-dimensional comparison to dependent quadratic NGRC features, identify when the approximation holds or fails, and give conditions for transferring the analysis to temporally dependent forecasting settings.

We conduct simulations and real-data experiments to validate the predicted regimes and examine adaptation, fitting–calibration tradeoffs, and uncertainty quantification across environmental, energy, air-quality, and financial time series.

2 Related Work

Reservoir computing and NGRC use fixed nonlinear or delay-coordinate representations with trained linear readouts Bollt (2021); Gauthier et al. (2021); Grigoryeva et al. (2025), motivating Bayesian and conformal UQ MacKay (1992); Tipping (2001); Shafer & Vovk (2008); Angelopoulos & Bates (2023); Lei et al. (2018). Burnaev & Vovk (2014) establish fixed-dimensional agreement for full conformalized ridge under standard Bayesian assumptions. We study split conformal around a common ridge predictor under residual misspecification.

High-dimensional full conformal is studied by Clarté & Zdeborová (2025); Gibbs & Candès (2025). Our focus is the same-center Bayesian–split-conformal discrepancy and dependent quadratic NGRC, building on ridge and random-matrix theory Dobriban & Wager (2018); Marčenko & Pastur (1967); Yaskov (2016). For temporal data we build on conformal methods for dependence and shift Chernozhukov et al. (2018); Xu & Xie (2023); Barber et al. (2023); Gibbs & Candes (2021); Zaffran et al. (2022), including ResCP Neglia et al. (2026). See Appendix A for details.

3 Method

We consider direct H-step forecasting of a scalar time series using an NGRC feature representation with a regularized linear readout. Bayesian and conformal uncertainty quantification are applied to the same fitted predictor.

3.1 NGRC Forecasting

Let x_t be a uniformly sampled scalar time series with response $y_t = x_{t+H}$. For delay length k and spacing s, define $z_t = (x_t, x_{t-s}, \ldots, x_{t-(k-1)s})^\top \in \mathbb{R}^k$. We use the quadratic NGRC feature map $r_t = \left(1, z_t^\top, z_{t,i} z_{t,j} : i \le j\right)^\top$ and $p = 1 + k + (k(k+1))/2$. Thus, the representation contains an intercept, delayed observations, and all unique quadratic interactions. For fit indices $\mathcal{I}_{\mathrm{fit}}$, let $R \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ denote the design matrix and response vector. The ridge readout is

$$\widehat{w} = \left(\frac{R^\top R}{n} + \lambda I_p \right)^{-1} \frac{R^\top y}{n}, \tag{1}$$

giving forecast $\widehat{y}_* = r_*^\top \widehat{w}$. We treat ridge regression as a prediction rule and stronger probabilistic assumptions are introduced only where required by the uncertainty model or theory.

3.2 Uncertainty Quantification

We compare Bayesian ridge and split conformal prediction using the same fitted NGRC readout such that differences in interval behavior are not confounded by different point predictors.

Bayesian ridge. Under $y \mid R, w \sim N(Rw, \sigma^2 I_n)$ and $w \sim N(0, \sigma^2 (n\lambda)^{-1} I_p)$, the posterior mean is equation 1 and $\Sigma_w = \sigma^2 (R^\top R + n\lambda I_p)^{-1}$. Thus $C_B(r_*) = r_*^\top \widehat{w} \pm z_{1-\alpha/2} \sqrt{\sigma^2 + r_*^\top \Sigma_w r_*}$. Outside the correctly specified Gaussian model we interpret this as a Gaussian readout interval; in the proportional theory, σ is known or consistently estimated separately.

Split conformal prediction. For a disjoint calibration set $\mathcal{I}_{\text{cal}}$ of size m, let $S_i = |y_i - r_i^\top \widehat{w}|$ and $S_{(1)} \le \cdots \le S_{(m)}$ denote the ordered scores. With $k_m = \lceil (m+1)(1-\alpha) \rceil$ and $S_{(m+1)} = +\infty$, define $\widehat{q}_{1-\alpha} = S_{(k_m)}$, and $C_C(r_*) = [r_*^\top \widehat{w} - \widehat{q}_{1-\alpha}, r_*^\top \widehat{w} + \widehat{q}_{1-\alpha}]$. Under exchangeability, $\Pr\{Y_* \in C_C(X_*)\} \ge 1 - \alpha$ Papadopoulos et al. (2002); Lei et al. (2018); this guarantee does not directly extend to general time series Chernozhukov et al. (2018); Xu & Xie (2023).

Adaptive empirical baselines. For temporal variation, we additionally evaluate time-weighted CP, which exponentially downweights older residuals Barber et al. (2023), adaptive CP, which updates the miscoverage level using recent coverage errors Gibbs & Candes (2021); Zaffran et al. (2022), and ResCP as a reservoir-based adaptive benchmark. Full update rules and implementation details are given in Appendix D.3.

4 Theory

We compare Bayesian and split-conformal intervals from fixed to proportional dimension. The assumptions differ across regimes: the fixed-dimensional result permits readout misspecification, whereas the proportional results assume exact linear readouts with Gaussian noise and stronger feature conditions. Thus the latter are high-dimensional benchmarks rather than misspecification-robust extensions of the former. Appendix C.2 summarizes the assumptions.

4.1 Fixed-Dimensional Bayesian–Conformal Comparison

Suppose $(r_t, y_t)_{t \in \mathbb{Z}}$ is strictly stationary and ergodic, with $r_t \in \mathbb{R}^p$ and fixed p, and assume $\mathbb{E}\|r_0\|^2 + \mathbb{E}y_0^2 < \infty$. Define $\Sigma = \mathbb{E}[r_0 r_0^\top]$, $w^\dagger = \Sigma^\dagger \mathbb{E}[r_0 y_0]$, $e_t = y_t - r_t^\top w^\dagger$, $\tau^2 = \mathbb{E}[e_0^2] \in (0, \infty)$, and let $\mathcal{S} = \operatorname{range}(\Sigma)$. Here $w^\dagger$ is the minimum-norm population least-squares projection. We do not assume that $\mathbb{E}[y_t \mid r_t] = r_t^\top w^\dagger$ or that Σ is nonsingular. This projection formulation is natural for NGRC, where a finite delay and polynomial feature map generally approximate rather than exactly specify the underlying dynamics.

Let $S_n = \frac{1}{n} R_n^\top R_n$, $\widehat{w}_n = (S_n + \lambda_n I_p)^{-1} \frac{R_n^\top y_n}{n}$, where $\lambda_n > 0$ and $\lambda_n \to 0$, and define $\widehat{\tau}_n^2 = \frac{1}{n} \sum_{t=1}^n (y_t - r_t^\top \widehat{w}_n)^2$. Let $z = z_{1-\alpha/2}$ and let q_e denote the $(1-\alpha)$ quantile of $|e_0|$. Assume that the distribution function of $|e_0|$ is continuous in a neighborhood of q_e and strictly crosses $1 - \alpha$ there. For calibration, let $\mathcal{I}_n = \{a_n + 1, \ldots, a_n + m_n\}$ be a deterministic contiguous block, disjoint from the observations used to fit $\widehat{w}_n$, where a_n is deterministic and $m_n \to \infty$. Define

$$W_{B,n}(r) = 2z\widehat{\tau}_n \sqrt{1 + \frac{1}{n} r^\top (S_n + \lambda_n I_p)^{-1} r},$$

and let $W_{C,n}$ denote the split-conformal width obtained from the absolute calibration residuals $|y_i - r_i^\top \widehat{w}_n|$, $i \in \mathcal{I}_n$.

Theorem 1 (Fixed-dimensional Bayesian–conformal width discrepancy)**.** *For every fixed $r \in \mathcal{S}$, $W_{B,n}(r) \xrightarrow{\mathbb{P}} 2z\tau$ and $W_{C,n} \xrightarrow{\mathbb{P}} 2q_e$. Consequently,*

$$W_{B,n}(r) - W_{C,n} \xrightarrow{\mathbb{P}} 2(z\tau - q_e).$$

In particular, the two widths are asymptotically equivalent if and only if $q_e = z\tau$.

The proof is given in Appendix C.4. Theorem 1 shows that low dimensionality alone does not imply Bayesian–conformal equivalence. The Gaussian readout interval is governed asymptotically

by the residual root-mean-square scale, whereas split conformal is governed by an absolute-residual quantile. Gaussian residuals provide an important case in which these summaries agree. The result does not require an exact conditional linear model, a nonsingular population feature covariance, or mixing of the calibration sequence. Stationarity and ergodicity are sufficient for the deterministic contiguous calibration blocks considered here.

Corollary 1 (Fixed-dimensional marginal coverage)**.** *Suppose the conditions of Theorem 1 hold, and let (r_*, y_*) be an independent draw from the stationary marginal distribution. Write $e_* = y_* - r_*^\top w^\dagger$, $F_{|e|}(x) = \mathbb{P}(|e_*| \le x)$, and assume additionally that $F_{|e|}$ is continuous at $z\tau$. Then*

$$\mathbb{P}\{y_* \in C_{B,n}(r_*)\} \longrightarrow F_{|e|}(z\tau) \;\text{ and }\; \mathbb{P}\{y_* \in C_{C,n}(r_*)\} \longrightarrow 1-\alpha.$$

In particular, if $q_e = z\tau$, the Bayesian interval also has asymptotic marginal coverage $1-\alpha$.

Corollary 1 concerns marginal coverage for an independent draw from the stationary marginal distribution; it does not claim conditional coverage given the current forecasting history. Thus, the same residual RMS–quantile mismatch that determines the limiting width difference also determines asymptotic Bayesian marginal miscalibration. An effective-dimension extension of the fixed-dimensional calculation is given in Appendix C.4.

4.2 High-Dimensional Bias, Variance, and Posterior Uncertainty

We next consider a sequence of problems for which $p = p_n$ grows with the training size n. Write $S = \frac{1}{n}R^\top R$, $A = (S+\lambda I_p)^{-1}$, and $\widehat{w} = A\frac{R^\top y}{n}$, where $\lambda > 0$ is fixed on the normalized covariance scale.

For the high-dimensional benchmark, suppose $y = Rw_0 + \varepsilon$, $\mathbb{E}[\varepsilon \mid R] = 0$, and $\mathrm{Cov}(\varepsilon \mid R) = \sigma^2 I_n$. Let r_* be an independent test feature with $\Sigma = \mathbb{E}[r_* r_*^\top]$, and let $y_* = r_*^\top w_0 + \varepsilon_*$, where ε_* is independent mean-zero test noise with variance σ^2.

Define $B_n = \lambda^2 w_0^\top A\Sigma A w_0$, $V_n = \frac{\sigma^2}{n}\operatorname{tr}(\Sigma ASA)$, and $P_n = \frac{\sigma^2}{n}\operatorname{tr}(\Sigma A)$. Here B_n is squared ridge prediction bias, V_n is frequentist estimation variance, and P_n is feature-averaged Bayesian posterior parameter uncertainty. Lemma 3 in Appendix C.5 gives the exact finite-sample identities $\mathbb{E}[(y_* - r_*^\top\widehat{w})^2 \mid R] = \sigma^2 + B_n + V_n$, and $P_n - V_n - B_n = \frac{\sigma^2\lambda}{n}\operatorname{tr}(\Sigma A^2) - B_n$.

These identities highlight the distinction between the two uncertainty calculations. After averaging over a new feature vector, Bayesian predictive variance contains noise plus posterior parameter uncertainty, whereas the frequentist prediction error contains noise, estimation variance, and ridge bias. Thus, the inequality $P_n \ge V_n$ alone does not determine the interval ordering. The prediction bias entering the conformal residual distribution can reverse it.

To translate this decomposition into interval widths, assume additionally that $\varepsilon \mid R \sim N(0, \sigma^2 I_n)$, that the calibration and test features are independent of the training sample and are i.i.d. $N(0,\Sigma)$, and that their observation noises are independent Gaussian variables. Assume that p/n is bounded, $\|\Sigma\|_{\mathrm{op}}$ and $\|w_0\|$ are uniformly bounded, and $m \to \infty$. For the conformal calculation, define the realized feature-averaged prediction-error variance $Q_n = (\widehat{w} - w_0)^\top\Sigma(\widehat{w} - w_0)$. Conditional on the fitted training sample, a fresh calibration or test residual is Gaussian with variance $\sigma^2 + Q_n$. Moreover, $\mathbb{E}(Q_n \mid R) = B_n + V_n$, and the proof in Appendix C.5 shows that $Q_n = B_n + V_n + O_{\mathbb{P}}(n^{-1/2})$. This concentration is what converts the conditional residual distribution into the finite-trace conformal width approximation below.

Theorem 2 (Finite-trace width comparison)**.** *For the common ridge center and known noise scale σ,*

$$\frac{W_{B,n}(r_*)^2}{4z^2} = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/2}), \qquad \frac{W_{C,n}^2}{4z^2} = \sigma^2 + B_n + V_n + O_{\mathbb{P}}(n^{-1/2} + m^{-1/2}).$$

Consequently,

$$\frac{W_{B,n}(r_*)^2 - W_{C,n}^2}{4z^2} = \frac{\sigma^2\lambda}{n}\operatorname{tr}(\Sigma A^2) - B_n + O_{\mathbb{P}}(n^{-1/2} + m^{-1/2}). \tag{2}$$

Equation 2 shows that either procedure may have the larger asymptotic width. The Bayesian contribution is governed by posterior parameter uncertainty, while the conformal residual scale contains

both estimation variance and ridge prediction bias. The general finite-trace comparison therefore depends on covariance-weighted traces and signal alignment and does not require isotropy.

The high-dimensional Bayesian result treats σ as known, or requires a separately justified consistent estimator. In proportional dimension, the training residual root mean square need not consistently estimate σ because shrinkage and leverage remain non-negligible; see Appendix C.5.3.

4.3 Quadratic NGRC Features

The finite-trace decomposition in Theorem 2 is a generic ridge result. Applying it to NGRC requires additional work because polynomial features constructed from the same delay vector are dependent. For a Gaussian delay vector G, coordinates such as G_jG_k and G_j^2-1 are not independent even after centering and normalization. Whitening therefore makes the coordinates uncorrelated but does not reduce the problem to the independent Gaussian feature model considered in the preceding high-dimensional benchmark.

To isolate this issue, let $G \sim N(0, I_d)$ and define the centered quadratic feature map

$$\phi_d(G) = \begin{pmatrix} G \\ \operatorname{svec}\left(\frac{GG^\top - I_d}{\sqrt{2}}\right) \end{pmatrix} \in \mathbb{R}^p, \qquad p = \frac{d(d+3)}{2},$$

where svec is an isometry from symmetric matrices under the Frobenius norm to Euclidean space. The quadratic coordinates consist of $(G_j^2-1)/\sqrt{2}$ and G_jG_k, $j<k$. Although these coordinates are dependent, $\mathbb{E}[\phi_d(G)] = 0$ and $\mathbb{E}[\phi_d(G)\phi_d(G)^\top] = I_p$.

The map ϕ_d is a population-normalized theoretical benchmark rather than the literal empirical NGRC representation, which includes an intercept and fit-sample normalization of raw lag and quadratic-product coordinates. Extending the theory to empirical normalization under dependent delays remains an open problem.

Isotropy alone is not sufficient for the preceding Gaussian-design argument. The extension requires quadratic-form concentration to control the spectral behavior of the dependent feature matrix and sufficient diffuseness of the quadratic prediction direction to obtain a Gaussian approximation for scalar prediction errors. These properties are established in Appendix C.6; the first also connects the dependent isotropic design to generalized Marchenko–Pastur behavior Yaskov (2016).

Let $r = \Sigma_p^{1/2}\phi_d(G)$, $\|\Sigma_p\|_{\mathrm{op}} \le L$, and assume $p/n \longrightarrow \gamma \in (0,\infty)$, $m \longrightarrow \infty$, and $\lambda > 0$. The rows of the training design R, the calibration features, and the test feature are independent copies of r, with independent Gaussian observation noise of variance σ^2.

For the first benchmark, consider the random-effects readout

$$w_0 \sim N\left(0, \frac{s_p^2}{p} I_p\right), \qquad y = r^\top w_0 + \varepsilon, \qquad \varepsilon \sim N(0,\sigma^2), \tag{3}$$

where s_p^2 is uniformly bounded and the readout, feature vectors, and noises are mutually independent. Let $S = R^\top R/n$, $A = (S+\lambda I_p)^{-1}$, and define $P_n = \frac{\sigma^2}{n}\operatorname{tr}(\Sigma_p A)$ and $T_n = \frac{s_p^2\lambda^2}{p}\operatorname{tr}(\Sigma_p A^2) + \frac{\sigma^2}{n}\operatorname{tr}(\Sigma_p ASA)$.

Theorem 3 (Quadratic-NGRC width approximation)**.** *Under the preceding conditions,*

$$\frac{W_{B,n}(r_*)^2}{4z^2} = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/4}), \qquad \frac{W_{C,n}^2}{4z^2} = \sigma^2 + T_n + O_{\mathbb{P}}(n^{-1/4} + m^{-1/2}).$$

Consequently,

$$\frac{W_{B,n}(r_*)^2 - W_{C,n}^2}{4z^2} = \left(\frac{\sigma^2\lambda}{n} - \frac{s_p^2\lambda^2}{p}\right)\operatorname{tr}(\Sigma_p A^2) + O_{\mathbb{P}}\left(n^{-1/4} + m^{-1/2}\right).$$

When $\Sigma_p = I_p$ and $s_p^2 \to s^2$, the same Marchenko–Pastur width limits and phase relation as in Corollary 2 hold despite the dependence among the quadratic feature coordinates. This follows from the quadratic-form concentration argument in Appendix C.6, rather than from Gaussian-row orthogonal invariance.

Theorem 3 provides the main link between the generic high-dimensional ridge comparison and quadratic NGRC. Although the features generated from a common delay vector are dependent, quadratic-form concentration controls the spectral behavior of the design, while diffuseness of the prediction direction controls the distribution of calibration and test residuals. Together these properties recover the finite-trace Bayesian and conformal width comparison for an explicit nonlinear NGRC feature model.

The random-effects assumption in equation 3 is a sufficient mechanism for obtaining a diffuse fitted prediction direction, but randomness of the underlying NGRC readout is not required. Appendix C.6.4 gives a deterministic-readout extension under the condition $\left\| \mathcal{Q}_d\left(\lambda \Sigma_p^{1/2} A w_0\right) \right\|_{\mathrm{op}} \longrightarrow 0$. Importantly, this condition concerns the quadratic component of the ridge-transformed bias direction, not the raw signal $\mathcal{Q}_d(w_0)$. The sample resolvent A and feature covariance Σ_p therefore enter the diffuseness requirement. This is a population-level sufficient condition rather than a directly observable diagnostic for a fitted NGRC model, since it depends on the unknown readout w_0 and population feature covariance. Its purpose is to identify the mechanism required by the deterministic-readout extension—diffuseness of the ridge-transformed quadratic bias direction—rather than to provide a data-driven model-selection criterion. We therefore interpret this result as an analytical benchmark for NGRC uncertainty rather than a complete calibration theory for deployed forecasting systems.

4.4 Signal-Dependent Width Ordering in Proportional Dimension

The finite-trace comparison admits a closed-form specialization under an isotropic Gaussian training design. In addition to the assumptions of Theorem 2, suppose that the rows of the training design R are i.i.d. $N(0, I_p)$ and independent of the training noises, calibration sample, and test pair. Assume further that $\Sigma = I_p$, $p/n \to \gamma \in (0, \infty)$, and $\|w_0\|^2 \to s^2$, where $w_0 = w_{0,p}$ is deterministic and $\lambda > 0$ is fixed. Define

$$I_2(\gamma, \lambda) = \int \frac{1}{(x+\lambda)^2} \, dF_\gamma^{\mathrm{MP}}(x),$$

where F_γ^{MP} is the Marchenko–Pastur law Marčenko & Pastur (1967).

Corollary 2 (Signal-dependent high-dimensional width ordering)**.** *Under the preceding conditions,*

$$W_{B,n}(r_*)^2 - W_{C,n}^2 \xrightarrow{\mathbb{P}} 4z^2 \lambda I_2(\gamma, \lambda)\big(\sigma^2 \gamma - s^2 \lambda\big).$$

Since $I_2(\gamma, \lambda) > 0$, the Bayesian interval is asymptotically wider, equal in width, or narrower than the conformal interval according as $s^2\lambda < \sigma^2\gamma$, $s^2\lambda = \sigma^2\gamma$, or $s^2\lambda > \sigma^2\gamma$.

Thus, the isotropic boundary $s^2\lambda = \sigma^2\gamma$ matches the true signal scale to that implied by the Bayesian prior and separates asymptotic overcoverage, nominal coverage, and undercoverage. This fixed-λ boundary is a specialization of the general finite-trace discrepancy, not an intrinsic failure of Bayesian inference in high dimension. Further interpretation, noise-scale estimation, and the sufficient separated-window dependence transfer are given in Appendices C.7–C.8. Data-dependent regularization requires separate analysis.

5 Simulations

We study residual-shape effects (Exp. 1), the high-dimensional phase transition (Exp. 2), quadratic-feature diffuseness and temporal dependence (Exp. 3), and robustness to variance shift (Exp. 4). Additional results appear in Appendix E. Experiments 1–2 are finite-sample diagnostics rather than estimates of asymptotic approximation rates. Unless stated otherwise, $\alpha = 0.05$, methods share the same ridge predictor, and error bars are ± 1.96 standard errors.

5.1 Experiment 1: Residual Shape and Fixed-Dimensional Agreement

We test Theorem 1 under Gaussian and non-Gaussian residual distributions. The simulations closely follow the predicted residual RMS–quantile mechanism. Gaussian errors yield nearly matching Bayesian and conformal widths with nominal Bayesian coverage, whereas non-Gaussian errors produce α-dependent width and coverage discrepancies. Split-conformal coverage remains near nominal throughout. Full settings and results are given in Appendix E.1.

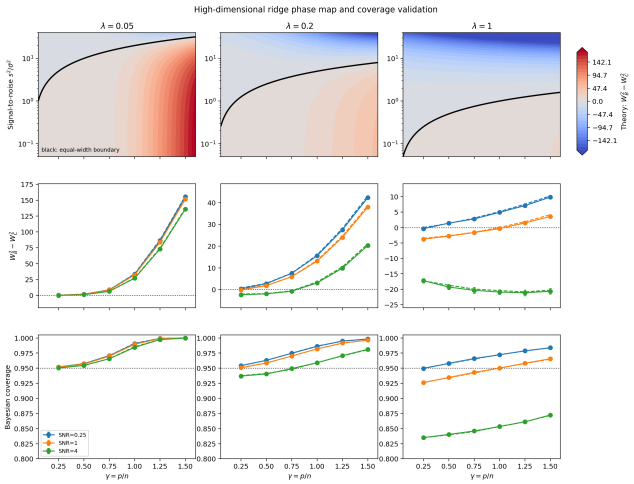


Figure 1: High-dimensional width and coverage phase transition. Top: Marchenko–Pastur limit for $W_B^2 - W_C^2$ and the equal-width boundary. Middle: empirical squared-width differences and finite-trace approximations. Bottom: empirical and theoretical Bayesian coverage; dotted lines mark 95% coverage.

5.2 Experiment 2: High-Dimensional Phase Transition and Coverage

Corollary 2 predicts a signal-dependent transition in the Bayesian–conformal width ordering and a corresponding change in Bayesian marginal coverage. We test this prediction under the isotropic Gaussian model $r_i \overset{\text{i.i.d.}}{\sim} N(0, I_p)$,$y_i = r_i^\top w_0 + \varepsilon_i$, with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. We set $p = \text{round}(\gamma n)$, and $\gamma \in \{0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$, with $\lambda \in \{0.05, 0.2, 1\}$ and $\text{SNR} = s^2/\sigma^2 \in \{0.25, 1, 4\}$, where $s^2 = \|w_0\|^2$ and $\sigma^2 = 1$. Each configuration uses $n = 400$ fitting observations, 400 calibration observations, and 1500 test observations, with results averaged over 100 replications. The Bayesian interval uses the known noise variance, and the realized ratio p/n is used in the theoretical calculations and the predicted equal-width boundary is $s^2/\sigma^2 = \gamma/\lambda$. We compare the empirical squared-width difference $W_B^2 - W_C^2$ with its finite-trace approximation and compare empirical Bayesian coverage with its asymptotic limit.

Figure 1 shows close agreement between theory and simulation across aspect ratios, ridge penalties, and signal strengths. The finite-trace approximation captures both the magnitude and sign of $W_B^2 - W_C^2$, while empirical Bayesian coverage follows the predicted calibration transition. In particular, configurations with narrower Bayesian intervals lie on the undercoverage side of the phase boundary. Individual Bayesian and conformal width approximations are verified separately in Appendix E.2.

5.3 Experiment 3: Quadratic NGRC Features, Diffuseness, and Temporal Transfer

Theorem 3 extends the high-dimensional comparison to dependent quadratic features, while Corollary 3 identifies diffuseness of the ridge-transformed quadratic bias direction as sufficient for Gaussian approximation.

For this experiment, we use the quadratic-only feature map $\psi_d(G) = \left(\frac{G_1^2-1}{\sqrt{2}}, \ldots, \frac{G_d^2-1}{\sqrt{2}}, \{G_i G_j : 1 \le i < j \le d\}\right)^\top$, and $p_q = \frac{d(d+1)}{2}$. The centered quadratic signal is generated as $G^\top M_d G - \text{tr}(M_d)$, with $M_d^{\text{conc}} = \sqrt{2}\, e_1 e_1^\top$, and $M_d^{\text{diff}} = \sqrt{2/d}\, I_d$.

Under the normalized theoretical convention $w_0^\top \psi_d(G) = (G^\top Q_d(w_0) G - \text{tr}\{Q_d(w_0)\})/\sqrt{2}$, the corresponding normalized coefficient matrices are $Q_d(w_0) = \sqrt{2}\, M_d$, and hence $Q_d^{\text{conc}}(w_0) = 2 e_1 e_1^\top$, and $Q_d^{\text{diff}}(w_0) = \frac{2}{\sqrt{d}} I_d$. Consequently, $\|w_0\|^2 = \|Q_d(w_0)\|_F^2 = 4$ for both signal configurations. The generating matrices M_d have equal squared Frobenius norm 2.

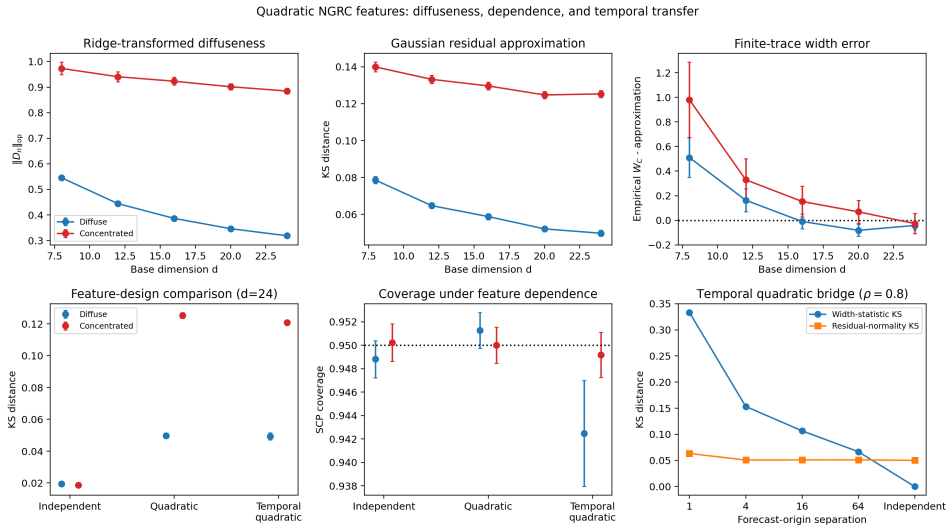


Figure 2: Quadratic NGRC diffuseness and temporal transfer. Panels show ridge-transformed diffuseness, residual Gaussianity, finite-trace width error, feature-design comparisons, SCP coverage, and convergence toward the independent-window benchmark. Error bars are ± 1.96 standard errors.

Experiment 3 uses the quadratic-only submap ψ_d, rather than the full linear-plus-quadratic map displayed in Theorem 3. Thus the simulated feature dimension is $p_q = d(d+1)/2$, whereas the full map has dimension $d(d+3)/2$. The same quadratic-form concentration and residual-approximation arguments apply after the linear block is omitted. The concentrated–diffuse deterministic construction is therefore intended primarily to diagnose the mechanism in Corollary 3, rather than to simulate the random-effects assumption of Theorem 3.

We use $d \in \{8, 12, 16, 20, 24\}$, $n_{\text{fit}} = n_{\text{cal}} = 2p_q$, $\lambda = 0.5$, $\sigma = 0.25$, and 150 replications. In each replication we record

$$\|D_n\|_{\text{op}} = \left\| Q_d\left(\lambda \Sigma_p^{1/2} A w_0\right) \right\|_{\text{op}},$$

the diffuseness diagnostic from the deterministic-readout theory. We also compare independent quadratic features with quadratic lag features from a stationary Gaussian AR(1) process. The temporal experiment uses separate stationary trajectories for fitting, calibration, and testing.

Figure 2 shows that decreasing $\|D_n\|_{\text{op}}$ is accompanied by improved residual Gaussianity, while the finite-trace width approximation can remain accurate even when the complete residual distribution is not close to Gaussian. Increasing forecast-origin separation also moves the dependent-window width distribution toward its independent-window benchmark. Smaller separations, which include overlapping windows, are empirical stress tests rather than settings covered by Proposition 1. Full settings, numerical values, and temporal-transfer diagnostics are given in Appendix E.3.2.

5.4 Experiment 4: Online Robustness Under Distribution Shift

As a robustness study, we evaluate one-step NGRC forecasting in a nonlinear Volterra system after the observation-noise standard deviation increases from 0.15 to 0.35. Frozen Bayesian and SCP intervals persistently undercover, whereas methods updated from recent forecast errors recover toward nominal coverage. An updated Gaussian baseline behaves similarly to the adaptive conformal procedures, indicating that sequential uncertainty updating is the primary mechanism. Full details are in Appendix E.4.

6 Real Data Analysis

We evaluate Bayesian ridge, SCP, adaptive CP, time-weighted CP, and ResCP on LA ozone, Solar power, Beijing PM_{10}, and Exchange returns using a chronological 40/40/20 fit/calibration/test split and leakage-safe tuning. Since $p/n_{\text{fit}} = 0.024$–0.040, these applications are closer to a low- or

Table 1: One-step NGRC results at 95% nominal coverage. Bold marks coverage closest to 0.95 and the lowest Winkler score.

Dataset	Method	Coverage	Width	Winkler
LA ozone	Bayesian ridge	0.993	2.047	2.083
	SCP	0.963	1.573	1.795
	Adaptive CP	**0.954**	1.485	1.784
	Time-weighted CP	**0.954**	1.477	**1.782**
	ResCP	0.929	1.369	1.797
Solar	Bayesian ridge	0.941	0.115	0.205
	SCP	0.940	0.115	0.205
	Adaptive CP	**0.943**	0.109	0.167
	Time-weighted CP	0.857	0.113	0.176
	ResCP	0.961	0.058	**0.075**
Beijing PM_{10}	Bayesian ridge	0.971	0.909	1.174
	SCP	0.960	0.776	1.138
	Adaptive CP	**0.947**	0.726	1.048
	Time-weighted CP	0.936	0.677	1.030
	ResCP	0.954	0.656	**0.933**
Exchange returns	Bayesian ridge	0.946	3.981	6.468
	SCP	0.983	5.407	6.997
	Adaptive CP	**0.954**	4.015	6.396
	Time-weighted CP	0.939	4.281	**6.242**
	ResCP	0.934	3.793	6.640

effective-dimensional regime than to the proportional-growth setting. However, they do not satisfy the assumptions of Theorem 1, since the empirical ridge penalties are data-selected and need not follow the vanishing-regularization sequence assumed there. We therefore use the theory only as a qualitative guide for interpreting residual shape, scale changes, regularization, and interval behavior. The residual diagnostic combines residual shape and fit-to-calibration scale change rather than isolating the stationary mechanism of Theorem 1. Full details are in Appendices D and F.1.

Table 1 shows substantial dataset dependence. Adaptive CP generally remains near nominal coverage, while ResCP is sharp for Solar and Beijing but undercovers for ozone and Exchange returns. The Bayesian–SCP width ordering agrees qualitatively with the residual-scale diagnostic. Runtime and further diagnostics are reported in Appendices D and F.

7 Conclusion

We characterize when Bayesian ridge and conformal prediction intervals agree or diverge for NGRC. In fixed dimension, their discrepancy is governed by the relationship between residual root-mean-square scale and absolute-error quantiles. In proportional dimension, posterior uncertainty, estimation variance, and ridge bias produce a signal-dependent phase transition. We extend this comparison to dependent quadratic NGRC features and give a sufficient transfer principle for temporally dependent forecast windows. Simulations and real-data experiments support these mechanisms and show that residual shape, dimensionality, regularization, and online adaptation can materially affect calibration and interval width.

Several directions remain open. First, extending the theory to multivariate responses and joint uncertainty quantification across multiple forecast horizons would require accounting for cross-coordinate and cross-horizon dependence. Second, sharper finite-sample guarantees under temporal dependence, nonstationarity, and distribution shift would better connect the independent-window theory to realistic forecasting settings. Third, the interaction between representation design and uncertainty quantification remains largely unexplored. Principled procedures for selecting lag length, nonlinear degree, seasonal features, and memory summaries using training data alone could help control both predictive error and calibration. More broadly, it would be useful to understand how these Bayesian–conformal comparisons change for richer nonlinear feature maps and other reservoir-style architectures.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and editing portions of the manuscript, generating and debugging code for simulation experiments and simulated data, and checking intermediate steps in mathematical derivations and proofs. All AI-assisted theoretical arguments were independently reviewed by the authors, and all simulation code and numerical results were verified and tested for correctness. The authors reviewed and revised all AI-generated text and retained full control over the scientific claims, methodology, interpretation, and presentation of the results. We take responsibility for the final content of this work, including any text, code, claims, or other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work is methodological and uses simulated systems, publicly available benchmark datasets, and the LA ozone data that will be made publicly available. We collect no new human-subject data, use no private or sensitive personal information, and study no applications involving direct intervention in high-stakes settings. The authors have read and adhere to the ICLR Code of Ethics and identify no substantial ethical or misuse concerns specific to this work.

REPRODUCIBILITY STATEMENT

The code to reproduce the experiments presented in this paper will be made publicly available.

AUTHOR CONTRIBUTIONS

Author 1 led the theoretical development, experiments, and evaluation of this work. Author 2 reviewed the manuscript and provided feedback on the presentation and content. Author 3 contributed to the theoretical analysis and proofs, and provided feedback on the presentation of the results and structure of the manuscript.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

P. Aram, D. R. Freestone, M. J. Cook, V. Kadirkamanathan, and D. B. Grayden. Model-based estimation of intra-cortical connectivity using electrophysiological data. *NeuroImage*, 118:563–575, September 2015. doi: 10.1016/j.neuroimage.2015.06.048.

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.

H. C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

Erik Bollt. On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(1), January 2021. ISSN 1089-7682. doi: 10.1063/5.0024890. URL http://dx.doi.org/10.1063/5.0024890.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 605–622. PMLR, 2014.

Song Chen. Beijing multi-site air quality, 2017.

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.

Lucas Clarté and Lenka Zdeborová. Building conformal prediction intervals with approximate message passing. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 798–820. PMLR, 2025.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. doi: 10.1214/17-AOS1549.

Rick Durrett. *Probability: Theory and Examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034.

Daniel J. Gauthier, Erik M. Bollt, Aaron Griffith, and Wendson A. S. Barbosa. Next generation reservoir computing. *Nature Communications*, 12:5564, 2021. doi: 10.1038/s41467-021-25801-2.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Characterizing the training-conditional coverage of full conformal inference in high dimensions, 2025.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Lyudmila Grigoryeva, Hannah Lim Jing Ting, and Juan-Pablo Ortega. Infinite-dimensional next-generation reservoir computing. *Physical Review E*, 111(3):035305, 2025. doi: 10.1103/PhysRevE.111.035305.

Bingyan Han and Hoi Ying Wong. Merton's portfolio problem under volterra heston model. *Finance Research Letters*, 39:101580, 2021. doi: 10.1016/j.frl.2020.101580.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104. ACM, 2018. doi: 10.1145/3209978.3210006.

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.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992. doi: 10.1162/neco.1992.4.3.415.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Roberto Neglia, Andrea Cini, Michael M. Bronstein, and Filippo Maria Bianchi. Rescp: Reservoir conformal prediction for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026.

Ivan Nourdin and Giovanni Peccati. *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*. Cambridge University Press, 2012.

Mikko S. Pakkanen, Xenia Miscouridou, Matthew J. Penn, Charles Whittaker, Tresnia Berah, Swapnil Mishra, Thomas A. Mellan, and Samir Bhatt. Unifying incidence and prevalence under a time-varying general branching process. *Journal of Mathematical Biology*, 87(2):35, 2023. doi: 10.1007/s00285-023-01958-w.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002. doi: 10.1007/3-540-36755-1_29.

Igor Patlashenko, Dan Givoli, and Paul Barbone. Time-stepping schemes for systems of volterra integro-differential equations. *Computer Methods in Applied Mechanics and Engineering*, 190 (43–44):5691–5718, 2001. doi: 10.1016/S0045-7825(01)00192-X.

Alvin C. Rencher and G. Bruce Schaalje. *Linear Models in Statistics*. John Wiley & Sons, 2 edition, 2008. ISBN 9780471754985. doi: 10.1002/9780470192610.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Ramon van Handel. Probability in high dimension. APC 550 Lecture Notes, Princeton University, 2016. URL `https://web.math.princeton.edu/~rvan/APC550.pdf`.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023.

Xuelin Yang, Baihe Huang, Yilong Hou, Guido Imbens, and Michael I. Jordan. Picpis: Prediction-interval-conditional prediction intervals, 2026. Manuscript, forthcoming on arXiv.

Pavel Yaskov. Necessary and sufficient conditions for the marchenko–pastur theorem. *Electronic Communications in Probability*, 21(73):1–8, 2016. doi: 10.1214/16-ECP4748.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

A Additional Related Work

Reservoir computing and NGRC. Reservoir computing (RC) combines a fixed dynamical representation with a trained linear readout. A substantial literature connects reservoir methods to nonlinear autoregressive representations and studies when fixed reservoirs provide useful nonlinear state embeddings Bollt (2021); Gauthier et al. (2021). Next generation reservoir computing (NGRC) makes this connection explicit by replacing the recurrent reservoir with delay-coordinate features and nonlinear transformations of observed states while retaining a linear output readout Gauthier et al. (2021). This explicit representation simplifies fitting and makes the readout amenable to standard regression tools. Recent work has also studied NGRC from high-dimensional and kernel perspectives, emphasizing how polynomial feature expansions interact with sample size and representation complexity Grigoryeva et al. (2025). Our focus is uncertainty quantification for this regularized readout rather than point-prediction consistency or representation design itself.

Bayesian, conformal, and fixed-dimensional ridge prediction. Because the NGRC feature map is fixed before the readout is estimated, uncertainty quantification can be introduced directly at the regression layer. Bayesian ridge regression combines a Gaussian likelihood with a shrinkage prior to obtain posterior predictive uncertainty MacKay (1992); Tipping (2001). Conformal prediction instead constructs prediction sets from conformity scores and provides finite-sample marginal validity under exchangeability Shafer & Vovk (2008); Angelopoulos & Bates (2023); Papadopoulos et al. (2002); Lei et al. (2018). Standard guarantees are marginal rather than covariate-conditional. Related work has therefore studied alternative local and conditional calibration objectives, including prediction-interval-conditional procedures Yang et al. (2026). These methods address a different inferential target from the marginal response prediction intervals studied here.

The closest classical comparison is Burnaev & Vovk (2014), who study the efficiency of conformalized ridge regression relative to Bayesian ridge prediction. Under the standard Bayesian linear-Gaussian assumptions, they show that the conformalized and Bayesian interval endpoints become asymptotically close in the classical fixed-dimensional regime. This is consistent with our fixed-dimensional result: under Gaussian residuals, the residual RMS and the corresponding absolute-error quantile agree at the Gaussian critical value. Our non-Gaussian and misspecified examples therefore do not contradict their result, because the assumptions differ.

There is also an algorithmic distinction. The conformalized ridge procedure of Burnaev & Vovk (2014) is a full-conformal construction: each candidate test response participates in the conformity-score calculation. We instead use split conformal prediction, fitting the ridge predictor on one sample and calibrating absolute residuals on a held-out sample. Consequently, our fixed-dimensional theorem should not be interpreted as extending the same conformal algorithm to weaker assumptions. Rather, it characterizes when a Bayesian ridge interval and a split conformal interval built around a common limiting predictor agree or disagree.

High-dimensional conformal regression and Bayesian comparisons. Recent work has begun to study conformal inference directly in proportional dimension. Clarté & Zdeborová (2025) develop approximate-message-passing algorithms for accelerating full conformal prediction in high-dimensional generalized linear regression. Their state-evolution analysis provides high-dimensional benchmarks for full-conformal interval size, and their synthetic experiments compare full-conformal intervals with Bayes-optimal posterior intervals under both matched and mismatched choices of loss and regularization. Their primary contribution is computational and concerns full conformal inference. Our analysis instead gives an explicit Bayesian–split-conformal width decomposition for a common ridge center and tracks posterior uncertainty, frequentist estimation variance, and regularization bias separately.

Gibbs & Candès (2025) study training-conditional coverage of full conformal regression when dimension and sample size grow proportionally. For a broad class of L_2-regularized linear regression models, they show that full conformal inference corrects the nonvanishing coverage bias that can arise from high-dimensional overfitting in simpler residual-based alternatives. Their inferential target is thus training-conditional coverage of full conformal inference. Our target is the marginal width and coverage discrepancy between Bayesian ridge and split conformal prediction using a common fitted readout. The results are therefore complementary: these works establish important high-dimensional behavior of full conformal procedures, while our theory isolates the bias–variance–posterior mechanism governing the relative widths of Bayesian and split-conformal intervals.

More broadly, high-dimensional ridge regression has been studied extensively under proportional asymptotics, $p/n \to \gamma > 0$, where prediction risk and related quantities admit deterministic random-matrix limits Dobriban & Wager (2018); Marčenko & Pastur (1967). Our finite-trace comparison uses this framework to distinguish Bayesian posterior parameter uncertainty from frequentist estimation variance and shrinkage bias, all of which can remain first order in proportional dimension.

Dependent nonlinear NGRC features. Polynomial NGRC features introduce an additional difficulty because squared lag variables and cross-products are dependent even when the underlying delay vector is Gaussian. Consequently, isotropy alone does not reduce the NGRC design to an independent-coordinate Gaussian model. Random-matrix limits can nevertheless hold for dependent-coordinate isotropic vectors under suitable concentration conditions Yaskov (2016). Our quadratic-NGRC analysis verifies the required quadratic-form concentration directly and addition-

ally identifies diffuseness of the relevant quadratic prediction direction as the condition governing Gaussian approximation of scalar forecast errors.

Conformal prediction under temporal dependence and distribution shift. Classical split conformal prediction relies on exchangeability, which generally fails for chronological time-series observations. A growing literature studies conformal prediction under temporal dependence, sequential sampling, and distribution shift Chernozhukov et al. (2018); Xu & Xie (2023); Barber et al. (2023). Adaptive conformal methods modify the effective miscoverage level according to recent coverage errors Gibbs & Candes (2021); Zaffran et al. (2022), while weighted procedures emphasize calibration observations judged more relevant to the current prediction problem Barber et al. (2023). These approaches motivate several adaptive baselines in our experiments.

Reservoir Conformal Prediction (ResCP) further uses a reservoir representation of forecast-error dynamics to assign state-dependent weights to past conformity scores Neglia et al. (2026). Our objective is not to propose a new adaptive weighting scheme. We instead study uncertainty quantification for the NGRC forecasting model itself and ask when Bayesian and split-conformal intervals agree, how dimension and regularization alter that comparison, and under what conditions an independent-window analysis can be transferred to temporally dependent forecast windows.

B Discretization and Connection to ARMA-Type Models

Memory-driven continuous-time systems can induce discrete-time dynamics with autoregressive structure and, under additional assumptions on the memory operator, moving-average components. However, discretizing a memory term expressed in past states does not produce a moving average of past innovations.

Let L denote the lag operator. Suppose that discretization of the local dynamics and the memory term yields a relation of the form $A(L)x_t = K(L)x_t + \varepsilon_t$, where $A(L)$ is the polynomial associated with the local dynamics, $K(L)$ is a linear state-memory operator and ε_t denotes the innovation term.

If the memory operator admits a rational representation $K(L) = B(L)/C(L)$, with finite-order polynomials $B(L)$ and $C(L)$, then multiplying by $C(L)$ gives $\{C(L)A(L) - B(L)\}x_t = C(L)\varepsilon_t$. After normalization, this can be written in the standard ARMA-type form $\Phi(L)x_t = \Theta(L)\varepsilon_t$, where $\Phi(L) = C(L)A(L) - B(L)$, $\Theta(L) = C(L)$. Equivalently, $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$, for suitable autoregressive and moving-average coefficients whenever $\Phi(L)$ and $\Theta(L)$ have finite degree.

The distinction is important. A finite approximation to a convolution of past states, $K(L)x_t \approx \sum_{j=1}^{J} \kappa_j x_{t-j}$, generally contributes additional lagged state terms and therefore enlarges the autoregressive component. It does not generate an MA(q) term in past innovations. The moving-average component arises only after the state-memory operator is eliminated in a representation that transfers part of the dynamics to the innovation side of the equation.

This connection provides a bridge between memory-driven dynamical systems and classical time-series models. In particular, systems with suitably structured memory kernels may admit finite-order ARMA or VARMA representations after discretization and elimination of the memory operator. More general memory kernels may instead lead to higher-order or infinite-order autoregressive representations. The NGRC framework does not require such a finite-order representation. Its delay-coordinate feature map directly models dependence on past observations and can incorporate nonlinear interactions among those lags.

C Proofs and Technical Results

Throughout the appendix, write $z = z_{1-\alpha/2}$. All limits are taken along the asymptotic sequence specified in the corresponding main-text result.

C.1 Background on Standard Technical Tools

For completeness, we briefly record references for several standard probabilistic and random-matrix tools used in the proofs. These results are not specific to the present setting; they are collected here to make the technical arguments easier to follow and to provide references for readers seeking additional background.

Ergodic theorem. If $(Z_t)_{t\in\mathbb{Z}}$ is strictly stationary and ergodic with $\mathbb{E}|Z_0| < \infty$, Birkhoff's ergodic theorem gives $\frac{1}{n}\sum_{t=1}^{n} Z_t \longrightarrow \mathbb{E}Z_0$ almost surely. The result applies componentwise to integrable finite-dimensional vector- and matrix-valued processes. We use this fact repeatedly in the fixed-dimensional analysis to obtain convergence of empirical feature moments, feature–residual cross moments, residual second moments, and calibration-block averages. See (Durrett, 2019, Theorem 6.2.1) for background on ergodic theorems.

Gaussian quadratic forms. If $Z \sim N(0, C)$ and M is symmetric, then $\mathbb{E}(Z^\top MZ) = \mathrm{tr}(MC)$ and $\mathrm{Var}(Z^\top MZ) = 2\,\mathrm{tr}\{(MC)^2\}$. Related formulas give, for deterministic b, $\mathrm{Var}\left(Z^\top MZ + 2b^\top MZ\right) = 2\,\mathrm{tr}\{(MC)^2\} + 4b^\top MCMb$. These standard Gaussian moment identities are used to control the fitted prediction error and Bayesian test-feature uncertainty; see, for example, (Rencher & Schaalje, 2008, Theorem 5.2a, Theorem 5.2c).

Empirical-CDF concentration. For i.i.d. observations with distribution function F and empirical distribution function $\widehat{F}_m$, the Dvoretzky–Kiefer–Wolfowitz–Massart inequality states that

$$\mathbb{P}\left(\sup_x |\widehat{F}_m(x) - F(x)| > \epsilon\right) \le 2e^{-2m\epsilon^2}.$$

We use this inequality after conditioning on the fitted training sample, so that the calibration residuals in the independent benchmark are conditionally i.i.d. The sharp two-sided constant is due to (Massart, 1990, Corollary 1).

Gaussian concentration and integration by parts. For $G \sim N(0, I_d)$ and a sufficiently regular function f, the Gaussian Poincaré inequality gives $\mathrm{Var}\{f(G)\} \le \mathbb{E}\|\nabla f(G)\|^2$; see (Boucheron et al., 2013, Theorem 3.20). We use this inequality to establish quadratic-form concentration for the explicit quadratic NGRC feature map. Gaussian integration by parts, $\mathbb{E}\{G_j f(G)\} = \mathbb{E}\left\{\frac{\partial f(G)}{\partial G_j}\right\}$, is given more generally in (van Handel, 2016, Lemma 6.10). Its connection with Stein's identity is discussed in (Nourdin & Peccati, 2012, Lemma 3.1.2).

Covering nets. For $\epsilon \in (0, 1)$, the Euclidean unit sphere in $\mathbb{R}^d$ admits an ϵ-net of cardinality at most $(1 + 2/\epsilon)^d$ (Vershynin, 2012, Lemma 5.2). For a symmetric matrix B and $\epsilon < 1/2$, $\|B\|_{\mathrm{op}} \le \frac{1}{1-2\epsilon}\max_{u\in\mathcal{N}_\epsilon} |u^\top Bu|$ (Vershynin, 2012, Lemma 5.4). In particular, a $1/4$-net can be chosen with cardinality at most 9^d and gives $\|B\|_{\mathrm{op}} \le 2\max_{u\in\mathcal{N}_{1/4}} |u^\top Bu|$. We use this device to control the largest quadratic direction of the fitted quadratic-NGRC prediction error.

Marchenko–Pastur limits. For an isotropic Gaussian design with $p/n \to \gamma \in (0, \infty)$, the empirical spectral distribution of the sample covariance converges to the Marchenko–Pastur law (Marčenko & Pastur, 1967, Theorem 1). Consequently, for fixed $\lambda > 0$, $\frac{1}{p}\,\mathrm{tr}(S + \lambda I)^{-1} \longrightarrow \int \frac{1}{x+\lambda}\, dF_\gamma^{\mathrm{MP}}(x)$, with analogous limits for $(S + \lambda I)^{-2}$ and $S(S + \lambda I)^{-2}$. These resolvent quantities are standard in high-dimensional ridge-regression theory; see also (Dobriban & Wager, 2018, Lemma 2.2).

For isotropic random vectors whose coordinates need not be independent, quadratic-form concentration provides a more general route to the Marchenko–Pastur law. In particular, Assumption (A1) of Yaskov (2016) requires $\frac{x_p^\top B_p x_p - \mathrm{tr}(B_p)}{p} \xrightarrow{\mathbb{P}} 0$ for every sequence of uniformly bounded positive-semidefinite B_p. Under proportional growth, (Yaskov, 2016, Theorem 3.3) shows that this condition implies the Marchenko–Pastur property.

Coupling under absolute regularity. For an absolutely regular, or β-mixing, process, Berbee's coupling principle allows a future observation block to be coupled to an independent copy with the

Table 2: Scope of the main theoretical results.

Result	Readout model	Feature/design regime	Interpretation
Thm. 1	Population least-squares projection; the conditional mean need not be linear.	Fixed p; stationary and ergodic observations under the stated moment and quantile conditions.	Allows readout misspecification.
Thm. 2	Exact linear readout $y = r^\top w_0 + \varepsilon$ with Gaussian noise for the width result.	Growing p; Gaussian calibration and test features independent of training.	General finite-trace high-dimensional benchmark.
Thm. 3	Exact linear readout in the theoretical quadratic feature representation with Gaussian noise.	Independent quadratic-Gaussian feature vectors with dependent coordinates; random-effects readout, with a deterministic diffuseness extension.	Explicit nonlinear NGRC benchmark.
Cor. 2	Exact linear Gaussian model.	Isotropic i.i.d. Gaussian training, calibration, and test designs with $p/n \to \gamma$.	Closed-form Marchenko–Pastur specialization.

same marginal distribution, with failure probability controlled by the relevant absolute-regularity coefficient. We use this result in Appendix C.8 to transfer independent-window probability statements to sufficiently separated NGRC observation windows. See (Berbee, 1979, Corollary 4.2.5).

C.2 Scope of the Theoretical Results

The theoretical results operate under deliberately different assumptions. Table 2 summarizes the main distinctions; the complete technical assumptions remain those stated with each result.

C.3 Quantile Stability

The following auxiliary result is used in the fixed-dimensional conformal argument. It shows that convergence of the calibration quantile requires only average, rather than uniform, convergence of the fitted conformity scores.

Lemma 1 (Quantile stability under average score perturbations)**.** *Let $S_1, \dots, S_m$ be estimated calibration scores and $S_1^0, \dots, S_m^0$ reference scores, with empirical distribution functions $\widehat{F}_m$ and $\widehat{F}_m^0$. Let $u = 1 - \alpha$ and $k_m = \lceil (m+1)u \rceil$. Write $S_{(1)} \le \cdots \le S_{(m)}$ for the order statistics of the estimated scores and adopt the convention $S_{(m+1)} = +\infty$. Define $\widehat{q}_m = S_{(k_m)}$. Suppose that $\widehat{F}_m^0(q) \to F(q)$ in probability at every continuity point of F, $\frac{1}{m}\sum_{i=1}^m |S_i - S_i^0| \xrightarrow{\mathbb{P}} 0$, and that F is continuous near its u-quantile q_u and strictly crosses u there. Then $\widehat{q}_m \xrightarrow{\mathbb{P}} q_u$.*

For fixed $\alpha > 0$ and $m \to \infty$, the edge case $k_m = m + 1$ eventually cannot occur, so the $+\infty$ convention does not alter the asymptotic argument below.

Proof. For $a > 0$, define $D_m(a) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{|S_i - S_i^0| > a\}$. Outside the exceptional set $\{|S_i - S_i^0| > a\}$, the implications $S_i^0 \le q - a \Longrightarrow S_i \le q$, $S_i \le q \Longrightarrow S_i^0 \le q + a$ hold. Summing the corresponding indicators and dividing by m gives the deterministic bounds $\widehat{F}_m^0(q-a) - D_m(a) \le \widehat{F}_m(q) \le \widehat{F}_m^0(q+a) + D_m(a)$. Moreover, Markov's inequality applied deterministically to the empirical average gives $D_m(a) \le \frac{1}{am}\sum_{i=1}^m |S_i - S_i^0| \xrightarrow{\mathbb{P}} 0$. Fix $t > 0$ sufficiently small. By continuity of F near q_u and the strict crossing assumption, we may choose $a \in (0, t)$ such that $F(q_u - t + a) < u < F(q_u + t - a)$. Applying the preceding empirical-CDF bounds at $q_u - t$ and $q_u + t$ gives $\widehat{F}_m(q_u - t) \le \widehat{F}_m^0(q_u - t + a) + D_m(a)$ and $\widehat{F}_m(q_u + t) \ge \widehat{F}_m^0(q_u + t - a) - D_m(a)$. Since $\widehat{F}_m^0(q) \to F(q)$ in probability at the two displayed continuity points and $D_m(a) \to 0$ in

probability $\widehat{F}_m(q_u - t) < u + o_{\mathbb{P}}(1)$, $\widehat{F}_m(q_u + t) > u + o_{\mathbb{P}}(1)$. Also, $\frac{k_m}{m} \longrightarrow u$. Therefore,

$$\mathbb{P}\left\{\widehat{F}_m(q_u - t) < \frac{k_m}{m} < \widehat{F}_m(q_u + t)\right\} \longrightarrow 1.$$

On this event, fewer than k_m observations are at or below $q_u - t$, whereas at least k_m observations are at or below $q_u + t$. By the definition of the k_mth order statistic, $q_u - t < \widehat{q}_m \leq q_u + t$. Hence $\mathbb{P}(|\widehat{q}_m - q_u| \leq t) \longrightarrow 1$. Since $t > 0$ is arbitrary, $\widehat{q}_m \xrightarrow{\mathbb{P}} q_u$. □

The same calculation also identifies an effective-dimensional extension. Let $d_{\text{eff},n} = \text{tr}\{S_n(S_n + \lambda_n I_p)^{-1}\}$. For the fitted observations,

$$\frac{1}{n}\sum_{t=1}^{n} W_{B,n}(r_t)^2 = 4z^2\widehat{\tau}_n^2\left(1 + \frac{d_{\text{eff},n}}{n}\right).$$

Thus the Bayesian parameter-uncertainty contribution also vanishes whenever $d_{\text{eff},n}/n \to 0$, provided the corresponding prediction and calibration errors are well behaved. This observation motivates the proportional regime, where the effective dimension need not be negligible.

C.4 Fixed-Dimensional Results

Lemma 2 (Projection and residual-scale consistency)**.** *Under the preceding conditions,* $\widehat{w}_n \longrightarrow w^\dagger$ *and* $\widehat{\tau}_n^2 \longrightarrow \tau^2$ *almost surely. The result does not require* Σ *to be invertible.*

Proof of Lemma 2. Let $\mathcal{S} = \text{range}(\Sigma) = \ker(\Sigma)^\perp$. If $v \in \ker(\Sigma)$, then $0 = v^\top \Sigma v = \mathbb{E}(v^\top r_0)^2$. Since $(v^\top r_0)^2 \geq 0$, it follows that $v^\top r_0 = 0$ almost surely. Thus $r_t \in \mathcal{S}$ almost surely for every t by strict stationarity. We next verify that the population projection coefficient is well defined on $\mathcal{S}$. For any $v \in \ker(\Sigma)$, $v^\top \mathbb{E}[r_0 y_0] = \mathbb{E}[(v^\top r_0)y_0] = 0$, and therefore $\mathbb{E}[r_0 y_0] \in \ker(\Sigma)^\perp = \mathcal{S}$. By definition, $w^\dagger = \Sigma^\dagger \mathbb{E}[r_0 y_0]$. The matrix $\Sigma\Sigma^\dagger$ is the orthogonal projector $P_{\mathcal{S}}$ onto $\mathcal{S}$, so $\Sigma w^\dagger = \Sigma\Sigma^\dagger \mathbb{E}[r_0 y_0] = P_{\mathcal{S}}\mathbb{E}[r_0 y_0] = \mathbb{E}[r_0 y_0]$. Writing $e_t = y_t - r_t^\top w^\dagger$, we therefore have $\mathbb{E}[r_0 e_0] = \mathbb{E}[r_0 y_0] - \Sigma w^\dagger = 0$.

Moreover, by Cauchy–Schwarz, $\mathbb{E}\|r_0 e_0\| \leq \{\mathbb{E}\|r_0\|^2\}^{1/2}\{\mathbb{E}e_0^2\}^{1/2} < \infty$. Because (r_t, y_t) is stationary and ergodic, each coordinate of $r_t r_t^\top$, $r_t e_t$, and e_t^2 is also stationary and ergodic. The ergodic theorem (Durrett, 2019, Theorem 6.2.1) therefore yields $S_n := \frac{1}{n}R_n^\top R_n \to \Sigma$, $\frac{1}{n}R_n^\top e_n \to 0$, $\frac{1}{n}\sum_{t=1}^{n} e_t^2 \to \tau^2$ almost surely. Since the dimension is fixed, entrywise convergence of S_n also implies $\|S_n - \Sigma\|_{\text{op}} \to 0$ almost surely. Every row of R_n belongs to $\mathcal{S}$ almost surely, and hence $R_n^\top y_n \in \mathcal{S}$. The ridge matrix $S_n + \lambda_n I$ leaves both $\mathcal{S}$ and $\mathcal{S}^\perp$ invariant, so $\widehat{w}_n = (S_n + \lambda_n I)^{-1}\frac{1}{n}R_n^\top y_n \in \mathcal{S}$. Also $w^\dagger \in \mathcal{S}$. Define $\Delta_n = \widehat{w}_n - w^\dagger$. Using $\frac{1}{n}R_n^\top y_n = S_n w^\dagger + \frac{1}{n}R_n^\top e_n$, we obtain

$$\begin{aligned}\Delta_n &= (S_n + \lambda_n I)^{-1}\left(S_n w^\dagger + \frac{1}{n}R_n^\top e_n\right) - w^\dagger \\ &= (S_n + \lambda_n I)^{-1}\left(\frac{1}{n}R_n^\top e_n - \lambda_n w^\dagger\right).\end{aligned}$$

Restricted to $\mathcal{S}$, the matrix Σ is positive definite. Let $c = \lambda_{\min}(\Sigma|_{\mathcal{S}}) > 0$. Since $\|S_n - \Sigma\|_{\text{op}} \to 0$ almost surely, Weyl's inequality implies that, almost surely for all sufficiently large n, $\lambda_{\min}(S_n|_{\mathcal{S}}) \geq \frac{c}{2}$. Because $\lambda_n \geq 0$, $\left\|(S_n + \lambda_n I)^{-1}|_{\mathcal{S}}\right\|_{\text{op}} \leq \frac{2}{c}$ eventually almost surely. Consequently,

$$\|\Delta_n\| \leq \frac{2}{c}\left\|\frac{1}{n}R_n^\top e_n - \lambda_n w^\dagger\right\| \longrightarrow 0$$

almost surely, since $n^{-1}R_n^\top e_n \to 0$ and $\lambda_n \to 0$. Hence $\widehat{w}_n \to w^\dagger$ almost surely. Finally, $y_n - R_n\widehat{w}_n = e_n - R_n\Delta_n$, so

$$\begin{aligned}\widehat{\tau}_n^2 &= \frac{1}{n}\|e_n - R_n\Delta_n\|^2 \\ &= \frac{1}{n}\sum_{t=1}^{n} e_t^2 - 2\Delta_n^\top\frac{1}{n}R_n^\top e_n + \Delta_n^\top S_n\Delta_n.\end{aligned}$$

The first term converges almost surely to τ^2. For the second term,

$$\left|\Delta_n^\top \frac{1}{n} R_n^\top e_n\right| \le \|\Delta_n\| \left\|\frac{1}{n} R_n^\top e_n\right\| \to 0$$

almost surely. Since $S_n \to \Sigma$, the sequence $\|S_n\|_{\mathrm{op}}$ is almost surely bounded, and therefore

$$0 \le \Delta_n^\top S_n \Delta_n \le \|S_n\|_{\mathrm{op}} \|\Delta_n\|^2 \to 0.$$

Thus $\widehat{\tau}_n^2 \to \tau^2$ almost surely. □

Proof of Theorem 1. Fix $r \in \mathcal{S}$. From the proof of Lemma 2, the restriction of $(S_n + \lambda_n I)^{-1}$ to $\mathcal{S}$ has almost surely bounded operator norm for all sufficiently large n. Hence

$$0 \le \frac{1}{n} r^\top (S_n + \lambda_n I)^{-1} r \le \frac{\|r\|^2}{n} \left\|(S_n + \lambda_n I)^{-1}|_{\mathcal{S}}\right\|_{\mathrm{op}} = O(n^{-1})$$

almost surely. Together with $\widehat{\tau}_n^2 \to \tau^2$, this gives

$$W_{B,n}(r) = 2z\widehat{\tau}_n \sqrt{1 + \frac{1}{n} r^\top (S_n + \lambda_n I)^{-1} r} \xrightarrow{\mathbb{P}} 2z\tau.$$

We next consider the conformal interval. For $i \in \mathcal{I}_{\mathrm{cal},n}$, define $S_i = |y_i - r_i^\top \widehat{w}_n|, \quad S_i^0 = |e_i| = |y_i - r_i^\top w^\dagger|$. The reverse triangle inequality gives

$$|S_i - S_i^0| \le \left|r_i^\top (\widehat{w}_n - w^\dagger)\right| \le \|r_i\| \, \|\widehat{w}_n - w^\dagger\|.$$

Therefore

$$\frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} |S_i - S_i^0| \le \|\widehat{w}_n - w^\dagger\| \frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} \|r_i\|.$$

Write the calibration block as $\mathcal{I}_{\mathrm{cal},n} = \{a_n + 1, \ldots, a_n + m_n\}$. By strict stationarity, $\frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} \|r_i\| \overset{d}{=} \frac{1}{m_n} \sum_{i=1}^{m_n} \|r_i\|$. Since $m_n \to \infty$, the ergodic theorem (Durrett, 2019, Theorem 6.2.1) gives $\frac{1}{m_n} \sum_{i=1}^{m_n} \|r_i\| \longrightarrow \mathbb{E}\|r_0\|$ almost surely.

Hence the calibration-block average satisfies $\frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} \|r_i\| \xrightarrow{\mathbb{P}} \mathbb{E}\|r_0\|$. Since $\widehat{w}_n - w^\dagger \to 0$ in probability, $\frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} |S_i - S_i^0| = o_{\mathbb{P}}(1)$. It remains to verify convergence of the empirical distribution of the reference scores. Let $F_e(q) = \mathbb{P}(|e_0| \le q)$. For every continuity point q of F_e, the process $\mathbf{1}\{|e_t| \le q\}$ is stationary, ergodic, and bounded.

Therefore $\frac{1}{m_n} \sum_{i=1}^{m_n} \mathbf{1}\{|e_i| \le q\} \longrightarrow F_e(q)$ almost surely. By stationarity, the corresponding calibration-block average has the same distribution, and hence $\frac{1}{m_n} \sum_{i \in \mathcal{I}_{\mathrm{cal},n}} \mathbf{1}\{|e_i| \le q\} \xrightarrow{\mathbb{P}} F_e(q)$. Thus the reference-score empirical CDF converges at every continuity point of F_e. Let q_e denote the $(1-\alpha)$-quantile of $|e_0|$. By the assumed continuity and strict crossing of F_e at q_e, all conditions of Lemma 1 are satisfied. Hence $\widehat{q}_{1-\alpha} \xrightarrow{\mathbb{P}} q_e$. The conformal width therefore satisfies $W_{C,n} = 2\widehat{q}_{1-\alpha} \xrightarrow{\mathbb{P}} 2q_e$. Combining the Bayesian and conformal limits gives

$$W_{B,n}(r) - W_{C,n} \xrightarrow{\mathbb{P}} 2(z\tau - q_e).$$

In particular, the two widths are asymptotically equivalent if and only if $q_e = z\tau$. □

Proof of Corollary 1. Let $e_* = y_* - r_*^\top w^\dagger$. Since $\widehat{w}_n \to w^\dagger$ in probability and $\mathbb{E}\|r_*\|^2 < \infty$, $r_*^\top (\widehat{w}_n - w^\dagger) = o_{\mathbb{P}}(1)$. Therefore $y_* - r_*^\top \widehat{w}_n = e_* + o_{\mathbb{P}}(1)$.

Theorem 1 and the same fixed-dimensional leverage argument, now applied to the random test feature r_*, give $W_{B,n}(r_*)/2 \xrightarrow{\mathbb{P}} z\tau$. By continuity of $F_{|e|}$ at $z\tau$ and Slutsky's theorem,

$$\mathbb{P}\{y_* \in C_{B,n}(r_*)\} = \mathbb{P}\left\{|y_* - r_*^\top \widehat{w}_n| \le \frac{1}{2} W_{B,n}(r_*)\right\} \longrightarrow F_{|e|}(z\tau).$$

Similarly, $W_{C,n}/2 \xrightarrow{\mathbb{P}} q_e$. The continuity and strict-crossing assumptions at q_e imply $F_{|e|}(q_e) = 1 - \alpha$. Hence $\mathbb{P}\{y_* \in C_{C,n}(r_*)\} \longrightarrow 1 - \alpha$. If $q_e = z\tau$, the Bayesian limit is also $1 - \alpha$. □

C.5 High-Dimensional Finite-Trace Results

Lemma 3 (Bias–variance–posterior decomposition)**.** *The out-of-sample prediction error satisfies*

$$\mathbb{E}[(y_* - r_*^\top \widehat{w})^2 \mid R] = \sigma^2 + B_n + V_n, \quad \textit{and} \quad P_n - V_n - B_n = \frac{\sigma^2 \lambda}{n} \operatorname{tr}(\Sigma A^2) - \lambda^2 w_0^\top A \Sigma A w_0. \tag{4}$$

For later use, define $Q_n = (\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0)$. Thus Lemma 3 implies $\mathbb{E}(Q_n \mid R) = B_n + V_n$. Theorem 2 requires the stronger fact that the realized quantity Q_n itself concentrates around this conditional mean.

C.5.1 Proof of Lemma 3

Proof. Since $A = (S + \lambda I)^{-1}$, $S = \frac{1}{n} R^\top R$, the ridge estimator satisfies $\widehat{w} = A\left(S w_0 + \frac{1}{n} R^\top \varepsilon\right)$. Using $AS = I - \lambda A$, we obtain $\widehat{w} - w_0 = -\lambda A w_0 + \xi$, $\xi = \frac{1}{n} A R^\top \varepsilon$. Conditionally on R, $\mathbb{E}(\xi \mid R) = 0$ and

$$\begin{aligned} \operatorname{Cov}(\xi \mid R) &= \frac{1}{n^2} A R^\top \operatorname{Cov}(\varepsilon) R A \\ &= \frac{\sigma^2}{n^2} A R^\top R A \\ &= \frac{\sigma^2}{n} A S A. \end{aligned}$$

Expanding the prediction-error quadratic form gives

$$(\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0) = \lambda^2 w_0^\top A \Sigma A w_0 - 2\lambda w_0^\top A \Sigma \xi + \xi^\top \Sigma \xi.$$

The conditional expectation of the cross term is zero: $\mathbb{E}[w_0^\top A \Sigma \xi \mid R] = w_0^\top A \Sigma \mathbb{E}(\xi \mid R) = 0$. For the quadratic term,

$$\begin{aligned} \mathbb{E}[\xi^\top \Sigma \xi \mid R] &= \operatorname{tr}\{\Sigma \operatorname{Cov}(\xi \mid R)\} \\ &= \frac{\sigma^2}{n} \operatorname{tr}(\Sigma A S A). \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}\left[(\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0) \,\middle|\, R\right] &= \lambda^2 w_0^\top A \Sigma A w_0 + \frac{\sigma^2}{n} \operatorname{tr}(\Sigma A S A) \\ &= B_n + V_n. \end{aligned}$$

For an independent test observation $y_* = r_*^\top w_0 + \varepsilon_*$, we have $y_* - r_*^\top \widehat{w} = r_*^\top (w_0 - \widehat{w}) + \varepsilon_*$. Conditionally on R and $\widehat{w}$, the two terms are independent, and the test noise satisfies $\mathbb{E}[\varepsilon_*] = 0$. Moreover, $\mathbb{E}\left[\{r_*^\top (w_0 - \widehat{w})\}^2 \mid R, \widehat{w}\right] = (\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0)$, while $\mathbb{E}(\varepsilon_*^2) = \sigma^2$. Taking the remaining conditional expectation gives $\mathbb{E}[(y_* - r_*^\top \widehat{w})^2 \mid R] = \sigma^2 + B_n + V_n$. Finally, since $A(S + \lambda I) = I$, we have $AS = I - \lambda A$, and, because A and S commute, $ASA = A - \lambda A^2$. Thus

$$\begin{aligned} P_n - V_n &= \frac{\sigma^2}{n} \operatorname{tr}(\Sigma A) - \frac{\sigma^2}{n} \operatorname{tr}(\Sigma A S A) \\ &= \frac{\sigma^2}{n} \operatorname{tr}\{\Sigma (A - A S A)\} \\ &= \frac{\sigma^2 \lambda}{n} \operatorname{tr}(\Sigma A^2). \end{aligned}$$

Subtracting B_n yields $P_n - V_n - B_n = \frac{\sigma^2 \lambda}{n} \operatorname{tr}(\Sigma A^2) - \lambda^2 w_0^\top A \Sigma A w_0$, which is the claimed bias–variance–posterior discrepancy. □

C.5.2 Proof of Theorem 2

Proof. By the assumed uniform bounds, choose constants $\Gamma, L, M < \infty$ such that, along the asymptotic sequence, $p/n \le \Gamma$, $\|\Sigma\|_{\mathrm{op}} \le L$, $\|w_0\| \le M$. Throughout this proof $\lambda > 0$ is fixed. Set $b = \lambda A w_0$, and $C = \frac{\sigma^2}{n} ASA$. Recall that $Q_n = (\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0)$.

From Lemma 3, $\widehat{w} - w_0 = -b + \xi, \xi \mid R \sim N(0, C)$. Hence $Q_n = b^\top \Sigma b - 2b^\top \Sigma \xi + \xi^\top \Sigma \xi$, and $\mathbb{E}(Q_n \mid R) = B_n + V_n$. Define $M_C = \Sigma^{1/2} C \Sigma^{1/2}$. For a centered Gaussian vector, the standard quadratic-form identity gives $\mathrm{Var}(\xi^\top \Sigma \xi \mid R) = 2\,\mathrm{tr}(M_C^2)$. The linear term satisfies $\mathrm{Var}(2b^\top \Sigma \xi \mid R) = 4b^\top \Sigma C \Sigma b$. Moreover, centered Gaussian odd moments vanish, so the covariance between the linear term and the centered quadratic term is zero. Therefore $\mathrm{Var}(Q_n \mid R) = 2\,\mathrm{tr}(M_C^2) + 4b^\top \Sigma C \Sigma b$.

Since S is positive semidefinite, $\|A\|_{\mathrm{op}} \le 1/\lambda$. Also, $ASA = S(S + \lambda I)^{-2}$, so

$$\|ASA\|_{\mathrm{op}} = \max_{x \in \mathrm{spec}(S)} \frac{x}{(x+\lambda)^2} \le \frac{1}{4\lambda},$$

where the scalar function $x/(x+\lambda)^2$ is maximized at $x = \lambda$. Consequently, $\|C\|_{\mathrm{op}} \le \sigma^2/4n\lambda$, and hence $\|M_C\|_{\mathrm{op}} \le \|\Sigma\|_{\mathrm{op}} \|C\|_{\mathrm{op}} \le \sigma^2 L/4n\lambda$. Similarly, $\mathrm{tr}(M_C) \le p\|M_C\|_{\mathrm{op}} \le \sigma^2 Lp/4n\lambda$. Because $\|\lambda A\|_{\mathrm{op}} \le 1$, we have $\|b\| = \|\lambda A w_0\| \le \|w_0\| \le M$, and therefore $B_n = b^\top \Sigma b \le LM^2$.

For the first variance term,

$$2\,\mathrm{tr}(M_C^2) \le 2p\|M_C\|_{\mathrm{op}}^2 \le \frac{\sigma^4 L^2 p}{8n^2\lambda^2}.$$

For the second,

$$\begin{aligned} 4b^\top \Sigma C \Sigma b &= 4(\Sigma^{1/2} b)^\top M_C (\Sigma^{1/2} b) \\ &\le 4B_n \|M_C\|_{\mathrm{op}} \\ &\le \frac{\sigma^2 L^2 M^2}{n\lambda}. \end{aligned}$$

Since p/n is uniformly bounded, $\mathrm{Var}(Q_n \mid R) = O(n^{-1})$ uniformly along the sequence. Conditional Chebyshev's inequality therefore implies $Q_n - (B_n + V_n) = O_{\mathbb{P}}(n^{-1/2})$, or equivalently, $Q_n = B_n + V_n + O_{\mathbb{P}}(n^{-1/2})$.

We now study the conformal interval. Let $\mathcal{D}_{\mathrm{tr}} = (R, y)$ denote the training sample and condition on $\mathcal{D}_{\mathrm{tr}}$. Under the Gaussian design and sample-splitting assumptions, the calibration observations are independent of $\mathcal{D}_{\mathrm{tr}}$ and are independent across calibration indices. For a calibration observation, $y_i - r_i^\top \widehat{w} = r_i^\top (w_0 - \widehat{w}) + \varepsilon_i$. Conditional on $\mathcal{D}_{\mathrm{tr}}$, the vector $w_0 - \widehat{w}$ is fixed, while $r_i \sim N(0, \Sigma)$, and $\varepsilon_i \sim N(0, \sigma^2)$ are independent. Hence $r_i^\top (w_0 - \widehat{w}) \mid \mathcal{D}_{\mathrm{tr}} \sim N(0, Q_n)$, and therefore $y_i - r_i^\top \widehat{w} \mid \mathcal{D}_{\mathrm{tr}} \sim N(0, \sigma^2 + Q_n)$. Thus the conditional CDF of the absolute residual is

$$F_{Q_n}(q) = 2\Phi\left(\frac{q}{\sqrt{\sigma^2 + Q_n}}\right) - 1, \quad q \ge 0.$$

Let $\widehat{F}_m$ denote the empirical CDF of the m calibration absolute residuals. The Dvoretzky–Kiefer–Wolfowitz–Massart inequality Massart (1990) gives, conditionally on $\mathcal{D}_{\mathrm{tr}}$,

$$\mathbb{P}\left(\sup_q |\widehat{F}_m(q) - F_{Q_n}(q)| > t \,\middle|\, \mathcal{D}_{\mathrm{tr}}\right) \le 2e^{-2mt^2}.$$

Hence $\sup_q |\widehat{F}_m(q) - F_{Q_n}(q)| = O_{\mathbb{P}}(m^{-1/2})$. The $(1-\alpha)$-quantile of F_{Q_n} is $q_n^0 = z\sqrt{\sigma^2 + Q_n}$, where $z = z_{1-\alpha/2}$. Its density is

$$f_{Q_n}(q) = \frac{2}{\sqrt{\sigma^2 + Q_n}} \phi\left(\frac{q}{\sqrt{\sigma^2 + Q_n}}\right), \quad q > 0.$$

In particular, $f_{Q_n}(q_n^0) = 2\phi(z)/\sqrt{\sigma^2 + Q_n}$. The preceding bounds imply $Q_n = O_{\mathbb{P}}(1)$, because $B_n = O(1)$ and $V_n = \mathrm{tr}(M_C) = O(1)$. Since $\sigma > 0$, it follows that, with probability tending to one, f_{Q_n} is bounded away from zero on a fixed neighborhood of q_n^0. Consequently the inverse-CDF map is locally Lipschitz there, and the empirical-CDF bound implies $\widehat{q}_{1-\alpha} = z\sqrt{\sigma^2 + Q_n} + O_{\mathbb{P}}(m^{-1/2})$.

Therefore,

$$\frac{W_{C,n}^2}{4z^2} = \frac{\widehat{q}_{1-\alpha}^2}{z^2} = \sigma^2 + Q_n + O_{\mathbb{P}}(m^{-1/2}),$$

where the last step uses $Q_n = O_{\mathbb{P}}(1)$. Substituting the expansion for Q_n gives

$$\frac{W_{C,n}^2}{4z^2} = \sigma^2 + B_n + V_n + O_{\mathbb{P}}(n^{-1/2} + m^{-1/2}).$$

We next consider the Bayesian interval. Conditional on R, the independent test feature satisfies $r_* \sim N(0, \Sigma)$, and hence $\mathbb{E}(r_*^\top A r_* \mid R) = \operatorname{tr}(\Sigma A)$. Therefore $\mathbb{E}\left[\frac{\sigma^2}{n} r_*^\top A r_* \,\middle|\, R\right] = \frac{\sigma^2}{n}\operatorname{tr}(\Sigma A) = P_n$. The Gaussian quadratic-form variance identity gives

$$\begin{aligned}
\operatorname{Var}\left(\frac{\sigma^2}{n} r_*^\top A r_* \,\middle|\, R\right) &= \frac{2\sigma^4}{n^2}\operatorname{tr}\left[(\Sigma^{1/2} A \Sigma^{1/2})^2\right] \\
&\le \frac{2\sigma^4 p}{n^2}\|\Sigma^{1/2} A \Sigma^{1/2}\|_{\mathrm{op}}^2 \\
&\le \frac{2\sigma^4 L^2 p}{n^2 \lambda^2} \\
&= O(n^{-1}).
\end{aligned}$$

Thus $\frac{\sigma^2}{n} r_*^\top A r_* = P_n + O_{\mathbb{P}}(n^{-1/2})$. Since the Bayesian predictive variance equals $\sigma^2 + \frac{\sigma^2}{n} r_*^\top A r_*$, we obtain

$$\frac{W_{B,n}(r_*)^2}{4z^2} = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/2}).$$

Finally,

$$\frac{W_{B,n}(r_*)^2 - W_{C,n}^2}{4z^2} = P_n - V_n - B_n + O_{\mathbb{P}}(n^{-1/2} + m^{-1/2}),$$

and Lemma 3 gives

$$P_n - V_n - B_n = \frac{\sigma^2 \lambda}{n}\operatorname{tr}(\Sigma A^2) - \lambda^2 w_0^\top A \Sigma A w_0.$$

□

Non-isotropic feature covariances. The isotropic result gives the clearest closed-form phase boundary, but the underlying finite-trace decomposition is more general. For non-isotropic features, the relevant posterior and frequentist estimation-variance terms are the covariance-weighted traces $\operatorname{tr}(\Sigma A)/n$ and $\operatorname{tr}(\Sigma A S A)/n$, whose difference is $\lambda \operatorname{tr}(\Sigma A^2)/n$.

These quantities cannot generally be replaced by unweighted integrals against the empirical spectral distribution of S. Moreover, the ridge-bias term $w_0^\top A \Sigma A w_0$ depends on the alignment of the signal with the eigenspaces of Σ. Consequently, the feature spectrum alone does not determine the Bayesian–conformal width ordering outside the isotropic setting. The general finite-trace comparison, rather than the scalar boundary $s^2\lambda = \sigma^2\gamma$, is therefore the appropriate statement for anisotropic designs.

C.5.3 Training residual scale in proportional dimension.

Let $\mathsf{H} = \frac{1}{n} R A R^\top$. Since $R\widehat{w} = \mathsf{H} y$, $y - R\widehat{w} = (I_n - \mathsf{H}) R w_0 + (I_n - \mathsf{H})\varepsilon$. Using $(I_n - \mathsf{H})R = \lambda R A$ and conditioning on R gives

$$\mathbb{E}\left[\frac{\|y - R\widehat{w}\|^2}{n} \,\middle|\, R\right] = \lambda^2 w_0^\top A S A w_0 + \frac{\sigma^2}{n}\operatorname{tr}(I_n - \mathsf{H})^2.$$

The resulting in-sample residual scale generally contains nonvanishing shrinkage and leverage effects and therefore need not converge to σ^2, which is why the high-dimensional Bayesian results assume known σ or a separately justified consistent estimator.

C.6 Proofs for the Quadratic NGRC Results

The proof of Theorem 3 uses two auxiliary results. The first verifies quadratic-form concentration despite dependence among the nonlinear feature coordinates; the second controls the Gaussian approximation of a linear-plus-quadratic Gaussian functional.

C.6.1 Quadratic-Form Concentration

Lemma 4. *Let* $G \sim N(0, I_d)$, $\phi_d(G) = \begin{pmatrix} G \\ \operatorname{svec}\left(\frac{GG^\top - I_d}{\sqrt{2}}\right) \end{pmatrix}$, *and* $p = \frac{d(d+3)}{2}$, *where* svec *is an isometry from symmetric matrices under the Frobenius norm to Euclidean vectors. Then* $\mathbb{E}\phi_d(G) = 0$, *and* $\mathbb{E}[\phi_d(G)\phi_d(G)^\top] = I_p$. *For every symmetric* $B \in \mathbb{R}^{p\times p}$,

$$\operatorname{Var}\{\phi_d(G)^\top B\phi_d(G)\} \le (4d^3 + 30d^2 + 38d)\|B\|_{\mathrm{op}}^2 \le 72d^3\|B\|_{\mathrm{op}}^2,$$

and consequently

$$\mathbb{E}\left[\left\{\frac{\phi_d(G)^\top B\phi_d(G) - \operatorname{tr}(B)}{p}\right\}^2\right] \le \frac{288}{d}\|B\|_{\mathrm{op}}^2.$$

Moreover, $\mathbb{E}\|\phi_d(G)\|^4 \le 7p^2$.

Proof. The mean and covariance identities follow from Gaussian second- and fourth-moment identities; all linear–quadratic cross moments vanish by symmetry.

Let $S_G = \|G\|^2$. Since $\|GG^\top - I_d\|_{\mathrm{F}}^2 = S_G^2 - 2S_G + d$, $\|\phi_d(G)\|^2 = S_G + \frac{1}{2}(S_G^2 - 2S_G + d) = \frac{1}{2}(S_G^2 + d)$. For $v \in \mathbb{R}^d$, $D\phi_d(G)[v] = \begin{pmatrix} v \\ \operatorname{svec}\left(\frac{Gv^\top + vG^\top}{\sqrt{2}}\right) \end{pmatrix}$, and hence $\|D\phi_d(G)[v]\|^2 = \|v\|^2 + \|G\|^2\|v\|^2 + (G^\top v)^2 \le (1 + 2S_G)\|v\|^2$. For $f(G) = \phi_d(G)^\top B\phi_d(G)$, $\nabla f(G) = 2D\phi_d(G)^\top B\phi_d(G)$. The Gaussian Poincaré inequality therefore gives Nourdin & Peccati (2012)

$$\begin{aligned}\operatorname{Var}\{f(G)\} &\le \mathbb{E}\|\nabla f(G)\|^2 \\ &\le 4\|B\|_{\mathrm{op}}^2\mathbb{E}[(1 + 2S_G)\|\phi_d(G)\|^2] \\ &= 2\|B\|_{\mathrm{op}}^2\mathbb{E}[(1 + 2S_G)(S_G^2 + d)].\end{aligned}$$

Using $\mathbb{E}S_G = d$, $\mathbb{E}S_G^2 = d(d+2)$, and $\mathbb{E}S_G^3 = d(d+2)(d+4)$, the final expression equals $(4d^3 + 30d^2 + 38d)\|B\|_{\mathrm{op}}^2$, which proves the variance bound. Because $p \ge d^2/2$, division by p^2 gives the normalized quadratic-form bound.

Finally, $\mathbb{E}S_G^4 = d(d+2)(d+4)(d+6)$, so

$$\begin{aligned}\mathbb{E}\|\phi_d(G)\|^4 &= \frac{1}{4}\mathbb{E}(S_G^2 + d)^2 \\ &= \frac{d^4 + 14d^3 + 49d^2 + 48d}{4} \le 7p^2,\end{aligned}$$

where the final inequality follows from

$$7p^2 - \frac{d^4 + 14d^3 + 49d^2 + 48d}{4} = \frac{1}{2}d(d-1)(d+3)(3d+8) \ge 0.$$

□

For independent rows distributed as $\phi_d(G)$, the preceding normalized quadratic-form concentration verifies the condition used in generalized Marchenko–Pastur results for isotropic vectors with dependent coordinates; see Yaskov (2016).

Quadratic prediction directions. For any $v \in \mathbb{R}^p$, write $v = (a, q)$, where $a \in \mathbb{R}^d$ and $q \in \mathbb{R}^{d(d+1)/2}$, and define $\mathcal{Q}_d(v) = \operatorname{svec}^{-1}(q)$. Then $v^\top\phi_d(G) = a^\top G + (G^\top\mathcal{Q}_d(v)G - \operatorname{tr}\{\mathcal{Q}_d(v)\})/\sqrt{2}$. Thus $\|\mathcal{Q}_d(v)\|_{\mathrm{op}}$ measures the size of the largest individual quadratic direction contributing to the scalar functional.

C.6.2 Gaussian Approximation for Quadratic Features

Lemma 5. *Let $F = a^\top G + \frac{G^\top BG - \mathrm{tr}(B)}{\sqrt{2}} + \sigma Z$, where $G \sim N(0, I_d)$, $Z \sim N(0,1)$ is independent of G, B is symmetric, and $\sigma > 0$. Define $V = \sigma^2 + \|a\|^2 + \mathrm{tr}(B^2)$. Then*

$$d_{\mathrm{K}}\left(\frac{F}{\sqrt{V}}, N(0,1)\right) \le \frac{\sqrt{\frac{9}{2}\|Ba\|^2 + 2\,\mathrm{tr}(B^4)}}{V} \le \frac{3}{\sqrt{2}\,\sigma}\|B\|_{\mathrm{op}}.$$

Proof. Gaussian integration by parts gives, for suitable differentiable f, $\mathbb{E}[Ff(F)] = \mathbb{E}[Tf'(F)]$, where $T = \sigma^2 + \|a\|^2 + \frac{3}{\sqrt{2}}a^\top BG + G^\top B^2 G$. Indeed, the linear and noise terms give $\mathbb{E}[(a^\top G + \sigma Z)f(F)] = \mathbb{E}[\{\|a\|^2 + \sqrt{2}a^\top BG + \sigma^2\}f'(F)]$, while the quadratic term gives

$$\mathbb{E}\left[\frac{G^\top BG - \mathrm{tr}(B)}{\sqrt{2}} f(F)\right] = \mathbb{E}\left[\left\{\frac{a^\top BG}{\sqrt{2}} + G^\top B^2 G\right\} f'(F)\right].$$

Thus $\mathbb{E}T = V$. The linear and centered quadratic terms in T are uncorrelated, giving

$$\mathrm{Var}(T) = \frac{9}{2}\|Ba\|^2 + 2\,\mathrm{tr}(B^4).$$

Applying the normal Stein equation for indicator test functions Nourdin & Peccati (2012) yields

$$d_{\mathrm{K}}(F/\sqrt{V}, N(0,1)) \le \mathbb{E}|1 - T/V| \le \frac{\sqrt{\mathrm{Var}(T)}}{V}.$$

Finally, $\|Ba\|^2 \le \|B\|_{\mathrm{op}}^2\|a\|^2$, and $\mathrm{tr}(B^4) \le \|B\|_{\mathrm{op}}^2\,\mathrm{tr}(B^2)$. Hence $\sqrt{\mathrm{Var}(T)} \le \frac{3}{\sqrt{2}}\|B\|_{\mathrm{op}}\sqrt{V}$. Since $V \ge \sigma^2$, the second bound follows. □

C.6.3 Proof of Theorem 3

Proof. Let $e = w_0 - \widehat{w}$. Conditionally on the training design R, Gaussianity of the random-effects readout and training noise gives $e \mid R \sim N(0, C_e)$, and $C_e = \frac{s_p^2\lambda^2}{p}A^2 + \frac{\sigma^2}{n}ASA$. Since $\|A\|_{\mathrm{op}} \le 1/\lambda$ and $\|ASA\|_{\mathrm{op}} \le 1/(4\lambda)$, $\|C_e\|_{\mathrm{op}} \le \frac{s_p^2}{p} + \frac{\sigma^2}{4n\lambda} =: c_n = O(n^{-1})$.

Set $v = \Sigma_p^{1/2}e$, $K_e = \Sigma_p^{1/2}C_e\Sigma_p^{1/2}$, and $Q_n = \|v\|^2$. Then $\|K_e\|_{\mathrm{op}} \le Lc_n$, and $\mathrm{tr}(K_e) = T_n = O(1)$. Thus $\mathbb{E}(Q_n \mid R) = T_n$, and $\mathrm{Var}(Q_n \mid R) = 2\,\mathrm{tr}(K_e^2) \le 2Lc_nT_n$. Therefore $Q_n = T_n + O_{\mathbb{P}}(n^{-1/2})$.

Next decompose the coefficient vector v according to the linear and quadratic parts of ϕ_d, writing $v^\top\phi_d(G) = a^\top G + \frac{G^\top BG - \mathrm{tr}(B)}{\sqrt{2}}$. For a unit vector $u \in \mathbb{R}^d$, $u^\top Bu$ is conditionally Gaussian, with variance at most Lc_n. Let $\mathcal{N}$ be a $1/4$-net of the unit sphere with $|\mathcal{N}| \le 9^d$. The standard net inequality gives $\|B\|_{\mathrm{op}} \le 2\max_{u\in\mathcal{N}}|u^\top Bu|$. A Gaussian union bound therefore yields, for every $t > 0$, $\mathbb{P}\left(\|B\|_{\mathrm{op}} > \sqrt{8Lc_n(d\log 9 + t)}\,\middle|\, R\right) \le 2e^{-t}$. Because $p = d(d+3)/2 \asymp n$, we have $d \asymp n^{1/2}$ and hence $\|B\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{\frac{d}{n}}\right) = O_{\mathbb{P}}(n^{-1/4})$. Condition on (w_0, R, y). A fresh signed prediction residual satisfies

$$y_* - r_*^\top\widehat{w} = a^\top G_* + \frac{G_*^\top BG_* - \mathrm{tr}(B)}{\sqrt{2}} + \sigma Z_*,$$

and has conditional variance $\sigma^2 + Q_n$. By Lemma 5,

$$\sup_x\left|\mathbb{P}(y_* - r_*^\top\widehat{w} \le x \mid w_0, R, y) - \Phi\left(\frac{x}{\sqrt{\sigma^2 + Q_n}}\right)\right| = O_{\mathbb{P}}(n^{-1/4}).$$

The corresponding absolute-residual CDF differs from the folded-Gaussian CDF by at most twice this amount.

Conditionally on (w_0, R, y), the calibration observations are independent. The Dvoretzky–Kiefer–Wolfowitz–Massart inequality Massart (1990), together with the preceding Gaussian approximation and $Q_n = T_n + O_{\mathbb{P}}(n^{-1/2})$, therefore gives $\widehat{q}_{1-\alpha} = z\sqrt{\sigma^2 + T_n} + O_{\mathbb{P}}(n^{-1/4} + m^{-1/2})$. Since $T_n = O(1)$ and $\sigma > 0$, the relevant folded-Gaussian density is uniformly bounded away from zero in a neighborhood of the target quantile. Squaring the preceding display gives $\frac{W_{C,n}^2}{4z^2} = \sigma^2 + T_n + O_{\mathbb{P}}(n^{-1/4} + m^{-1/2})$.

For the Bayesian interval, conditionally on R, $r_*^\top A r_* = \phi_d(G_*)^\top \Sigma_p^{1/2} A \Sigma_p^{1/2} \phi_d(G_*)$. Because $\|\Sigma_p^{1/2} A \Sigma_p^{1/2}\|_{\mathrm{op}} \le \frac{L}{\lambda}$, Lemma 4 gives

$$\mathbb{E}\left[\left\{\frac{r_*^\top A r_* - \mathrm{tr}(\Sigma_p A)}{n}\right\}^2 \,\middle|\, R\right] \le \frac{72 L^2 d^3}{\lambda^2 n^2}.$$

Since $d^2 \asymp n$, $\frac{1}{n} r_*^\top A r_* = \frac{1}{n}\mathrm{tr}(\Sigma_p A) + O_{\mathbb{P}}(n^{-1/4})$, and hence $\frac{W_{B,n}(r_*)^2}{4z^2} = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/4})$. Finally, using $ASA = A - \lambda A^2$,

$$\begin{aligned} P_n - T_n &= \frac{\sigma^2}{n}\mathrm{tr}(\Sigma_p A) - \frac{\sigma^2}{n}\mathrm{tr}(\Sigma_p ASA) - \frac{s_p^2\lambda^2}{p}\mathrm{tr}(\Sigma_p A^2) \\ &= \left(\frac{\sigma^2\lambda}{n} - \frac{s_p^2\lambda^2}{p}\right)\mathrm{tr}(\Sigma_p A^2). \end{aligned}$$

□

Approximation rate. The $O_{\mathbb{P}}(n^{-1/4})$ terms in Theorem 3 describe approximation by the finite-sample trace quantities and are not claimed as rates for convergence of those traces to their Marchenko–Pastur limits.

When $\Sigma_p = I_p$, Lemma 4 verifies the quadratic-form concentration condition required for the Marchenko–Pastur limit for isotropic vectors with dependent coordinates Yaskov (2016). Therefore $P_n \to \sigma^2\gamma I_1$ and $T_n \to s^2\lambda^2 I_2 + \sigma^2\gamma J$, which yields the same limiting width formulas and phase boundary as Corollary 2. This conclusion follows here from the dependent-feature concentration argument, not from the Gaussian training-row assumption used in the proof of that corollary.

C.6.4 Deterministic Readout Extension

The random-effects assumption in equation 3 provides a sufficient mechanism ensuring that the fitted prediction direction is spread across many quadratic coordinates. Randomness of the underlying NGRC readout is not required, however. For a fixed dynamical system, a more natural formulation treats w_0 as deterministic and directly controls the concentration of its quadratic ridge-bias direction.

For a deterministic readout, define $B_n = \lambda^2 w_0^\top A \Sigma_p A w_0$, $V_n = \frac{\sigma^2}{n}\mathrm{tr}(\Sigma_p ASA)$, and let $D_n = \mathcal{Q}_d\left(\lambda \Sigma_p^{1/2} A w_0\right)$.

Corollary 3 (Deterministic quadratic-NGRC readout)**.** *Suppose the feature, noise, proportional-growth, and regularization conditions of Theorem 3 hold, but let $w_0 = w_{0,p}$ be deterministic and satisfy $\sup_p \|w_{0,p}\| < \infty$. Assume that, for some deterministic sequence $\rho_n \downarrow 0$,*

$$\|D_n\|_{\mathrm{op}} = O_{\mathbb{P}}(\rho_n). \tag{5}$$

Then

$$\frac{W_{B,n}(r_*)^2}{4z^2} = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/4}), \quad \frac{W_{C,n}^2}{4z^2} = \sigma^2 + B_n + V_n + O_{\mathbb{P}}\left(n^{-1/4} + \rho_n + m^{-1/2}\right).$$

Consequently,

$$\frac{W_{B,n}(r_*)^2 - W_{C,n}^2}{4z^2} = \frac{\sigma^2\lambda}{n}\mathrm{tr}(\Sigma_p A^2) - B_n + O_{\mathbb{P}}\left(n^{-1/4} + \rho_n + m^{-1/2}\right).$$

In particular, if

$$\|D_n\|_{\mathrm{op}} = O_{\mathbb{P}}(n^{-1/4}),$$

then the deterministic-readout extension has the same approximation order as the random-effects result.

The condition in equation 5 is imposed on the ridge-transformed quadratic bias direction $\mathcal{Q}_d(\lambda\Sigma_p^{1/2}Aw_0)$. Diffuseness of the raw quadratic signal $\mathcal{Q}_d(w_0)$ alone does not imply this condition.

Proof. Let $e = w_0 - \widehat{w}$. Since $\widehat{w} = A\left(Sw_0 + \frac{1}{n}R^\top\varepsilon\right)$ and $AS = I - \lambda A$, then we have $e = \lambda Aw_0 - \xi$ and $\xi = \frac{1}{n}AR^\top\varepsilon$. Conditionally on R, $\xi \sim N(0, C)$ and $C = \frac{\sigma^2}{n}ASA$.

Set $b = \lambda Aw_0$ and $Q_n = e^\top\Sigma_p e$. Then $\mathbb{E}(Q_n \mid R) = b^\top\Sigma_p b + \mathrm{tr}(\Sigma_p C) = B_n + V_n$. As in the proof of Theorem 2, the Gaussian quadratic-form identities give

$$\mathrm{Var}(Q_n \mid R) = 2\,\mathrm{tr}\left[\left(\Sigma_p^{1/2}C\Sigma_p^{1/2}\right)^2\right] + 4b^\top\Sigma_p C\Sigma_p b.$$

Because $\|A\|_{\mathrm{op}} \le \lambda^{-1}$, $\|ASA\|_{\mathrm{op}} \le 1/4\lambda$, and $\|\Sigma_p\|_{\mathrm{op}}$ and $\|w_0\|$ are uniformly bounded, the right-hand side is $O(n^{-1})$. Hence $Q_n = B_n + V_n + O_{\mathbb{P}}(n^{-1/2})$. It remains to control the distribution of a new prediction residual. Define $v = \Sigma_p^{1/2}e = v_b - v_\xi$, where $v_b = \lambda\Sigma_p^{1/2}Aw_0$ and $v_\xi = \Sigma_p^{1/2}\xi$. By definition, $D_n = \mathcal{Q}_d(v_b)$, and by linearity of $\mathcal{Q}_d$, $\mathcal{Q}_d(v) = D_n - \mathcal{Q}_d(v_\xi)$. By assumption, $\|D_n\|_{\mathrm{op}} = O_{\mathbb{P}}(\rho_n)$. The first term satisfies $\|\mathcal{Q}_d(v_b)\|_{\mathrm{op}} = O_{\mathbb{P}}(\rho_n)$ by assumption.

Conditionally on R, v_ξ is Gaussian with covariance $K_\xi = (\sigma^2/n)\Sigma_p^{1/2}ASA\Sigma_p^{1/2}$, and therefore $\|K_\xi\|_{\mathrm{op}} = O(n^{-1})$. For each fixed unit vector $u \in \mathbb{R}^d$, $u^\top\mathcal{Q}_d(v_\xi)u$ is a centered Gaussian linear functional of v_ξ with variance $O(n^{-1})$. Applying the same $1/4$-net argument used in the random-effects proof gives $\|\mathcal{Q}_d(v_\xi)\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{d/n}\right)$. Since $p = d(d+3)/2 \asymp n$, we have $d \asymp n^{1/2}$ and hence $\|\mathcal{Q}_d(v_\xi)\|_{\mathrm{op}} = O_{\mathbb{P}}(n^{-1/4})$. Consequently, $\|\mathcal{Q}_d(v)\|_{\mathrm{op}} = O_{\mathbb{P}}(\rho_n + n^{-1/4})$.

Conditionally on the fitted training sample, a fresh signed prediction residual can be written as

$$y_* - r_*^\top\widehat{w} = a^\top G_* + \frac{G_*^\top\mathcal{Q}_d(v)G_* - \mathrm{tr}\{\mathcal{Q}_d(v)\}}{\sqrt{2}} + \sigma Z_*,$$

and its conditional variance is $\sigma^2 + \|v\|^2 = \sigma^2 + Q_n$. Lemma 5 therefore yields

$$\sup_x \left|\mathbb{P}(y_* - r_*^\top\widehat{w} \le x \mid R, y) - \Phi\left(\frac{x}{\sqrt{\sigma^2 + Q_n}}\right)\right| = O_{\mathbb{P}}(\rho_n + n^{-1/4}).$$

The same bound, up to a constant factor, holds for the corresponding absolute-residual distribution.

Conditionally on the training sample, the calibration observations are independent. The Dvoretzky–Kiefer–Wolfowitz–Massart inequality and the preceding Gaussian approximation therefore give

$$\widehat{q}_{1-\alpha} = z\sqrt{\sigma^2 + Q_n} + O_{\mathbb{P}}\left(\rho_n + n^{-1/4} + m^{-1/2}\right).$$

Since $Q_n = O_{\mathbb{P}}(1)$ and $\sigma > 0$, squaring and using $Q_n = B_n + V_n + O_{\mathbb{P}}(n^{-1/2})$ gives

$$\frac{W_{C,n}^2}{4z^2} = \sigma^2 + B_n + V_n + O_{\mathbb{P}}\left(\rho_n + n^{-1/4} + m^{-1/2}\right).$$

The Bayesian calculation does not require randomness of w_0. The quadratic-form concentration argument already used in the random-effects case gives $r_*^\top Ar_*/n = \mathrm{tr}(\Sigma_p A)/n + O_{\mathbb{P}}(n^{-1/4})$, and hence $W_{B,n}(r_*)^2/4z^2 = \sigma^2 + P_n + O_{\mathbb{P}}(n^{-1/4})$. Finally, $P_n - V_n = \sigma^2\lambda\,\mathrm{tr}(\Sigma_p A^2)/n$, so

$$\frac{W_{B,n}(r_*)^2 - W_{C,n}^2}{4z^2} = \frac{\sigma^2\lambda}{n}\,\mathrm{tr}(\Sigma_p A^2) - B_n + O_{\mathbb{P}}\left(\rho_n + n^{-1/4} + m^{-1/2}\right),$$

which gives the stated width-difference expansion. Finally, if $\rho_n = O(n^{-1/4})$, then $\rho_n + n^{-1/4} = O(n^{-1/4})$, so the deterministic-readout extension has the same approximation order as the random-effects result. □

C.6.5 WHY QUADRATIC DIFFUSENESS IS NEEDED.

Consider $f_d(G) = (G_1^2 - 1)/\sqrt{2}$. For every d, $\mathrm{Var}\{f_d(G)\} = 1$, but its distribution is a centered and scaled χ_1^2 distribution and does not approach a Gaussian distribution as $d \to \infty$. The associated quadratic matrix is $B_d = e_1 e_1^\top$, and $\|B_d\|_{\mathrm{op}} = 1$ so all quadratic variation remains concentrated in a single direction.

By contrast, consider the diffuse quadratic functional

$$\widetilde{f}_d(G) = \frac{1}{\sqrt{d}} \sum_{j=1}^{d} \frac{G_j^2 - 1}{\sqrt{2}} = \frac{\|G\|^2 - d}{\sqrt{2d}}.$$

Its associated quadratic matrix is $\widetilde{B}_d = I_d/\sqrt{d}$, so $\|\widetilde{B}_d\|_{\mathrm{op}} = d^{-1/2} \longrightarrow 0$, and $\mathrm{tr}(\widetilde{B}_d^2) = 1$. Thus its total variance remains of constant order while its largest individual quadratic direction vanishes. The central limit theorem for the centered chi-square statistic yields $\widetilde{f}_d(G) \xrightarrow{d} N(0,1)$.

The two examples show what equation 5 rules out. High dimensionality alone does not imply Gaussian prediction errors if the signal remains concentrated in a small number of quadratic directions.

C.7 ISOTROPIC MARCHENKO–PASTUR SPECIALIZATION

For the isotropic proportional benchmark, define

$$I_1(\gamma, \lambda) = \int \frac{1}{x + \lambda}\, dF_\gamma^{\mathrm{MP}}(x), \qquad I_2(\gamma, \lambda) = \int \frac{1}{(x + \lambda)^2}\, dF_\gamma^{\mathrm{MP}}(x),$$

and

$$J(\gamma, \lambda) = I_1 - \lambda I_2 = \int \frac{x}{(x + \lambda)^2}\, dF_\gamma^{\mathrm{MP}}(x),$$

where F_γ^{MP} denotes the Marchenko–Pastur law Marčenko & Pastur (1967). These quantities are the deterministic limits of the resolvent traces appearing in the finite-trace comparison.

C.7.1 PROOF OF COROLLARY 2

Proof. By the additional assumptions of Corollary 2, the rows of R are i.i.d. $N(0, I_p)$. Hence $S = \frac{1}{n} R^\top R$ is a Gaussian sample-covariance matrix. Write $S = U \Lambda U^\top$, and $A = (S + \lambda I_p)^{-1} = U(\Lambda + \lambda I_p)^{-1} U^\top$. The Marchenko–Pastur theorem Marčenko & Pastur (1967); Dobriban & Wager (2018), together with bounded continuity of

$$x \mapsto \frac{1}{x + \lambda}, \qquad x \mapsto \frac{1}{(x + \lambda)^2}, \qquad x \mapsto \frac{x}{(x + \lambda)^2},$$

gives

$$\frac{1}{p}\,\mathrm{tr}(A) \xrightarrow{\mathbb{P}} I_1, \qquad \frac{1}{p}\,\mathrm{tr}(A^2) \xrightarrow{\mathbb{P}} I_2, \qquad \frac{1}{p}\,\mathrm{tr}(ASA) \xrightarrow{\mathbb{P}} J.$$

Since $p/n \to \gamma$,

$$P_n = \frac{\sigma^2}{n}\,\mathrm{tr}(A) \xrightarrow{\mathbb{P}} \sigma^2 \gamma I_1, \qquad V_n = \frac{\sigma^2}{n}\,\mathrm{tr}(ASA) \xrightarrow{\mathbb{P}} \sigma^2 \gamma J.$$

It remains to identify the limit of the bias term $B_n = \lambda^2 w_0^\top A^2 w_0$. Let $D = (\Lambda + \lambda I_p)^{-2}$, so that $A^2 = UDU^\top$. Because the isotropic Gaussian design is orthogonally invariant, conditional on the eigenvalues Λ, the direction $u = U^\top w_0/\|w_0\|$ is uniformly distributed on the unit sphere in $\mathbb{R}^p$. Hence $w_0^\top A^2 w_0 = \|w_0\|^2 u^\top D u$.

For u uniformly distributed on the unit sphere, $\mathbb{E}(u_i^2) = 1/p$, $\mathbb{E}(u_i^4) = 3/p(p+2)$, and $E(u_i^2 u_j^2) = 1/p(p+2)$ for $i \neq j$. Therefore, conditional on Λ, $\mathbb{E}\left[w_0^\top A^2 w_0 \,\middle|\, \Lambda\right] = (\|w_0\|^2/p)\,\mathrm{tr}(A^2)$. The same sphere-moment identities give

$$\mathrm{Var}\left(w_0^\top A^2 w_0 \,\middle|\, \Lambda\right) = \frac{2\|w_0\|^4}{p(p+2)} \left\{ \mathrm{tr}(A^4) - \frac{\mathrm{tr}(A^2)^2}{p} \right\} \leq \frac{2\|w_0\|^4}{p(p+2)}\,\mathrm{tr}(A^4).$$

Since $\|A\|_{\mathrm{op}} \le \lambda^{-1}$, we have $\mathrm{tr}(A^4) \le p\|A\|_{\mathrm{op}}^4 \le p/\lambda^4$, and hence $\mathrm{Var}\left(w_0^\top A^2 w_0 \,\middle|\, \Lambda\right) \le 2\|w_0\|^4/(p+2)\lambda^4$. Because $\|w_0\|^2 \to s^2 < \infty$, the right-hand side converges to zero. Conditional Chebyshev's inequality therefore yields

$$w_0^\top A^2 w_0 - \frac{\|w_0\|^2}{p}\,\mathrm{tr}(A^2) \xrightarrow{\mathbb{P}} 0.$$

Together with $\|w_0\|^2 \to s^2$ and $\mathrm{tr}(A^2)/p \xrightarrow{\mathbb{P}} I_2$, this gives $B_n = \lambda^2 w_0^\top A^2 w_0 \xrightarrow{\mathbb{P}} s^2\lambda^2 I_2$.

Theorem 2 now implies

$$\frac{W_{B,n}(r_*)^2}{4z^2} \xrightarrow{\mathbb{P}} \sigma^2(1+\gamma I_1), \qquad \frac{W_{C,n}^2}{4z^2} \xrightarrow{\mathbb{P}} \sigma^2(1+\gamma J) + s^2\lambda^2 I_2.$$

Finally, $ASA = A - \lambda A^2$, so $J = I_1 - \lambda I_2$. Therefore

$$\begin{aligned}
&\sigma^2(1+\gamma I_1) - \left\{\sigma^2(1+\gamma J) + s^2\lambda^2 I_2\right\} \\
&\quad = \sigma^2\gamma(I_1 - J) - s^2\lambda^2 I_2 \\
&\quad = \sigma^2\gamma\lambda I_2 - s^2\lambda^2 I_2 \\
&\quad = \lambda I_2(\sigma^2\gamma - s^2\lambda).
\end{aligned}$$

Multiplying by $4z^2$ yields the stated limit in Corollary 2. Since $\lambda > 0$ and $I_2 > 0$, the sign of the limiting squared-width difference is determined by $\sigma^2\gamma - s^2\lambda$, giving the three stated asymptotic ordering regimes. □

C.7.2 Prior-Scale Interpretation of the Equality Boundary

The preceding interpretation conditions on a fixed deterministic ridge parameter λ. If λ is selected from the data, for example by cross-validation or empirical Bayes, its limiting behavior and dependence on the fitted signal must be incorporated into the analysis. Consequently, the fixed-λ phase diagram does not by itself predict the calibration of a data-adaptively regularized procedure. In the real-data experiments, ridge regularization is selected using pre-test fitting data. Those experiments therefore provide empirical evidence under tuning rather than a direct test of Corollary 2.

Under the Bayesian ridge prior, $w \sim N\left(0, \frac{\sigma^2}{n\lambda} I_p\right)$. Hence $\mathbb{E}\|w\|^2 = p\sigma^2/n\lambda \longrightarrow \sigma^2\gamma/\lambda$. Therefore, the equality condition $s^2\lambda = \sigma^2\gamma$ is equivalent to $s^2 = \sigma^2\gamma/\lambda$, so the limiting true signal strength matches the limiting signal scale implied by the isotropic Bayesian prior. This interpretation is specific to the isotropic benchmark and is not required for the more general finite-trace comparison.

C.7.3 Bayesian coverage in the proportional regime.

Let $v_B = \sigma^2(1+\gamma I_1)$, and
$v_C = \sigma^2(1+\gamma J) + s^2\lambda^2 I_2$. The preceding proof shows that the Bayesian squared radius satisfies $W_{B,n}^2/4 \xrightarrow{\mathbb{P}} z^2 v_B$, and hence $W_{B,n}/2 \xrightarrow{\mathbb{P}} z\sqrt{v_B}$.

To identify the limiting distribution of the prediction error, define $E_n = y_* - r_*^\top \widehat{w}$. Conditional on the fitted training sample $\mathcal{D}_{\mathrm{tr}}$, the Gaussian design gives $E_n \mid \mathcal{D}_{\mathrm{tr}} \sim N(0, \sigma^2 + Q_n)$, where $Q_n = (\widehat{w} - w_0)^\top \Sigma (\widehat{w} - w_0)$. The proof of Theorem 2 and the preceding Marchenko–Pastur limits give $\sigma^2 + Q_n \xrightarrow{\mathbb{P}} v_C$. For each fixed $t \in \mathbb{R}$,

$$\mathbb{E}[e^{itE_n} \mid \mathcal{D}_{\mathrm{tr}}] = \exp\left\{-\frac{t^2}{2}(\sigma^2 + Q_n)\right\} \xrightarrow{\mathbb{P}} \exp\left(-\frac{t^2 v_C}{2}\right).$$

The conditional characteristic functions are bounded in absolute value by one. Hence bounded convergence along subsequences, or equivalently dominated convergence together with convergence in probability, gives $\mathbb{E}e^{itE_n} \longrightarrow \exp\left(-t^2 v_C/2\right)$. By Lévy's continuity theorem, $E_n \xrightarrow{d} N(0, v_C)$. Since $W_{B,n}/2 \xrightarrow{\mathbb{P}} z\sqrt{v_B}$, Slutsky's theorem gives

$$\mathbb{P}\left(|E_n| \le \frac{W_{B,n}}{2}\right) \longrightarrow \mathbb{P}\left(|N(0, v_C)| \le z\sqrt{v_B}\right).$$

Therefore

$$\mathbb{P}\{y_* \in C_B(r_*)\} \longrightarrow 2\Phi\left(z\sqrt{\frac{v_B}{v_C}}\right) - 1.$$

C.7.4 Optimal Ridge Penalty and the Equality Boundary

For $s^2 > 0$, the limiting prediction risk is

$$\mathcal{E}(\lambda) = \sigma^2 + \int \frac{s^2\lambda^2 + \sigma^2\gamma x}{(x+\lambda)^2}\, dF_\gamma^{\mathrm{MP}}(x).$$

Differentiating under the integral gives $\mathcal{E}'(\lambda) = 2(s^2\lambda - \sigma^2\gamma)\int \frac{x}{(x+\lambda)^3}\, dF_\gamma^{\mathrm{MP}}(x)$. The integral is positive, so the unique minimizer is $\lambda_* = \frac{\sigma^2\gamma}{s^2}$. This is the standard optimal-ridge scaling in the isotropic proportional regime; see Dobriban & Wager (2018). At this value, $s^2\lambda_* = \sigma^2\gamma$, which is exactly the equality boundary in Corollary 2.

C.7.5 Relation to the fixed-dimensional regime.

The proportional-growth result does not reduce automatically to Theorem 1 by taking $\gamma \downarrow 0$. The high-dimensional analysis holds the normalized ridge penalty $\lambda > 0$ fixed. Under this scaling, a nonvanishing shrinkage bias may persist even as the aspect ratio decreases. Recovering the fixed-dimensional result therefore additionally requires $\lambda \longrightarrow 0$ on the normalized covariance scale, or more generally requires the resulting prediction bias to vanish. The limits $p/n \to 0$ and $\lambda \to 0$ therefore represent distinct asymptotic operations.

C.8 Extension to Temporally Dependent Observations

The preceding high-dimensional results use an independent-sample benchmark. We now give one sufficient condition under which probability statements for independent NGRC observation windows can be transferred to a dependent time series.

Suppose the underlying process is strictly stationary and absolutely regular with coefficient $\beta(h)$, using the total-variation convention described in Appendix C.1. A feature–response pair at forecast origin t depends on observations in the window $[t-L_n, t+H]$, $L_n = (k_n-1)s_n$. Let $t_1 < \cdots < t_N$ denote retained forecast origins and suppose

$$t_{j+1} - t_j \geq L_n + H + h_n, \qquad h_n \geq 1. \tag{6}$$

Proposition 1 (Transfer under absolute regularity)**.** *Under equation 6, the N retained observation windows can be coupled to mutually independent copies with the same marginal window distribution such that the probability of at least one coupling failure is at most $(N-1)\beta(h_n)$. Consequently, if an event has failure probability at most δ under the corresponding independent-window model, then its failure probability under the dependent process is at most $\delta + (N-1)\beta(h_n)$. In particular, whenever the independent copies satisfy the assumptions required for the standard split-conformal guarantee, $\mathbb{P}\{y_* \in C_C(r_*)\} \geq 1 - \alpha - (N-1)\beta(h_n)$.*

Proof. The latest observation in the window associated with t_j is $t_j + H$, while the earliest observation in the next retained window is $t_{j+1} - L_n$. Therefore equation 6 implies $(t_{j+1} - L_n) - (t_j + H) \geq h_n$. By Berbee's coupling principle Berbee (1979), the $(j+1)$st window can be coupled to an independent copy having the same marginal distribution, with failure probability at most $\beta(h_n)$. Applying the coupling successively over the $N-1$ transitions and using the union bound gives total coupling failure probability at most $(N-1)\beta(h_n)$.

On the event that all couplings succeed, every statistic computed from the retained windows agrees with the corresponding statistic computed from the independent copies. Hence, if $\mathcal{E}$ has failure probability at most δ under the independent-window model,

$$\mathbb{P}_{\mathrm{dep}}(\mathcal{E}^c) \leq \mathbb{P}_{\mathrm{ind}}(\mathcal{E}^c) + \mathbb{P}(\text{coupling failure}) \leq \delta + (N-1)\beta(h_n).$$

Taking $\mathcal{E}$ to be the split-conformal coverage event gives the stated bound. □

Scope of the transfer result. The separation condition applies to every pair of successive retained forecast windows, not merely to the boundaries between fitting, calibration, and test blocks. Thus the proposition describes a thinned sequence of sufficiently separated NGRC forecast origins. It does not provide an independence approximation for the ordinary overlapping delay windows used at every consecutive forecast origin.

Sampling cost under quadratic features. The separation requirement can be costly in the proportional quadratic-NGRC regime. With a full quadratic lag map, $p \asymp k_n^2$. If $p \asymp n$, then $k_n \asymp n^{1/2}$. For fixed lag spacing, $L_n = (k_n - 1)s_n \asymp n^{1/2}$. Retaining $N \asymp n$ windows while enforcing $t_{j+1} - t_j \geq L_n + H + h_n$ therefore requires an underlying time series of length at least order $NL_n \asymp n^{3/2}$, even before accounting for any additional mixing gap h_n. The proposition should therefore be viewed as a sufficient theoretical bridge for heavily separated windows rather than as a direct asymptotic description of ordinary overlapping NGRC forecasts.

The result is also only a transfer principle. It transfers probability statements from an independent-window model to sufficiently separated dependent windows, but does not establish Gaussian feature marginals, the linear readout model, quadratic-feature concentration, or a Marchenko–Pastur law for an arbitrary dependent NGRC process.

For example, if $\beta(h) \leq Ce^{-ch}$, choosing $h_n \geq \frac{1}{c} \log \left\{ \frac{C(N-1)}{\eta_n} \right\}$ makes the additional dependence error at most η_n. Thus preventing overlap between adjacent NGRC windows is not sufficient for approximate independence: an additional gap is required relative to the decay of temporal dependence.

The coupling result addresses dependence across retained windows only. It does not establish the Gaussian feature distribution, the high-dimensional readout model, or the preprocessing assumptions used elsewhere in the theory. Those conditions must be verified separately. For related conformal methods for dependent data, see Chernozhukov et al. (2018).

D Experimental Details

D.1 Datasets

We use four univariate forecasting targets. Table 3 reports the number of observations after construction of each target and the horizon-specific fixed-origin splits used in the primary analysis. All splits are chronological. After constructing the direct H-step NGRC rows, the first 40% are used for fitting, the next 40% for calibration, and the final 20% for testing. Rows are assigned by target time. Thus, a feature row is never moved into an earlier block merely because some of its lagged inputs precede a split boundary.

LA ozone. We analyze the `LA_O3` object contained in `LA_O3.RData`. The exact data file used in our experiments is included in the supplementary submission archive. The file provides daily ozone observations for a location near Los Angeles at longitude -118.742064 and latitude 34.0206066. The forecasting target is the `value` field. Observations run from January 1, 1980, through December 31, 2023, yielding 16,071 daily values. If multiple records share a date, their `value` measurements are averaged before construction of the regular daily grid. The resulting series contains no missing dates.

Solar. We use the Solar-Energy benchmark distributed with the multivariate time-series repository of Lai et al. (2018). The source matrix contains 137 solar-power series from Alabama sampled every 10 minutes during 2006. The forecasting target is the cross-sectional mean across the 137 series at each time point, yielding 52,560 observations. The source matrix used in the experiments contains no missing entries. Because there are 144 observations per day and 1,008 per week, these periods are also used for deterministic seasonal features.

Beijing PM_{10}. We use the Beijing Multi-Site Air Quality dataset from the UCI Machine Learning Repository (Chen, 2017). The dataset contains hourly air-pollution measurements from 12 monitoring sites in Beijing from March 1, 2013, through February 28, 2017. The target is the mean PM$*10$ concentration across the sites reporting at a given hour, in $\mu\mathrm{g\,m^{-3}}$. The regular hourly grid contains

Table 3: Processed series lengths and chronological fitting/calibration/test sample sizes. Each triplet is reported as $n_{\mathrm{fit}}/n_{\mathrm{cal}}/n_{\mathrm{test}}$. Counts vary slightly with H because direct H-step feature construction removes the final H forecast origins.

Dataset	Observations	$H = 1$	$H = 3$	$H = 7$
LA ozone	16,071	6,420/6,420/3,210	6,419/6,419/3,210	6,417/6,417/3,210
Solar	52,560	21,009/21,009/10,506	21,008/21,008/10,506	21,007/21,007/10,504
Beijing PM_{10}	35,064	14,016/14,016/7,008	14,015/14,015/7,008	14,013/14,013/7,008
Exchange returns	7,587	3,029/3,029/1,515	3,028/3,028/1,515	3,026/3,026/1,515

Table 4: Dataset-specific NGRC configurations. The last column gives the ridge penalties selected at horizons $H = 1, 3, 7$, respectively.

Dataset	Delays k	Degree	Features p	λ at $H = (1, 3, 7)$
LA ozone	21	2	253	(0.03, 1, 0.3)
Solar	36	2	711	(0.001, 0.001, 0.001)
Beijing PM_{10}	24	2	333	(0.003, 0.001, 0.001)
Exchange returns	14	2	120	(10, 10, 10)

Solar uses two Fourier harmonics at periods 144 and 1,008 observations; Beijing uses two harmonics at periods 24 and 168 hours. Ozone and Exchange returns use no Fourier terms.

35,064 observations. After spatial averaging, 83 hours have no available PM_{10} measurement at any site; these values are filled causally using the most recent earlier hourly value.

Exchange returns. We use the Exchange-Rate benchmark distributed with the multivariate time-series repository of Lai et al. (2018), available at `https://github.com/laiguokun/multivariate-time-series-data`. The dataset contains eight daily exchange-rate series from 1990 to 2016, listed as Australia, Britain, Canada, Switzerland, China, Japan, New Zealand, and Singapore. Following this ordering, we analyze the first series (array index 0), corresponding to Australia.

Let P_t denote the observed exchange-rate level. We transform the positive level series to scaled log returns, $r_t = 100\big[\log(P_t) - \log(P_{t-1})\big]$.

The 7,588 level observations therefore produce 7,587 daily returns. The selected series contains no missing values.

The repeated-origin sensitivity analysis uses three shifted chronological windows beginning at 0%, 10%, and 20% of the valid NGRC rows. Each window contains 40% fitting, 20% calibration, and 20% test rows; the resulting windows overlap and are used to assess temporal sensitivity rather than to create independent replications.

D.2 NGRC Configuration and Point Predictors

For a direct forecast at horizon H, the delay vector at forecast origin i is

$$d_i = (x_i, x_{i-1}, \ldots, x_{i-k+1})^\top,$$

and the target is x_{i+H}. Thus, a prediction for target time t uses observations only through $t-H$. Lag spacing is one observation for every dataset. The default NGRC feature vector contains an intercept, the k linear lags, and the $k(k+1)/2$ unique upper-triangular quadratic products $d_{i,a}d_{i,b}$ for $a \le b$. No cubic terms are included. For Solar and Beijing, deterministic Fourier terms are added as

$$\sin(2\pi ji/P), \qquad \cos(2\pi ji/P), \qquad j \in \{1, 2\},$$

for each listed period P. These terms depend only on the integer time index; no frequency or phase is estimated from future responses.

The NGRC coefficients minimize

$$\frac{1}{n_{\mathrm{fit}}}\|y - Rw\|_2^2 + \lambda\|D^{1/2}w\|_2^2, \qquad D = \mathrm{diag}(0, 1, \ldots, 1),$$

so the intercept is not penalized. This is the empirical convention used in every reported real-data fit. It differs from the theoretical convention in Equation 1, where all coordinates are penalized through λI_p. Under the normalized ridge objective and a Gaussian likelihood with variance σ^2, the empirical penalty corresponds to a flat prior on the intercept and independent Gaussian priors with precision $n_{\text{fit}}\lambda/\sigma^2$ on the remaining coefficients. The theoretical results use the proper isotropic Gaussian prior on all coordinates; the unpenalized intercept is an empirical modeling convention and is not used in the theoretical derivations.

D.2.1 Additional Point-Forecasting Baselines

The exploratory architecture comparison uses the same standardized target, chronological fitting/calibration/test blocks, and one-step forecast targets as the NGRC analysis. Sequence-model lookback lengths equal the dataset-specific NGRC delay lengths: 21 observations for LA ozone, 36 for Solar, 24 for Beijing PM_{10}, and 14 for Exchange returns.

The ARIMA baseline is an $\mathrm{ARIMA}(3,1,3)$ model with no deterministic trend. Its parameters are estimated from the fitting prefix. During calibration and testing, the fitted state is updated sequentially after each newly observed response, without re-estimating the model order or using future responses.

The GRU consists of one univariate GRU layer with hidden dimension 32, followed by a linear readout from the final hidden state. The causal Transformer first embeds each scalar observation into dimension 32 and adds a learned positional representation. It uses three Transformer encoder layers, two attention heads, feed-forward dimension 64, GELU activations, dropout probability 0.1, and an upper-triangular causal attention mask. A linear readout is applied to the representation at the final sequence position.

Both neural models are trained with the mean absolute error loss and the Adam optimizer with learning rate 10^{-3}. Training uses batch size 32, no within-epoch shuffling, and at most 200 epochs. The final 20% of the fitting block is reserved as a chronological early-stopping set; the calibration block is not used for neural-model selection. Training stops after 20 epochs without an improvement exceeding 10^{-7} in validation loss. NumPy and PyTorch are initialized with seed 123. Once training and early stopping are complete, all point-model parameters are frozen before uncertainty quantification.

D.3 Uncertainty-Quantification Methods

All primary UQ methods are applied to the same NGRC point forecasts $\widehat{y}_t = r_t^\top \widehat{w}$. For Bayesian ridge, let $G_\lambda = R_{\text{fit}}^\top R_{\text{fit}} + n_{\text{fit}}\lambda D$, and $\widehat{\sigma}^2_{\text{fit}} = \frac{1}{n_{\text{fit}}}\sum_{i\in\text{fit}}(y_i - \widehat{y}_i)^2$. Let $s_t = \widehat{\sigma}_{\text{fit}}\sqrt{1 + r_t^\top G_\lambda^{-1} r_t}$. The plug-in Gaussian predictive interval is $C_{\mathrm{B}}(r_t) = \left[\widehat{y}_t - z_{1-\alpha/2}s_t, \widehat{y}_t + z_{1-\alpha/2}s_t\right]$. The scale is estimated from fitting residuals and is frozen before calibration and testing.

Ordinary split conformal prediction uses calibration scores $S_j = |y_j - \widehat{y}_j|, \quad j = 1,\dots,m$. Let $S_{(1)} \le \cdots \le S_{(m)}$ denote their order statistics and define $k_\alpha = \lceil (m+1)(1-\alpha) \rceil$, with the convention $S_{(m+1)} = +\infty$. The split-conformal radius is $q_{1-\alpha} = S_{(k_\alpha)}$, and the resulting interval is $C_{\mathrm{SCP}}(r_t) = [\widehat{y}_t - q_{1-\alpha},\ \widehat{y}_t + q_{1-\alpha}]$. Under exchangeability, this convention gives the standard finite-sample marginal coverage guarantee. The asymmetric SCP diagnostic instead applies equal-tail quantiles at $\alpha/2$ and $1-\alpha/2$ to the signed calibration residuals.

D.3.1 Time-Weighted Conformal Prediction

Let $S_j = \left|y_j - r_j^\top \widehat{w}\right|$ denote a previously observed forecast residual, and let $S_1,\dots,S_{M_t}$ be the scores available when the prediction at time t is issued, ordered from oldest to newest. We assign exponentially decaying weights $\omega_{j,t} = \gamma^{M_t - j} / \sum_{\ell=1}^{M_t}\gamma^{M_t-\ell}$, with $0 < \gamma \le 1$, and define the weighted empirical quantile

$$\widehat{q}^{\,w}_{t,1-\alpha} = \inf\left\{ q : \sum_{j=1}^{M_t} \omega_{j,t}\mathbf{1}\{S_j \le q\} \ge 1-\alpha \right\}.$$

The corresponding interval is $C_{\mathrm{TW}}(r_t) = \left[r_t^\top \widehat{w} - \widehat{q}^{w}_{t,1-\alpha},\ r_t^\top \widehat{w} + \widehat{q}^{w}_{t,1-\alpha}\right]$. We tune $\gamma \in \{0.95, 0.98, 0.99, 0.995, 0.997, 0.999\}$ using the calibration block: the first 60% supplies the initial score history and the final 40% is used for chronological validation. The value minimizing validation Winkler score is frozen before testing.

D.3.2 Adaptive Conformal Prediction

Let α_t denote the effective miscoverage level used for the interval at time t, initialized at $\alpha_1 = \alpha$, and define $\mathrm{err}_t = \mathbf{1}\{y_t \notin C_t(1-\alpha_t)\}$. After the corresponding response becomes observable, we update $\alpha_{t+1} = \Pi_{[0.001,0.999]}\left[\alpha_t + \eta(\alpha - \mathrm{err}_t)\right]$. At each forecast origin, the conformal radius is recomputed from all currently available absolute residuals using level $1-\alpha_t$. We tune $\eta \in \{0.001, 0.003, 0.005, 0.01, 0.02, 0.05\}$ with the same chronological 60%/40% calibration split and Winkler-score criterion used for time-weighted conformal prediction.

Adaptive CP and time-weighted CP are used here as adaptive empirical baselines. The projection of α_t onto $[0.001, 0.999]$ differs from the unprojected adaptive-conformal recursion for which a telescoping long-run coverage argument is available. We therefore do not claim that this argument transfers unchanged to the projected implementation. Likewise, the time-weighted method normalizes weights over the revealed historical residuals only and does not assign test-point mass at infinity. It is a weighted residual-quantile baseline, not the finite-sample weighted conformal construction, and no corresponding finite-sample coverage guarantee is claimed.

For direct H-step forecasting, neither adaptive nor time-weighted conformal prediction uses a score immediately after its forecast is issued. A test residual and, for adaptive conformal prediction, its coverage error enter the online history only after the corresponding target becomes observable, giving an H-step feedback delay. The unobservable final $H-1$ calibration targets are revealed sequentially under the same rule.

D.3.3 Reservoir Conformal Prediction

We additionally evaluate Reservoir Conformal Prediction (ResCP) Neglia et al. (2026). ResCP encodes the revealed signed-residual history in a random reservoir and compares the current reservoir state with previously recorded states. Sampling probabilities combine a temperature-scaled softmax of cosine similarities with the official implementation's linear recency weights. Residuals are sampled with replacement according to these probabilities, and empirical lower and upper quantiles of the sampled residuals define a potentially asymmetric interval around the same point forecast used by the other uncertainty-quantification methods.

We use the official stochastic sampler and reservoir implementation through a local residual-bundle adapter. All primary real-data experiments use a fixed reservoir configuration: reservoir dimension 512, spectral radius 1.2, leak rate 0.9, input scaling 0.25, connectivity 0.2, and zero reservoir noise. These reservoir parameters are not tuned. Residual inputs to the reservoir are standardized using only calibration residuals available at the corresponding forecast origin, while sampled residuals remain on the original standardized-target scale. We use cosine similarity, linear recency decay, and 100 candidate quantile pairs. The adaptive-miscoverage parameter internal to ResCP is set to zero.

Only the sampling temperature and residual-history policy are selected. The temperature grid is $T \in \{0.05, 0.10, 0.15, 0.25\}$. The candidate numerical history windows are $W \in \{100, 250, 500, 1000, 1500, 3200, 3800, 7600\}$, subject to the requirement that W not exceed the residual history available in the chronological tuning prefix. An all-history policy is included as a separate candidate. Thus, for example, the numerical window $W = 7600$ is not considered for the LA-ozone experiments because it exceeds the available tuning history.

For a direct H-step forecast, the first $H-1$ calibration targets are treated as boundary guards because their forecasts would have been issued before the fitting block closed. The final $H-1$ calibration outcomes are excluded from hyperparameter selection because they are not yet observable when the first test interval is issued. Temperature and history policy are selected on the final 10% of this availability-safe calibration sequence by first minimizing the validation Winkler score. Ties in Winkler score are broken by smaller absolute validation coverage error, with any remaining ties resolved by candidate-grid order. The selected configuration is then frozen before test evaluation.

Table 5: Hyperparameters selected for the primary fixed-origin experiments at nominal coverage 0.95. Here T is the ResCP sampling temperature and W is its residual-history window.

Dataset and horizon	λ	Adaptive η	TWCP γ	ResCP T	ResCP W
Beijing PM_{10}, $H = 1$	0.003	0.05	0.95	0.1	all
Beijing PM_{10}, $H = 3$	0.001	0.02	0.95	0.15	all
Beijing PM_{10}, $H = 7$	0.001	0.01	0.99	0.15	all
Exchange returns, $H = 1$	10	0.01	0.95	0.15	1500
Exchange returns, $H = 3$	10	0.02	0.95	0.25	100
Exchange returns, $H = 7$	10	0.01	0.95	0.15	1000
LA ozone, $H = 1$	0.03	0.001	0.999	0.1	3200
LA ozone, $H = 3$	1	0.001	0.999	0.1	all
LA ozone, $H = 7$	0.3	0.001	0.999	0.25	3800
Solar, $H = 1$	0.001	0.05	0.95	0.05	all
Solar, $H = 3$	0.001	0.001	0.999	0.05	7600
Solar, $H = 7$	0.001	0.001	0.999	0.05	7600

During testing, a forecast residual and its associated forecast-origin reservoir state enter the ResCP history only after the corresponding target has become observable. The initially unavailable calibration-tail residuals are revealed sequentially under the same rule. Consequently, ResCP never uses a response before its observation time.

Each primary dataset–horizon configuration uses one prespecified deterministic seed for the reservoir and stochastic sampler, recorded in the experiment manifest. The primary-table entries are conditional on these seeds. We separately assess stochastic sensitivity by repeating the final ResCP evaluation over five seeds while holding the point forecasts, data splits, and selected hyperparameters fixed; see Appendix F.8. The FIFO history variant used specifically in the variance-change simulation is described in Appendix E.4.3.

D.4 Hyperparameter Selection

All hyperparameters are selected chronologically, without using test responses. For NGRC, the ridge penalty is chosen by validation RMSE on the final 20% of the fitting block, as described in Appendix D.2. For $H > 1$, hyperparameter selection excludes the first $H - 1$ calibration targets used as boundary guards and the final $H - 1$ calibration targets that are not yet observable when the first test forecast is issued. For adaptive and time-weighted conformal prediction, the first 60% of the calibration block supplies the initial residual history and the final 40% is used for validation. The candidate minimizing validation Winkler score is retained, with ties resolved by candidate-grid order. ResCP uses the first 90% of the calibration block as its initial history and the final 10% for chronological validation. ResCP candidates are ordered first by validation Winkler score, then by absolute validation coverage error, and finally by candidate-grid order.

The candidate grids are

$$\begin{aligned}
\lambda &\in \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\},\\
\eta &\in \{0.001, 0.003, 0.005, 0.01, 0.02, 0.05\},\\
\gamma &\in \{0.95, 0.98, 0.99, 0.995, 0.997, 0.999\},\\
T_{\mathrm{ResCP}} &\in \{0.05, 0.10, 0.15, 0.25\},\\
W_{\mathrm{ResCP}} &\in \{100, 250, 500, 1000, 1500, 3200, 3800, 7600, \mathrm{all}\},
\end{aligned}$$

where a numerical ResCP window is omitted whenever it exceeds the residual history available in the first 90% of the availability-safe calibration sequence used as the chronological tuning prefix. Table 5 gives the values selected for the primary 95% fixed-origin experiments. The symbol all means that the complete revealed residual history is retained.

The exploratory forecaster comparison fixes ARIMA order and neural-network architectures in advance rather than selecting them on the test data. Neural training is stopped using only the fitting-block validation segment. The architecture and optimization settings are reported in Appendix D.2.

D.5 Evaluation Metrics

Let $\{(y_i, \widehat{y}_i, L_i, U_i)\}_{i=1}^n$ denote the test responses, point forecasts, and lower and upper prediction limits. Point accuracy is measured by

$$\mathrm{RMSE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \right\}^{1/2}.$$

Empirical coverage is

$$\widehat{\mathrm{Cov}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{L_i \le y_i \le U_i\}.$$

Mean prediction-interval width is

$$\widehat{\mathrm{Width}} = \frac{1}{n} \sum_{i=1}^{n} (U_i - L_i).$$

Tables report coverage error in percentage points as

$$\Delta\,\mathrm{Cov} = 100\{\widehat{\mathrm{Cov}} - (1-\alpha)\}.$$

Thus, positive values indicate overcoverage and negative values indicate undercoverage relative to the nominal level.

For one observation, the Winkler interval score is

$$W_{\alpha,i} = \begin{cases} U_i - L_i, & L_i \le y_i \le U_i, \\ (U_i - L_i) + \dfrac{2}{\alpha}(L_i - y_i), & y_i < L_i, \\ (U_i - L_i) + \dfrac{2}{\alpha}(y_i - U_i), & y_i > U_i. \end{cases}$$

The reported Winkler score is $n^{-1} \sum_{i=1}^{n} W_{\alpha,i}$; smaller values are better because the score penalizes both wide intervals and missed observations. The main tables use nominal 95% coverage ($\alpha = 0.05$). The one-step calibration curves additionally use nominal coverage levels 80%, 85%, 90%, and 97.5%.

D.6 Experimental Protocol

The primary real-data experiment uses direct forecast horizons $H \in \{1, 3, 7\}$. For each dataset and horizon, valid feature–target rows are ordered by target time and divided once into consecutive 40% fitting, 40% calibration, and 20% test blocks. NGRC fitting and ridge selection use only the fitting block. UQ tuning and initialization use only the calibration block, and all reported accuracy and uncertainty metrics are computed once on the untouched test block. The point predictor is fixed before calibration and testing. The calibration counts reported in Table 3 are allocated block sizes. For $H > 1$, the initially available calibration-score set contains $n_{\mathrm{cal}} - 2(H-1)$ scores after removing the two boundary segments.

Availability-safe multi-horizon protocol. For a direct H-step forecast, a residual becomes available only when its target is observed. The first $H-1$ calibration targets are boundary guards: their forecasts would have been issued before the fitting block closed, so they are not used as calibration scores. The final $H-1$ calibration targets are not observable when the first test interval is issued. Consequently, the initially available calibration-score set contains $n_{\mathrm{cal}} - 2(H-1)$ scores. Frozen split conformal prediction uses this initial set and remains fixed throughout testing. Hyperparameter selection uses the same availability-safe sequence. For adaptive CP, time-weighted CP, and ResCP, the initially unavailable calibration-tail residuals and subsequent test residuals enter the online history only when their targets become observable. Ridge validation likewise excludes the first $H-1$ validation targets following its inner fitting boundary. Thus no fitting, tuning, or uncertainty procedure uses a response before its observation time.

Runtime summaries use explicitly separated scopes. NGRC fitting time includes fitting-block scaling, ridge selection, and the final ridge refit, but excludes feature construction and disk input/output.

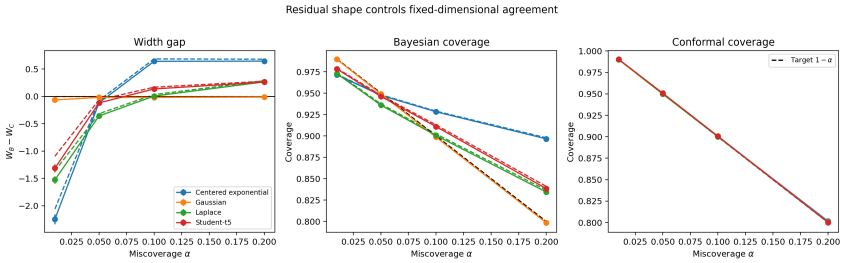


Figure 3: Fixed-dimensional width and coverage results across residual distributions and nominal miscoverage levels.

UQ runtime includes hyperparameter selection, where applicable, and construction of test intervals, but excludes point-predictor fitting. The exploratory forecaster benchmark additionally reports end-to-end pipeline time, including preprocessing, point-model fitting, forecasting, ResCP selection, and interval construction. All timings are wall-clock measurements from the same local experimental workflow and are intended for within-study comparison.

Computing environment. The final experiments were executed on an Apple M1 Pro processor with eight CPU cores and 16 GB of memory. Neural-network calculations used the CPU. The software environment used Python 3.11.1, NumPy 1.26.4, pandas 2.2.3, SciPy 1.16.3, scikit-learn 1.6.1, PyTorch 2.9.1, statsmodels 0.14.6, and Matplotlib 3.10.8. Exact dependencies and executable commands are included in the reproducibility repository.

E Additional Simulation Results

E.1 Experiment 1: Fixed-Dimensional Simulation Details

For each replication of Experiment 5.1, we generate

$$r_i \overset{\text{i.i.d.}}{\sim} N(0, I_p), \qquad y_i = r_i^\top w_0 + e_i,$$

with

$$p = 5, \qquad w_0 = (-0.4, -0.2, 0, 0.2, 0.4)^\top, \qquad \mathbb{E}[e_i^2] = 1.$$

We consider four unit-variance residual distributions: Gaussian $N(0,1)$, Laplace with scale $1/\sqrt{2}$, standardized Student-t_5 errors $\sqrt{3/5}\,T_5$ and centered exponential errors $E - 1$ with $E \sim \text{Exp}(1)$.

We use

$$\alpha \in \{0.01, 0.05, 0.10, 0.20\}, \qquad n_{\text{fit}} = 800, \qquad n_{\text{cal}} = 800, \qquad n_{\text{test}} = 3000,$$

with normalized ridge penalty $\lambda = 0.01$. Results are averaged over 300 independent replications.

For each configuration, we record the Bayesian–conformal width difference and marginal coverage of both procedures. The empirical width difference and Bayesian coverage are compared with the limits in Theorem 1 and Corollary 1, while split-conformal coverage is compared with $1 - \alpha$.

Figure 3 confirms the residual-shape mechanism. For Gaussian errors, $q_e = z\tau$, so the two interval widths nearly coincide and Bayesian coverage is close to nominal. For the non-Gaussian distributions, the residual root-mean-square scale and the relevant absolute-residual quantile generally differ, producing the predicted α-dependent width and Bayesian coverage discrepancies. Split-conformal coverage remains close to its nominal level across the configurations.

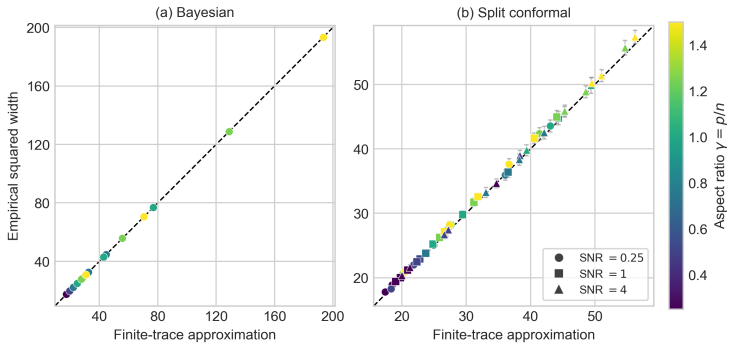


Figure 4: Individual finite-trace width verification for Experiment 2. Points compare empirical and theoretical squared widths for Bayesian intervals (left) and split-conformal intervals (right). Error bars are 95% confidence intervals over 100 replications. Color denotes $\gamma = p/n$; split-conformal markers denote signal-to-noise ratio. The dashed line denotes equality.

Table 6: Raw and ridge-transformed quadratic diffuseness diagnostics for Experiment 3. Entries are means $\pm$ standard errors over 150 replications. The generating-matrix norm refers to the unnormalized matrix M_d in $G^\top M_d G - \mathrm{tr}(M_d)$. The transformed diagnostic $\|D_n\|_{\mathrm{op}}$ uses the normalized theoretical convention $Q_d(w_0) = \sqrt{2}M_d$.

d	Signal structure	Generating-matrix $\|M_d\|_{\mathrm{op}}$	$\|D_n\|_{\mathrm{op}}$	Residual KS
8	Concentrated	1.414	0.973 ± 0.013	0.140 ± 0.001
8	Diffuse	0.500	0.545 ± 0.004	0.079 ± 0.001
12	Concentrated	1.414	0.941 ± 0.010	0.133 ± 0.001
12	Diffuse	0.408	0.444 ± 0.002	0.065 ± 0.001
16	Concentrated	1.414	0.923 ± 0.008	0.130 ± 0.001
16	Diffuse	0.354	0.386 ± 0.002	0.059 ± 0.001
20	Concentrated	1.414	0.901 ± 0.006	0.125 ± 0.001
20	Diffuse	0.316	0.346 ± 0.001	0.052 ± 0.001
24	Concentrated	1.414	0.885 ± 0.006	0.125 ± 0.001
24	Diffuse	0.289	0.318 ± 0.001	0.050 ± 0.001

E.2 Experiment 2: Verification of Individual Width Approximations

For the isotropic Gaussian setting of Section 5.2, we separately verify the finite-trace approximations $W_B^2 \approx 4z^2(\sigma^2 + P_n)$ and $W_C^2 \approx 4z^2(\sigma^2 + B_n + V_n)$ where $z = z_{1-\alpha/2}$. Figure 4 shows close agreement between empirical and theoretical squared widths across aspect ratios, ridge penalties, and signal strengths. Thus the accuracy of the approximation for $W_B^2 - W_C^2$ does not arise from cancellation between errors in the two individual width approximations.

E.3 Experiment 3: Additional Quadratic-Feature Diagnostics

E.3.1 Ridge-Transformed Diffuseness Diagnostic

The concentrated and diffuse constructions in Section 5.3 specify raw quadratic coefficient matrices with equal Frobenius norm. Corollary 3, however, imposes its diffuseness condition on the ridge-transformed bias direction rather than on the raw signal. We therefore record, in every replication of the standardized independent quadratic experiment, $\|D_n\|_{\mathrm{op}} = \left\| \mathcal{Q}_d\left(\lambda \Sigma_p^{1/2} A_n w_0\right) \right\|_{\mathrm{op}}$, and $A_n = \left(R^\top R/n_{\mathrm{fit}} + \lambda I_p\right)^{-1}$. For the centered and variance-standardized independent quadratic feature map, $\Sigma_p = I_p$. The map $\mathcal{Q}_d$ is the inverse of the isometric svec convention: diagonal coordinates are unchanged, while an off-diagonal upper-triangular coordinate is divided by $\sqrt{2}$ and placed symmetrically in the corresponding two matrix entries. Thus the reported value is the matrix operator norm appearing directly in the deterministic-readout corollary.

Table 7: Width, coverage, and residual-normality results for Experiment 3 at the largest base dimension, $d = 24$. Entries are means $\pm$ standard errors over 150 replications. Trace denotes the finite-trace approximation, and Residual KS is the Kolmogorov–Smirnov distance between the standardized prediction residual and a standard Gaussian distribution.

Feature design	Signal	Empirical W_B	Trace W_B	Empirical W_C	Trace W_C	Bayesian coverage	SCP coverage	Residual KS
Independent Gaussian	Concentrated	1.165 ± 0.000	1.165 ± 0.000	3.732 ± 0.016	3.715 ± 0.010	0.462 ± 0.001	0.950 ± 0.001	0.019 ± 0.000
Independent Gaussian	Diffuse	1.165 ± 0.000	1.166 ± 0.000	3.743 ± 0.015	3.747 ± 0.011	0.459 ± 0.002	0.949 ± 0.001	0.020 ± 0.000
Quadratic Gaussian	Concentrated	1.169 ± 0.000	1.173 ± 0.000	3.843 ± 0.035	3.870 ± 0.023	0.542 ± 0.002	0.950 ± 0.001	0.125 ± 0.001
Quadratic Gaussian	Diffuse	1.169 ± 0.000	1.173 ± 0.000	3.830 ± 0.015	3.872 ± 0.010	0.448 ± 0.001	0.951 ± 0.001	0.050 ± 0.001
Temporal quadratic	Concentrated	1.054 ± 0.001	1.055 ± 0.001	2.889 ± 0.024	2.886 ± 0.013	0.637 ± 0.002	0.949 ± 0.001	0.121 ± 0.001
Temporal quadratic	Diffuse	1.055 ± 0.001	1.057 ± 0.001	2.336 ± 0.032	2.373 ± 0.025	0.633 ± 0.005	0.942 ± 0.002	0.049 ± 0.001

Table 6 shows that the transformed direction becomes increasingly diffuse under the diffuse construction: its mean operator norm decreases from 0.545 at $d = 8$ to 0.318 at $d = 24$. The concentrated construction remains substantially larger, decreasing only from 0.973 to 0.885. Over the same dimensions, the diffuse residual-normality KS distance decreases from 0.079 to 0.050, whereas the concentrated value remains 0.125 at $d = 24$. These finite-dimensional results directly diagnose the condition in Corollary 3 and are consistent with the Gaussian-approximation mechanism in Lemma 5; they are not presented as proof of the asymptotic condition.

Table 7 reports the complementary width and coverage diagnostics at $d = 24$. The empirical and finite-trace Bayesian widths differ by at most 0.004, and the conformal-width approximations are also close. Split-conformal coverage remains near the nominal 0.95 level. The substantially smaller residual KS distance for the diffuse quadratic signal is consistent with the corresponding reduction in $\|D_n\|_{\mathrm{op}}$ reported in Table 6.

E.3.2 Supplementary Temporal-Transfer Stress Test

To supplement the quadratic-feature experiment 5.3, we examine whether a temporally dependent quadratic-window design approaches its same-marginal independent-window benchmark as the separation between forecast origins increases.

We generate a stationary Gaussian AR(1) process, $x_t = \rho x_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t$, with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$, and $\rho = 0.8$. At each retained forecast origin, we form a lag vector of base dimension $d = 20$ and construct its centered, variance-standardized quadratic coordinates. The resulting feature dimension is $p = (d(d+1))/2 = 210$. The signal is the diffuse diagonal quadratic form used in the main quadratic-feature experiment, with total signal magnitude 2. We use $n = m = 2p = 420$ fitting and calibration observations, 2,000 test observations, noise standard deviation $\sigma = 0.25$, normalized ridge penalty $\lambda = 0.5$, and miscoverage level $\alpha = 0.05$.

Forecast origins are retained at separations $s \in \{1, 4, 16, 64\}$. For comparison, we also generate independent Gaussian lag windows with the same within-window AR(1) covariance. This benchmark therefore preserves the marginal feature-response distribution while removing dependence across forecast origins. Results are averaged over 150 independent simulation replications.

We assess temporal transfer using three diagnostics. First, we compare the empirical split-conformal width W_C with the finite-trace approximation $W_C^{\mathrm{tr}} = 2z_{1-\alpha/2}\sqrt{\sigma^2 + B_n + V_n}$, where B_n and V_n are the finite-sample squared-bias and estimation-variance terms. Second, for each separation, we compare the distribution of W_C across replications with its independent-window counterpart using the two-sample Kolmogorov–Smirnov distance $D_{\mathrm{KS}} = \sup_x \left|\widehat{F}_s(x) - \widehat{F}_{\mathrm{ind}}(x)\right|$. We additionally report the KS distance between the standardized prediction residuals and a standard Gaussian distribution. Third, we report empirical split-conformal coverage at every separation.

Figure 5 shows that the empirical and finite-trace conformal widths remain close throughout the bridge. At separation one, the empirical and approximated widths are 2.397 and 2.364, respectively. At separations 4, 16, and 64, the corresponding empirical widths are 2.221, 2.236, and 2.212, compared with finite-trace approximations 2.228, 2.251, and 2.237. The independent-window benchmark has empirical width 2.218 and approximated width 2.240.

The KS distance between the conformal-width distribution and the independent-window benchmark decreases from 0.333 at separation one to 0.153, 0.107, and 0.067 at separations 4, 16, and 64, respectively. The residual-normality KS diagnostic similarly decreases from approximately 0.063

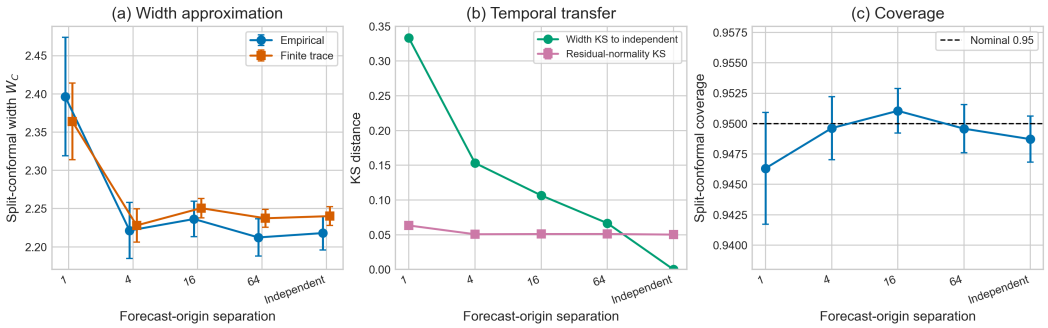


Figure 5: Temporal quadratic-window bridge for a stationary Gaussian AR(1) process with $\rho = 0.8$ and base dimension $d = 20$. Panel (a) compares the empirical split-conformal width with its finite-trace approximation. Panel (b) reports the KS distance between the conformal-width distribution and its independent-window benchmark, together with the residual-normality KS diagnostic. Panel (c) reports split-conformal coverage, with the dashed line denoting the nominal level 0.95. Error bars, where shown, are 95% confidence intervals for cell means over 150 replications.

at separation one to approximately 0.051 at the larger separations, close to the independent-window value 0.050.

Mean split-conformal coverage remains close to the nominal level throughout: 0.946, 0.950, 0.951, and 0.950 at separations 1, 4, 16, and 64, respectively, compared with 0.949 for independent windows. Increasing forecast-origin separation substantially reduces the cross-window dependence targeted by Proposition 1, while the width comparison and split-conformal calibration are relatively insensitive to the remaining dependence in this experiment.

E.3.3 Absolute-Regularity Penalties Under Temporal Dependence

We also examine the absolute-regularity correction terms appearing in Proposition 1. For the stationary Gaussian AR(1) process with $\rho = 0.8$, the lag-h absolute-regularity coefficient is computed numerically from the Gaussian transition distribution.

The current temporal bridge uses lag windows of base dimension $d = 20$. If consecutive forecast origins are separated by Δ time steps, the distance between the latest observation in the earlier window and the earliest observation in the later window is $h(\Delta) = \Delta - d + 1 = \Delta - 19$. Consequently, the simulated separations $\Delta \in \{1, 4, 16\}$ produce overlapping lag windows, and an absolute-regularity transfer bound between disjoint windows is not applicable to those configurations. The first nonoverlapping configuration occurs at $\Delta = 20$, for which $h = 1$. At the largest simulated separation, $\Delta = 64$, the corresponding boundary lag is $h = 45$.

The empirical conformal-width statistic uses $n = 420$ fitting and $m = 420$ calibration windows. Using the pooled count $N_W = n + m = 840$, the corresponding distributional correction is $(N_W - 1)\beta(h) = 839\beta(h)$. A single split-conformal coverage event additionally involves one test window, giving $N_C = n + m + 1 = 841$ and $\mathbb{P}\{Y_* \notin C_C(X_*)\} \le \alpha + (N_C - 1)\beta(h) = 0.05 + 840\beta(h)$. These pooled counts are conservative for the simulation implementation, which generates the fitting, calibration, and test blocks from separate AR(1) trajectories while retaining temporal dependence within each block.

Figure 6 shows that the sufficient transfer bounds are highly conservative at short nonoverlapping separations. The untruncated corrections remain greater than one until approximately $\Delta = 45$, corresponding to boundary lag $h = 26$. At the largest simulated separation, $\Delta = 64$, numerical integration gives $\beta(45) \approx 1.39 \times 10^{-5}$. The resulting width-distribution correction is approximately 0.0117, and the split-conformal miscoverage upper bound is approximately $0.05 + 840\beta(45) \approx 0.0617$. The observed mean split-conformal miscoverage at this separation is approximately 0.0504, consistent with this bound.

The empirical two-sample KS distance in Figure 5 is an estimate based on 150 replications per distribution and therefore need not lie below the population coupling bound pointwise. In particular, sim-

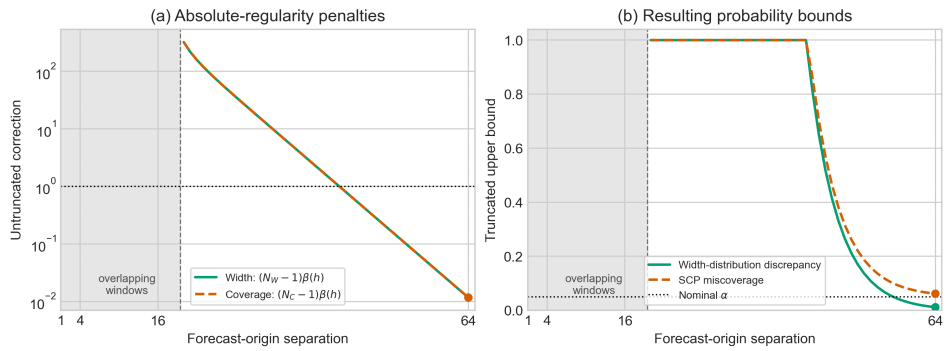


Figure 6: Absolute-regularity corrections for the temporal quadratic-window experiment with $\rho = 0.8$ and $d = 20$. The shaded region contains origin separations $\Delta < 20$, for which consecutive lag windows overlap and the disjoint-window transfer bound is not applied. Panel (a) shows the untruncated corrections $(N_W - 1)\beta(h)$ and $(N_C - 1)\beta(h)$; the horizontal dotted line marks one, above which the corresponding probability correction is vacuous. Panel (b) shows the resulting truncated upper bounds for the width-distribution discrepancy and split-conformal miscoverage. The horizontal dotted line in panel (b) marks the nominal miscoverage level $\alpha = 0.05$. Vertical dotted lines identify the four forecast-origin separations used in the simulation.

ulation variation can make the empirical KS statistic larger than the bound on the underlying distributional discrepancy. For the independent-window benchmark, $\beta(h) = 0$, so the width-distribution correction is zero and the miscoverage upper bound reduces to the nominal level $\alpha = 0.05$.

E.4 Experiment 4: Additional Online-Robustness Details

This section provides implementation details for the sequential uncertainty methods used in Experiment 5.4. All methods use the same fixed NGRC ridge point predictor. After the initial fit, the regression coefficients are not updated, so only the uncertainty estimates are allowed to adapt as test outcomes become available.

E.4.1 Updated Gaussian and Rolling SCP Baselines

Let $e_t = y_t - \widehat{y}_t$ denote the one-step-ahead forecast residual at test time t. When the interval for y_t is constructed, only residuals whose outcomes have already been observed are available. The sequential methods use a rolling window containing the $M = 100$ most recent available residuals. At the beginning of the test period, this window is initialized using the most recent residuals from the pre-test calibration block.

For comparison, the frozen Bayesian ridge interval is $C_t^{\mathrm{B}} = \left[\widehat{y}_t \pm z_{1-\alpha/2}\widehat{\tau}\sqrt{1+\ell_t}\right]$ with $\ell_t = r_t^\top \left(R_{\mathrm{fit}}^\top R_{\mathrm{fit}} + n_{\mathrm{fit}}\lambda I\right)^{-1} r_t$, where $\widehat{\tau}$ is the root-mean-square residual estimated from the fitting sample. The scale $\widehat{\tau}$ remains fixed after deployment, although the interval width may vary with the predictive-leverage term ℓ_t.

Updated Gaussian interval. Let $\mathcal{I}_t$ denote the indices of the M residuals available immediately before forecast t. The sequential forecast-error scale is estimated as $\widehat{\tau}_t^2 = \frac{1}{M}\sum_{j\in\mathcal{I}_t} e_j^2$. The updated Gaussian interval is $C_t^{\mathrm{UG}} = \left[\widehat{y}_t \pm z_{1-\alpha/2}\widehat{\tau}_t\right]$. Here $\widehat{\tau}_t$ estimates the total out-of-sample forecast-error scale, so an additional Bayesian leverage factor is not applied. This diagnostic therefore isolates the effect of replacing the frozen training-residual scale by a sequential estimate based on recently observed forecast errors.

Rolling split conformal prediction. Rolling split conformal prediction uses the same fixed point predictor but updates its residual quantile sequentially. For $j \in \mathcal{I}_t$, define the conformity scores $s_j = |e_j|$. Let $s_{t,(1)} \le \cdots \le s_{t,(M)}$ denote the order statistics of the M scores in the current residual window, and define $k = \lceil (M+1)(1-\alpha) \rceil$, with the convention $s_{t,(M+1)} = +\infty$. Let $q_t = s_{t,(k)}$

under this convention. The resulting interval is $C_t^{\mathrm{RSCP}} = [\widehat{y}_t - q_t, \widehat{y}_t + q_t]$. For $M = 100$ and $\alpha = 0.05$, $k = 96$, so the infinite-score convention is not active in the experiments reported here and the numerical results are unchanged. After y_t is observed, the residual e_t is appended to the rolling history and the oldest residual is discarded. Thus, the interval for y_t never uses its own outcome or any future response.

The updated Gaussian and rolling SCP procedures are diagnostic baselines. Comparing the updated Gaussian interval with frozen Bayesian ridge isolates the effect of sequential scale estimation, while comparing rolling SCP with frozen SCP isolates the effect of updating the empirical residual distribution.

Adaptive CP and time-weighted CP follow the update rules in Appendix D.3. In this experiment, adaptive CP uses learning rate $\eta = 0.01$, while time-weighted CP uses decay factor 0.98.

E.4.2 Numerical Simulation of the Nonlinear Volterra System

The exponentially damped oscillatory memory kernel in Experiment 5.4 admits a recursive state-space representation. Define

$$u(t) = \int_0^t e^{-0.2(t-s)} \cos\{2\pi(t-s)\} x(s)\, ds, \qquad v(t) = \int_0^t e^{-0.2(t-s)} \sin\{2\pi(t-s)\} x(s)\, ds.$$

Then $M(t) = u(t)$, and differentiation of the two convolution states gives

$$u'(t) = x(t) - 0.2u(t) - 2\pi v(t), \qquad v'(t) = 2\pi u(t) - 0.2v(t).$$

Hence the Volterra integro-differential equation can be written as the three-dimensional system

$$\begin{aligned} x'(t) &= \sin(t) - 0.1x(t) + 0.5\tanh\{u(t)\}, \\ u'(t) &= x(t) - 0.2u(t) - 2\pi v(t), \\ v'(t) &= 2\pi u(t) - 0.2v(t), \end{aligned}$$

with $x(0) = u(0) = v(0) = 0$.

We integrate all three states using forward Euler with internal step $h = 0.005$. Writing $t_j = jh$, the internal recursion is

$$\begin{aligned} x_{j+1} &= x_j + h\left[\sin(t_j) - 0.1x_j + 0.5\tanh(u_j)\right], \\ u_{j+1} &= u_j + h\left[x_j - 0.2u_j - 2\pi v_j\right], \\ v_{j+1} &= v_j + h\left[2\pi u_j - 0.2v_j\right]. \end{aligned}$$

The homogeneous Euler update for the two memory states has eigenvalues $1 - 0.2h \pm 2\pi h\,\mathrm{i}$, and therefore amplification factor $r(h) = \sqrt{(1-0.2h)^2 + (2\pi h)^2}$.

Stability of this update requires

$$r(h) < 1 \quad \Longleftrightarrow \quad h < \frac{2(0.2)}{0.2^2 + (2\pi)^2} \approx 0.01012.$$

Thus, applying the printed Euler recursion directly at the observation spacing 0.1 would be unstable because $r(0.1) \approx 1.164$. The selected internal step $h = 0.005$ satisfies the stability condition ($r(0.005) \approx 0.99949$).

We retain every 20th internal state, so the observation spacing remains $\Delta t = 20h = 0.1$. The first 500 retained states (equivalently, the first 10,000 internal Euler steps) are discarded as burn-in before constructing the fitting, calibration, and test samples. Observation noise is added only after generating the latent trajectory, so the constant-variance and variance-change conditions use

the same underlying path. Within each paired replication, the two conditions also use the same standardized Gaussian noise innovations. They differ only because the noise standard deviation for test observations 160 through 799 is 0.35 in the variance-change condition and remains 0.15 in the constant-variance condition. As defined in Section 5.4, observations 160 through 799 in the variance-change condition form the high-noise period.

As a numerical-accuracy check, we repeated the complete 100-replication experiment with the internal step halved to $h = 0.0025$, retaining every 40th state so that the observation grid remained unchanged. Across the 3,221 retained latent states, the maximum absolute trajectory difference was 1.26×10^{-3} and the root-mean-square difference was 8.86×10^{-4}, or 0.044% of the trajectory range. Across the downstream summaries, the largest absolute change in a coverage quantity was 3.6×10^{-4} and the largest change in an interval-width quantity was 6.3×10^{-5}. Within the variance-change condition, all recovery fractions were unchanged, and the largest change in a conditional mean recovery delay was 2.24 observations. The independently selected ResCP configurations were also identical. Thus, the retained trajectories and downstream uncertainty results are insensitive to this further refinement of the integration step.

E.4.3 ResCP Implementation Details

We use the ResCP construction described in Appendix D.3.3. For Experiment 5.4, the principal difference is the tuning protocol. The common reservoir and sampling hyperparameters are selected once on an independent constant-variance pilot sequence by minimizing validation Winkler score, and are then held fixed across all constant-variance and variance-change evaluation replications.

We report an all-history version that retains every revealed residual and the distribution-shift treatment described for ResCP, which initializes a calibration set with the $N = 800$ pre-test residuals and keeps its size fixed by FIFO replacement during testing. Both versions apply ResCP's linear recency weighting. The two implementations therefore differ only in whether the calibration history grows or remains fixed at 800 residuals. As for the other online methods, a forecast residual is incorporated only after its corresponding outcome has been observed.

E.4.4 Recovery Criterion and Summary Statistics

The post-shift high-noise period consists of test observations 160–799. We report coverage over the first 50 high-noise observations, coverage and mean interval width over the complete high-noise period, and sustained-recovery diagnostics.

Recovery delay is measured from the noise increase at test step 160. It is the number of subsequent test observations until coverage over the trailing window of 100 test observations first enters $[0.93, 0.97]$ and remains in that interval for every remaining test observation. Recovery fraction is the number of replications satisfying this sustained-recovery criterion. This is a stringent retrospective stability criterion rather than a direct estimate of an intrinsic adaptation time: sampling variation can cause a calibrated procedure to leave the band, and the reported recovery delay is defined only among replications that eventually satisfy the criterion. We therefore interpret recovery delay jointly with the corresponding recovery fraction.

Because the change occurs at test index 160, an absolute recovery index can be obtained by adding 160 to the reported delay. For example, Adaptive CP's mean delay of 546.6 corresponds to mean test index 706.6 among recovering replications, while all-history ResCP's mean delay of 610.3 corresponds to index 770.3.

Figure 7 and Table 8 show that most methods are close to nominal coverage under constant variance, although both ResCP variants undercover in this configuration. Behavior during the high-noise period differs substantially according to whether uncertainty is updated online. Frozen SCP and frozen-scale Bayesian ridge exhibit persistent undercoverage after the noise increase, with neither method recovering in any replication. Both ResCP variants also suffer substantial initial undercoverage.

FIFO updating gives a modest improvement over all-history ResCP: full high-noise coverage rises from 0.815 to 0.830, and the recovery fraction rises from $11/100$ to $17/100$, at the cost of increasing high-noise width from 1.127 to 1.177. Thus the shift treatment helps here, but adaptation remains partial.

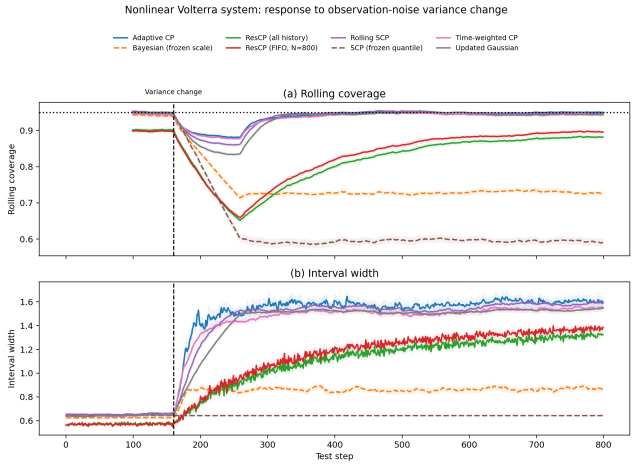


Figure 7: Online response to an observation-noise variance change in the nonlinear Volterra system. The vertical dashed line marks the increase in observation-noise standard deviation from 0.15 to 0.35 at test step 160, and the horizontal dotted line denotes nominal 95% coverage. Panel (a) reports trailing-window coverage and panel (b) reports mean interval width. Curves are averages over 100 paired replications and shaded regions are 95% confidence intervals.

Table 8: Performance under constant observation-noise variance and after the noise increase in the nonlinear Volterra system. First-50 high-noise coverage is calculated over test observations 160–209. Full high-noise coverage and high-noise width use test observations 160–799. Recovery fraction is the number of replications satisfying the sustained-recovery criterion. Observations to recovery counts the number of test observations from the noise increase until sustained recovery and is reported as mean $\pm$ standard error among replications that recover. All other entries are means $\pm$ standard errors over 100 paired replications.

Method	Constant coverage	Constant width	First-50 high-noise coverage	Full high-noise coverage	High-noise width	Recovery fraction	Observations from noise increase to recovery
Bayesian (frozen scale)	0.941 ± 0.001	0.624 ± 0.001	0.702 ± 0.005	0.726 ± 0.001	0.858 ± 0.003	$0/100$	–
SCP (frozen quantile)	0.949 ± 0.001	0.645 ± 0.002	0.606 ± 0.008	0.595 ± 0.003	0.645 ± 0.002	$0/100$	–
Rolling SCP	0.950 ± 0.000	0.656 ± 0.002	0.794 ± 0.004	0.935 ± 0.001	1.502 ± 0.006	$62/100$	592.6 ± 5.1
Updated Gaussian	0.947 ± 0.001	0.643 ± 0.002	0.756 ± 0.005	0.929 ± 0.001	1.458 ± 0.005	$58/100$	587.2 ± 5.5
Adaptive CP	0.949 ± 0.000	0.650 ± 0.002	0.832 ± 0.003	0.938 ± 0.000	1.544 ± 0.007	$80/100$	546.6 ± 9.8
Time-weighted CP	0.945 ± 0.000	0.642 ± 0.002	0.828 ± 0.004	0.933 ± 0.000	1.476 ± 0.006	$76/100$	574.4 ± 7.6
ResCP (all history)	0.911 ± 0.001	0.583 ± 0.002	0.602 ± 0.006	0.815 ± 0.001	1.127 ± 0.005	$11/100$	610.3 ± 13.7
ResCP (shift-aware FIFO, $N = 800$)	0.901 ± 0.001	0.571 ± 0.002	0.602 ± 0.006	0.830 ± 0.001	1.177 ± 0.005	$17/100$	603.3 ± 9.8

In contrast, adaptive CP, rolling SCP, time-weighted CP, and the updated Gaussian interval widen after the noise increase and recover much closer to nominal coverage, although at the cost of wider intervals. Under the sustained-recovery criterion, Adaptive CP has the shortest conditional mean recovery delay and the largest recovery fraction in this experiment. The comparison with Updated Gaussian shows that once the Gaussian interval is allowed to update its scale online, its behavior is similar to the adaptive conformal procedures. Thus, access to recently observed forecast errors and sequential updating of the uncertainty estimate are the primary mechanisms associated with improved robustness to the noise increase.

E.5 Experiment 5: Fitting–Calibration Tradeoff Under Controlled Data Budgets

Finally, we examine how reserving observations for conformal calibration affects forecasting performance under a controlled data budget. We consider $y_i = x_i^\top \beta_0 + \varepsilon_i$ with $x_i \sim N(0, I_p)$ and $\varepsilon_i \sim N(0, \sigma^2)$. We use a total pre-test budget $B = 400$, $\sigma^2 = 0.01$, and $\|\beta_0\|^2 = 1$. We compare a low-dimensional setting with $p = 10$ ($p/B = 0.025$) and a high-dimensional setting with $p = 300$ ($p/B = 0.75$). Each replication uses an independent test set of 2,000 observations.



Figure 8: Fitting–calibration tradeoff under the fixed total budget $B = 400$. Bayesian ridge uses all observations for fitting, while split conformal varies the fitting fraction ρ. Columns show coverage, interval width, Winkler score, and test RMSE for $p = 10$ and $p = 300$. Results are means over 200 paired replications with 95% confidence intervals.

Throughout this experiment, ridge regression uses the same normalized convention $\widehat{\beta} = \arg\min_\beta \left\{ \frac{1}{n_{\text{fit}}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$, with $\lambda = 2.5 \times 10^{-4}$. The normalized penalty is held fixed as the fitting allocation changes, so differences across ρ reflect the fitting–calibration tradeoff rather than an implicit change in regularization strength.

Equal total budget. Bayesian ridge uses all $B = 400$ observations for fitting, whereas split conformal prediction allocates ρB observations to fitting and $(1-\rho)B$ observations to calibration, with $\rho \in \{0.5, 0.6, 0.7, 0.8\}$. Within each replication, all allocation fractions use the same pre-test sample and test set. Results are averaged over 200 paired replications, and we report coverage, average interval width, Winkler score, and test RMSE.

Figure 8 shows that the allocation cost is small in low dimension but becomes substantial when the fitting problem is high dimensional. Increasing ρ improves split-conformal prediction by providing more observations for fitting, but across the considered allocation range its interval width, Winkler score, and RMSE remain larger than those of Bayesian ridge while both procedures maintain near-nominal coverage. Thus, reserving observations for calibration can carry a meaningful efficiency cost when the feature dimension is large relative to the available fitting sample.

Unequal total budget. As a sensitivity analysis, we fix the split-conformal budget at $N_C = 400$ and vary the Bayesian budget over $N_B/N_C \in \{0.5, 0.75, 1, 1.5, 2\}$. Split conformal is evaluated at $\rho \in \{0.5, 0.6, 0.7, 0.8\}$, while Bayesian samples are nested across budget ratios within each replication. All procedures share the same test set.

Figure 9 shows that the equal-budget Bayesian advantage does not persist uniformly as its fitting budget changes. In the high-dimensional setting with $p = 300$, the first three Bayesian budgets give $p/N_B = 1.5$, 1, and 0.75, respectively. Thus the two smallest budgets place the estimator at or beyond the interpolation threshold, producing wide, conservative intervals and substantially worse prediction performance. Performance improves sharply once the fitting sample becomes sufficiently large relative to the feature dimension. The unequal-budget comparison is therefore primarily an aspect-ratio sensitivity analysis rather than evidence of uniform Bayesian dominance or inferiority.

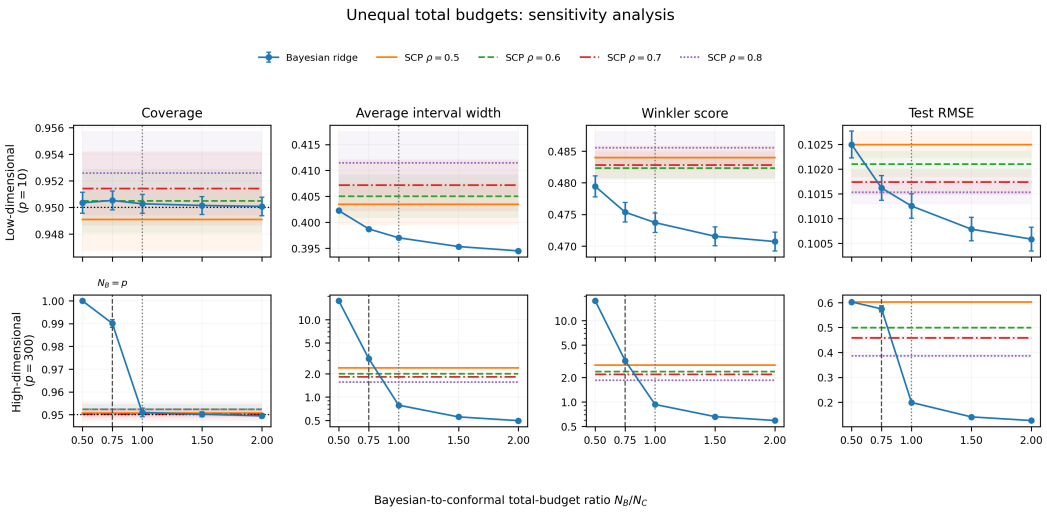


Figure 9: Sensitivity to unequal total data budgets. Split conformal uses $N_C = 400$, while Bayesian ridge varies over $N_B/N_C \in \{0.5, 0.75, 1, 1.5, 2\}$. The vertical dotted line marks equal budgets, and the dashed line in the high-dimensional row marks $N_B = p = 300$. Results are means over 200 paired replications with 95% confidence intervals.

F Additional Real-Data Details

This section provides implementation details and supplementary analyses for Section 6. We use the same chronological, leakage-safe evaluation protocol described in the main text. The material below records dataset-specific feature construction, tuning procedures, sequential update rules, and robustness analyses that are supplemental to the main text. The dataset definitions, preprocessing steps, NGRC construction, uncertainty- quantification procedures, hyperparameter selection, and leakage-safe chronological protocol are described in Appendix D. This section reports supplementary real-data diagnostics and robustness analyses.

F.1 Supplementary Theory-Facing Diagnostics

The theoretical analysis penalizes every coordinate through λI_p. The empirical NGRC fits instead leave the intercept unpenalized, using $D = \mathrm{diag}(0, 1, \ldots, 1)$. Accordingly, for the empirical diagnostics we define $d_{\mathrm{eff}} = \mathrm{tr}\{S(S + \lambda D)^{-1}\}$, and $S = \frac{R^\top R}{n_{\mathrm{fit}}}$. This is equivalently the trace of the actual fitted smoother and is the expression used to generate the reported diagnostic. The use of D here matches the empirical fitting convention; the theoretical effective-dimension calculations continue to use I_p.

For each dataset, Table 9 reports p/n_{fit}, the regularization-adjusted ratio $d_{\mathrm{eff}}/n_{\mathrm{fit}}$, residual RMS scales in the fitting, calibration, and test blocks, the calibration residual quantile $\widehat{q}_{1-\alpha,\mathrm{cal}}$, and $\rho_q = \widehat{q}_{1-\alpha,\mathrm{cal}}/z_{1-\alpha/2}\widehat{\tau}_{\mathrm{fit}}$.

The diagnostic combines a within-calibration residual-shape effect with a between-block scale effect. Exactly,

$$\rho_q = \underbrace{\frac{\widehat{q}_{1-\alpha,\mathrm{cal}}}{z_{1-\alpha/2}\widehat{\tau}_{\mathrm{cal}}}}_{\text{within-calibration shape factor}} \quad \underbrace{\frac{\widehat{\tau}_{\mathrm{cal}}}{\widehat{\tau}_{\mathrm{fit}}}}_{\text{fit-to-calibration scale factor}}.$$

The first factor compares the calibration residual quantile with the Gaussian radius based on the calibration RMS itself, while the second measures a change in residual RMS between the fitting and calibration periods. Thus ρ_q does not by itself isolate residual shape. Predictive leverage and other finite-sample effects additionally prevent it from being an exact Bayesian–SCP width ratio.

Table 9: Theory-facing diagnostics for the one-step, 95% fixed-origin analysis. RMS scales and residual quantiles are on the standardized-response scale.

Dataset	n_{fit}	p/n_{fit}	$d_{\mathrm{eff}}/n_{\mathrm{fit}}$	$\widehat{\tau}_{\mathrm{fit}}$	$\widehat{\tau}_{\mathrm{cal}}$	$\widehat{\tau}_{\mathrm{test}}$	$\widehat{q}_{.95,\mathrm{cal}}$	$\widehat{q}/(z\widehat{\tau}_{\mathrm{fit}})$
LA ozone	6420	0.0394	0.0287	0.521	0.392	0.360	0.787	0.771
Solar	21009	0.0338	0.0038	0.029	0.026	0.031	0.057	1.001
Beijing PM_{10}	14016	0.0238	0.0116	0.231	0.248	0.182	0.388	0.858
Exchange returns	3029	0.0396	0.0038	0.993	1.445	1.284	2.704	1.389

Table 10: Decomposition of the residual diagnostic at 95% nominal coverage, computed from the underlying quantities in Table 9.

Dataset	Shape factor	Scale factor	ρ_q
LA ozone	1.024	0.752	0.771
Solar	1.119	0.897	1.001
Beijing PM_{10}	0.798	1.074	0.858
Exchange returns	0.955	1.455	1.389

LA ozone and Beijing have $\rho_q = 0.771$ and 0.858, respectively, so the fit-block Gaussian RMS scale is large relative to the calibration residual quantile and the Bayesian intervals are correspondingly wider. Solar has $\rho_q = 1.001$, indicating nearly equal Gaussian and conformal residual radii, while Exchange returns has $\rho_q = 1.389$, for which SCP is wider. The differences among $\widehat{\tau}_{\mathrm{fit}}$, $\widehat{\tau}_{\mathrm{cal}}$, and $\widehat{\tau}_{\mathrm{test}}$ also provide direct evidence of temporal changes in the residual scale.

The decomposition sharpens the interpretation of the empirical width ordering. For LA ozone, the within-calibration shape factor is close to one while the calibration RMS is substantially smaller than the fitting RMS, so the Bayesian–SCP difference is driven mainly by a between-block scale change. For Exchange returns, the shape factor is again close to one, whereas the calibration RMS is about 46% larger than the fitting RMS, making temporal scale change the dominant contribution to $\rho_q > 1$. Beijing PM_{10} shows a more substantial within-block shape effect, while Solar exhibits offsetting shape and scale contributions. Hence these data support a broader residual-distribution interpretation of the fixed-dimensional comparison, but they do not constitute a clean empirical isolation of the stationary residual-shape mechanism.

For reference, the one-step fitting and calibration sample sizes and feature dimensions are

Dataset	n_{fit}	n_{cal}	p
LA ozone	6420	6420	253
Solar	21009	21009	711
Beijing PM_{10}	14016	14016	333
Exchange returns	3029	3029	120

where the feature dimension includes the intercept and any deterministic seasonal covariates used by the corresponding NGRC specification. These dimensions confirm that the empirical applications are much closer to the fixed- or effective-low-dimensional regime than to the proportional-growth benchmark developed in Section 4.

F.2 Complete Fixed-Origin Results

Table 11 gives the complete fixed-origin results underlying the condensed comparison in Table 1. In addition to coverage, interval width, and Winkler score, we report point RMSE, signed coverage error, and uncertainty-quantification runtime. The asymmetric SCP baseline is included here to isolate the effect of signed residual asymmetry from the additional localization, weighting, and sequential adaptation used by ResCP.

Symmetric and asymmetric SCP are nearly indistinguishable for these applications, indicating that signed-residual asymmetry alone does not explain the larger differences between SCP and ResCP. Computational costs also vary substantially: Bayesian ridge and ordinary SCP require little additional work after fitting the shared NGRC predictor, whereas adaptive CP, time-weighted CP,

Table 11: Complete fixed-origin one-step NGRC uncertainty-quantification results for the 40% fit/40% calibration/20% test split at 95% nominal coverage. RMSE is reported once per dataset because all methods use the same NGRC point forecasts. The signed coverage error is $\Delta_{\mathrm{cov}} = 100(\widehat{\mathrm{Cov}} - 0.95)$ percentage points. UQ time includes method-specific calibration, tuning, sequential updating, and interval construction, but excludes the shared NGRC fit.

Dataset	Method	RMSE	Coverage	Δ_{cov} (pp)	Width	Winkler	UQ time (s)
LA ozone	Bayesian ridge	0.360	0.993	4.25	2.047	2.083	0.015
	SCP		0.963	1.29	1.573	1.795	< 0.001
	Asymmetric SCP		0.963	1.29	1.574	1.791	0.001
	Adaptive CP		0.954	0.36	1.485	1.784	6.523
	Time-weighted CP		0.954	0.36	1.477	1.782	8.677
	ResCP		0.929	-2.13	1.369	1.797	18.831
Solar	Bayesian ridge	0.031	0.941	-0.94	0.115	0.205	0.115
	SCP		0.940	-0.98	0.115	0.205	0.001
	Asymmetric SCP		0.933	-1.70	0.112	0.209	0.002
	Adaptive CP		0.943	-0.66	0.109	0.167	73.168
	Time-weighted CP		0.857	-9.31	0.113	0.176	99.132
	ResCP		0.961	1.08	0.058	0.075	256.402
Beijing PM_{10}	Bayesian ridge	0.182	0.971	2.10	0.909	1.174	0.025
	SCP		0.960	1.00	0.776	1.138	0.001
	Asymmetric SCP		0.960	1.05	0.775	1.138	0.001
	Adaptive CP		0.947	-0.32	0.726	1.048	32.816
	Time-weighted CP		0.936	-1.36	0.677	1.030	45.449
	ResCP		0.954	0.43	0.656	0.933	125.954
Exchange returns	Bayesian ridge	1.284	0.946	-0.41	3.981	6.468	0.002
	SCP		0.983	3.35	5.407	6.997	< 0.001
	Asymmetric SCP		0.983	3.35	5.391	6.999	< 0.001
	Adaptive CP		0.954	0.38	4.015	6.396	1.420
	Time-weighted CP		0.939	-1.14	4.281	6.242	1.846
	ResCP		0.934	-1.6	3.793	6.640	4.890

and ResCP require sequential processing and, where applicable, pre-test hyperparameter selection. These runtimes are UQ-only measurements and should not be compared directly with the end-to-end architecture runtimes in Table 16.

F.3 Joint Coverage and Efficiency

Figure 10 displays coverage error against both mean interval width and Winkler score for the fixed-origin analysis. Narrow intervals should be interpreted jointly with calibration: points to the left of zero undercover and therefore should not be interpreted as more efficient at equal coverage. In particular, ResCP's relatively narrow ozone and Exchange intervals accompany coverage shortfalls, whereas its Solar and Beijing intervals combine smaller widths with coverage closer to the target.

F.4 Repeated chronological evaluation

Table 1 provides the detailed comparison for one fixed forecast period. To assess whether its conclusions depend on that particular period, we separately repeat the comparison over three shifted 40% fit/20% calibration/20% test windows. Table 12 reports the mean and standard deviation across these forecast periods. Because this robustness analysis uses a shorter calibration block, its values should not be compared row-for-row with Table 1; its purpose is to quantify sensitivity to temporal shifts.

The repeated evaluation changes the emphasis relative to a single favorable test period. Adaptive CP is comparatively stable across forecast periods. Mean 95% coverage ranges from 0.943 to 0.953, with small between-origin variation for ozone, Solar, and Beijing. Bayesian ridge is conservative for ozone and Beijing but undercovers on average for Solar and Exchange returns, consistent with the block-to-block RMS changes in Table 9. SCP and asymmetric SCP are nearly indistinguishable for these series, so asymmetry alone does not explain the larger differences observed for ResCP.

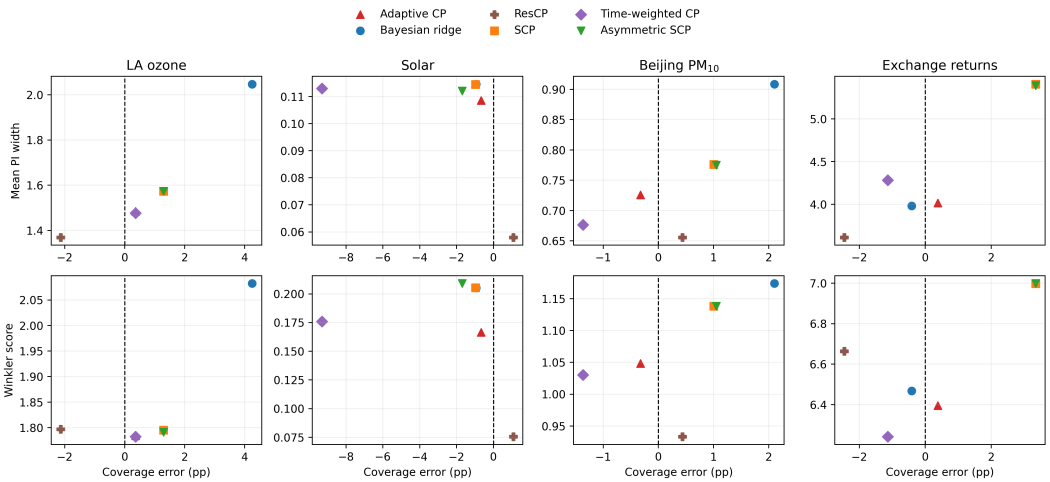


Figure 10: Coverage error versus mean interval width (top) and Winkler score (bottom) at 95% nominal coverage. The vertical line marks exact marginal coverage. All methods use the same NGRC point forecasts.

Table 12: Repeated chronological one-step results at 95% nominal coverage. Entries are means ± standard deviations over three shifted forecast periods. RMSE is reported once per dataset because all UQ methods use the same NGRC point predictor within each origin. Because the windows overlap, the standard deviations quantify forecast-period sensitivity rather than sampling uncertainty or standard errors.

Dataset	Method	RMSE	Coverage	Width	Winkler
LA ozone	Bayesian ridge	0.383 ± 0.003	0.981 ± 0.006	1.927 ± 0.123	2.054 ± 0.087
	SCP		0.959 ± 0.008	1.639 ± 0.057	1.929 ± 0.008
	Asymmetric SCP		0.958 ± 0.006	1.630 ± 0.047	1.927 ± 0.011
	Adaptive CP		0.953 ± 0.001	1.587 ± 0.005	1.921 ± 0.011
	Time-weighted CP		0.953 ± 0.001	1.586 ± 0.002	1.921 ± 0.011
	ResCP		0.947 ± 0.005	1.542 ± 0.032	1.936 ± 0.010
Solar	Bayesian ridge	0.034 ± 0.008	0.905 ± 0.064	0.111 ± 0.004	0.233 ± 0.081
	SCP		0.920 ± 0.038	0.117 ± 0.007	0.221 ± 0.061
	Asymmetric SCP		0.922 ± 0.030	0.116 ± 0.008	0.217 ± 0.055
	Adaptive CP		0.943 ± 0.003	0.119 ± 0.022	0.168 ± 0.009
	Time-weighted CP		0.875 ± 0.015	0.123 ± 0.026	0.182 ± 0.032
	ResCP		0.922 ± 0.025	0.057 ± 0.006	0.086 ± 0.005
Beijing PM_{10}	Bayesian ridge	0.203 ± 0.028	0.970 ± 0.003	0.879 ± 0.046	1.260 ± 0.127
	SCP		0.953 ± 0.007	0.723 ± 0.097	1.239 ± 0.157
	Asymmetric SCP		0.953 ± 0.007	0.723 ± 0.101	1.239 ± 0.158
	Adaptive CP		0.945 ± 0.000	0.711 ± 0.027	1.059 ± 0.075
	Time-weighted CP		0.936 ± 0.002	0.684 ± 0.047	1.073 ± 0.103
	ResCP		0.948 ± 0.007	0.614 ± 0.013	0.966 ± 0.097
Exchange returns	Bayesian ridge	1.318 ± 0.268	0.930 ± 0.062	4.106 ± 0.124	7.653 ± 2.801
	SCP		0.957 ± 0.060	5.416 ± 0.814	8.276 ± 2.010
	Asymmetric SCP		0.956 ± 0.059	5.415 ± 0.823	8.280 ± 1.993
	Adaptive CP		0.951 ± 0.011	4.780 ± 1.606	6.997 ± 1.948
	Time-weighted CP		0.946 ± 0.003	4.591 ± 1.468	6.541 ± 1.548
	ResCP		0.925 ± 0.050	4.168 ± 0.865	7.355 ± 2.268

The Exchange results have much larger between-origin variation for every nonadaptive method, illustrating the practical importance of temporal shift.

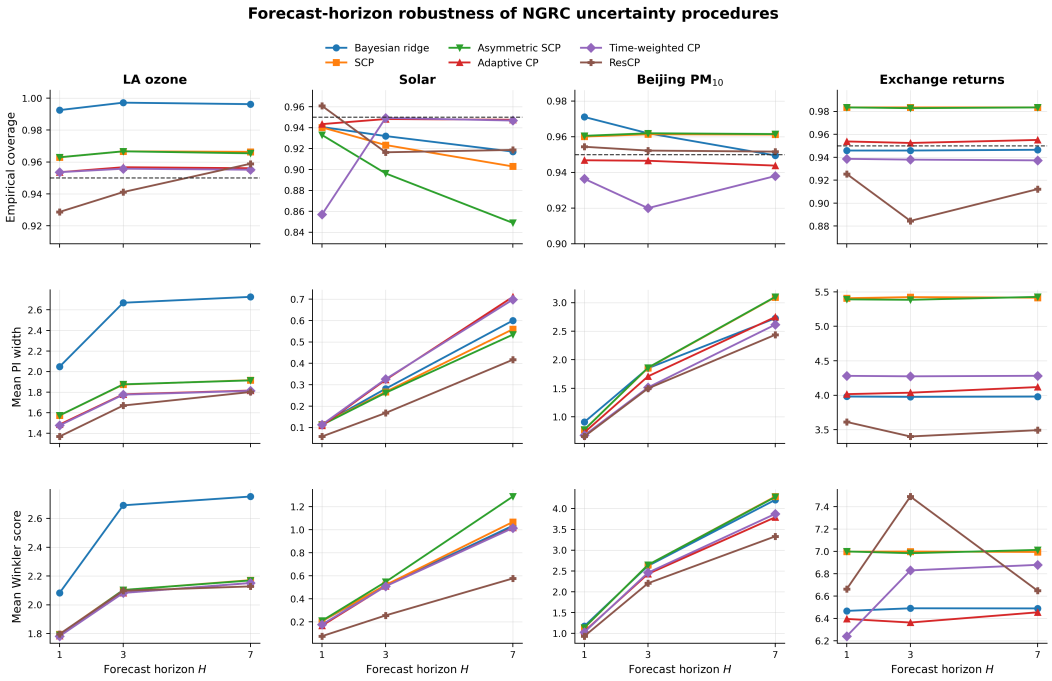


Figure 11: Forecast-horizon robustness of the NGRC uncertainty procedures for $H \in \{1, 3, 7\}$. Columns correspond to datasets; rows report empirical coverage, mean prediction-interval width, and mean Winkler score. The dashed horizontal line in the coverage panels marks the 95% nominal level. All sequential updates are delayed until the corresponding target is observed.

F.5 Forecast-horizon robustness

Using the same fixed-origin split, we additionally evaluate direct forecast horizons $H \in \{1, 3, 7\}$ to assess whether the one-step conclusions persist as the prediction horizon increases. Figure 11 reports empirical coverage, mean interval width, and Winkler score for all six procedures. For $H > 1$, every sequential method uses the delayed-update protocol described above: a residual enters the online calibration state only after its target has been observed.

Under the availability-safe boundary rule, Adaptive CP remains close to nominal coverage across horizons. ResCP has relatively low Winkler scores for LA ozone and Beijing PM_{10}, although its Exchange-return coverage is 0.894 at $H = 3$ and 0.931 at $H = 7$. For Solar, uncertainty grows rapidly with the horizon; ResCP coverage is 0.915 at $H = 3$ and 0.919 at $H = 7$, while ordinary SCP also undercovers. ResCP hyperparameter selection considers only numerical history windows that do not exceed the residual history available in the chronological tuning prefix. Under this protocol, the LA-ozone $H = 7$ configuration selects $T = 0.25$ and $W = 3800$. Thus good one-step performance does not imply uniform multi-horizon calibration.

F.6 Calibration across nominal levels

Figure 12 compares empirical and nominal coverage at nominal levels 80%, 85%, 90%, 95%, and 97.5%. A method whose 95% result appears satisfactory can still show systematic miscalibration elsewhere on the curve. No procedure is uniformly calibrated across datasets and levels.

F.7 Feature-construction ablation

The feature ablation compares linear lags, quadratic NGRC features, three single-scale memory summaries, and a multiscale memory representation. Table 14 reports point and interval metrics rather than RMSE alone.

Table 13: Complete availability-safe multi-horizon NGRC uncertainty-quantification results at 95% nominal coverage.

LA ozone					**Solar**				
H	Method	Cov.	Width	Winkler	H	Method	Cov.	Width	Winkler
1	Bayesian ridge	0.993	2.047	2.083	1	Bayesian ridge	0.941	0.115	0.205
	SCP	0.963	1.573	1.795		SCP	0.940	0.115	0.205
	Adaptive CP	0.954	1.485	1.784		Adaptive CP	0.943	0.109	0.167
	Time-weighted CP	0.954	1.477	1.782		Time-weighted CP	0.857	0.113	0.176
	ResCP	0.929	1.369	1.797		ResCP	0.961	0.058	0.075
3	Bayesian ridge	0.997	2.669	2.691	3	Bayesian ridge	0.932	0.281	0.513
	SCP	0.967	1.874	2.102		SCP	0.923	0.267	0.520
	Adaptive CP	0.957	1.778	2.083		Adaptive CP	0.948	0.323	0.509
	Time-weighted CP	0.956	1.775	2.082		Time-weighted CP	0.949	0.326	0.509
	ResCP	0.953	1.761	2.102		ResCP	0.915	0.168	0.256
7	Bayesian ridge	0.996	2.726	2.752	7	Bayesian ridge	0.917	0.600	1.035
	SCP	0.966	1.916	2.171		SCP	0.903	0.560	1.067
	Adaptive CP	0.956	1.816	2.151		Adaptive CP	0.948	0.711	1.018
	Time-weighted CP	0.955	1.812	2.152		Time-weighted CP	0.947	0.698	1.014
	ResCP	0.950	1.765	2.123		ResCP	0.919	0.417	0.578
Beijing PM_{10}					**Exchange returns**				
H	Method	Cov.	Width	Winkler	H	Method	Cov.	Width	Winkler
1	Bayesian ridge	0.971	0.909	1.174	1	Bayesian ridge	0.946	3.981	6.468
	SCP	0.960	0.776	1.138		SCP	0.983	5.407	6.997
	Adaptive CP	0.947	0.726	1.048		Adaptive CP	0.954	4.015	6.396
	Time-weighted CP	0.936	0.677	1.030		Time-weighted CP	0.939	4.281	6.242
	ResCP	0.954	0.656	0.933		ResCP	0.934	3.793	6.640
3	Bayesian ridge	0.962	1.851	2.622	3	Bayesian ridge	0.946	3.976	6.491
	SCP	0.961	1.856	2.645		SCP	0.983	5.423	6.998
	Adaptive CP	0.946	1.709	2.429		Adaptive CP	0.952	4.037	6.364
	Time-weighted CP	0.920	1.515	2.459		Time-weighted CP	0.938	4.274	6.830
	ResCP	0.951	1.496	2.208		ResCP	0.894	3.335	7.202
7	Bayesian ridge	0.949	2.725	4.208	7	Bayesian ridge	0.947	3.980	6.490
	SCP	0.961	3.095	4.288		SCP	0.983	5.423	6.999
	Adaptive CP	0.944	2.748	3.790		Adaptive CP	0.954	4.007	6.403
	Time-weighted CP	0.938	2.614	3.871		Time-weighted CP	0.937	4.282	6.879
	ResCP	0.953	2.436	3.322		ResCP	0.931	3.565	6.585

Memory-feature definitions. For a decay parameter $a > 0$, define the causal exponential-memory state by $m_t^{(a)} = e^{-a} m_{t-1}^{(a)} + (1 - e^{-a})\, x_t, \qquad m_0^{(a)} = x_0$. The short-, medium-, and long-memory representations append one state with $a = 0.20$, 0.05, and 0.01, respectively, corresponding to exponential time constants $1/a = 5$, 20, and 100 observations. The multiscale representation appends all three states. The states are computed using only observations available through time t and are appended after the quadratic lag terms. Quadratic cross-products involving the memory states are not included. As with the other non-intercept features, the memory states are standardized using the fitting block only.

Table 14 reports point-prediction RMSE together with coverage, interval width, and Winkler score for Bayesian ridge and ordinary SCP. This analysis is intended as a robustness check rather than as a search for an optimal feature representation.

The qualitative Bayesian–SCP comparison is largely stable across these representations. Changes in feature construction can alter point error and interval metrics, but the differences among the uncertainty procedures are not explained solely by fine tuning of the NGRC representation. Similar point RMSE also does not imply identical interval behavior.

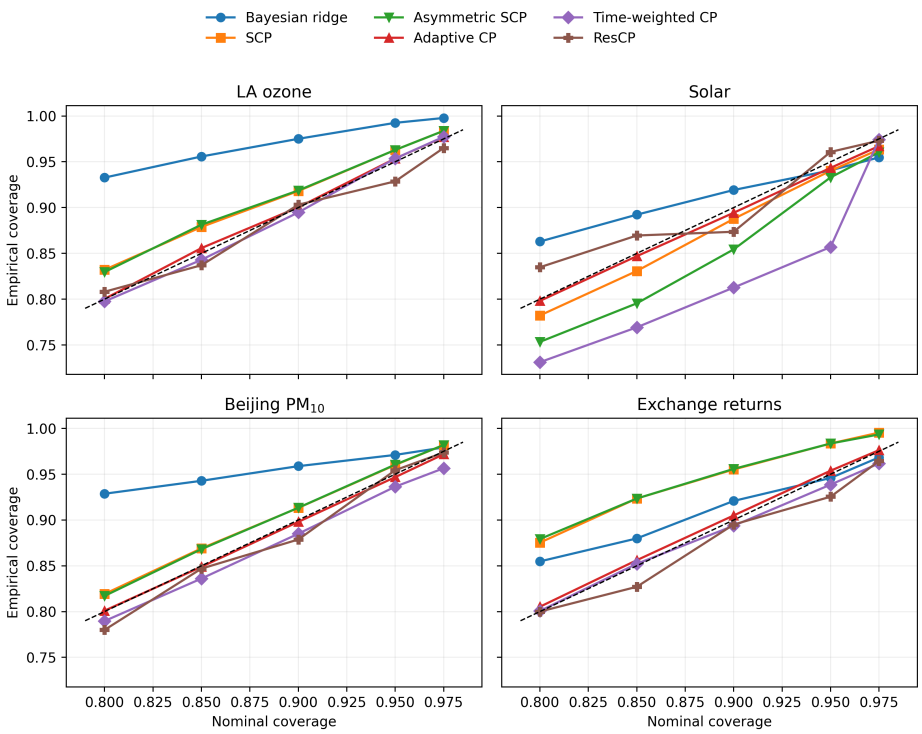


Figure 12: Fixed-origin empirical versus nominal coverage for all six NGRC UQ procedures. The dashed diagonal denotes exact marginal calibration.

Table 14: Feature-construction ablation at $H = 1$ and 95% nominal coverage. RMSE is reported once per feature construction because Bayesian ridge and SCP use the same fitted point predictor.

			Bayesian ridge			SCP		
Dataset	Features	RMSE	Coverage	Width	Winkler	Coverage	Width	Winkler
LA ozone	Linear lags	0.357	0.993	2.098	2.132	0.964	1.583	1.806
	Quadratic NGRC	0.360	0.993	2.047	2.083	0.963	1.573	1.795
	Short-memory NGRC	0.360	0.993	2.047	2.082	0.963	1.574	1.796
	Medium-memory NGRC	0.360	0.993	2.047	2.082	0.963	1.572	1.795
	Long-memory NGRC	0.355	0.992	2.040	2.079	0.963	1.555	1.783
	Multi-scale NGRC	0.354	0.992	2.036	2.076	0.962	1.558	1.783
Solar	Linear lags	0.033	0.936	0.120	0.227	0.933	0.116	0.228
	Quadratic NGRC	0.031	0.941	0.115	0.205	0.940	0.115	0.205
	Short-memory NGRC	0.031	0.941	0.114	0.205	0.940	0.114	0.205
	Medium-memory NGRC	0.031	0.941	0.115	0.205	0.940	0.114	0.205
	Long-memory NGRC	0.031	0.941	0.115	0.205	0.940	0.114	0.205
	Multi-scale NGRC	0.031	0.941	0.114	0.205	0.940	0.114	0.205
Beijing PM_{10}	Linear lags	0.179	0.973	0.931	1.185	0.958	0.751	1.126
	Quadratic NGRC	0.182	0.971	0.909	1.174	0.960	0.776	1.138
	Short-memory NGRC	0.182	0.971	0.909	1.174	0.960	0.775	1.138
	Medium-memory NGRC	0.182	0.971	0.909	1.174	0.960	0.773	1.137
	Long-memory NGRC	0.182	0.971	0.909	1.174	0.960	0.772	1.137
	Multi-scale NGRC	0.182	0.971	0.909	1.174	0.960	0.773	1.137
Exchange returns	Linear lags	1.276	0.948	3.912	6.401	0.983	5.405	6.983
	Quadratic NGRC	1.284	0.946	3.981	6.468	0.983	5.407	6.997
	Short-memory NGRC	1.284	0.946	3.981	6.466	0.983	5.404	6.994
	Medium-memory NGRC	1.284	0.946	3.981	6.467	0.983	5.407	6.997
	Long-memory NGRC	1.284	0.946	3.981	6.467	0.983	5.403	6.994
	Multi-scale NGRC	1.283	0.946	3.980	6.466	0.983	5.405	6.994

F.8 ResCP stochastic-seed robustness

ResCP uses random reservoir initialization and stochastic residual sampling. We therefore repeated its final test evaluation using five prespecified seeds: the seed used for the primary result and four

Table 15: Sensitivity of ResCP to its stochastic seed. For each dataset and horizon, the NGRC point forecasts, chronological split, and selected ResCP hyperparameters are fixed. Entries are means $\pm$ standard deviations over five seeds (the primary seed and four prespecified alternatives); the coverage range gives the minimum and maximum empirical coverage across those seeds. Standard deviations describe computational seed sensitivity, not sampling standard errors.

Dataset	Horizon	Coverage	Coverage range	Width	Winkler score
LA ozone	$H = 1$	0.933 ± 0.003	[0.929, 0.936]	1.375 ± 0.006	1.818 ± 0.017
	$H = 3$	0.954 ± 0.001	[0.953, 0.955]	1.755 ± 0.012	2.095 ± 0.006
	$H = 7$	0.952 ± 0.001	[0.950, 0.953]	1.772 ± 0.008	2.111 ± 0.012
Solar	$H = 1$	0.960 ± 0.003	[0.955, 0.963]	0.059 ± 0.001	0.078 ± 0.002
	$H = 3$	0.928 ± 0.010	[0.915, 0.937]	0.183 ± 0.012	0.258 ± 0.009
	$H = 7$	0.928 ± 0.009	[0.919, 0.936]	0.419 ± 0.008	0.576 ± 0.010
Beijing PM_{10}	$H = 1$	0.954 ± 0.004	[0.947, 0.958]	0.648 ± 0.024	0.933 ± 0.011
	$H = 3$	0.954 ± 0.005	[0.948, 0.960]	1.532 ± 0.062	2.146 ± 0.056
	$H = 7$	0.957 ± 0.003	[0.953, 0.961]	2.525 ± 0.081	3.339 ± 0.060
Exchange returns	$H = 1$	0.937 ± 0.003	[0.934, 0.941]	3.830 ± 0.070	6.654 ± 0.037
	$H = 3$	0.891 ± 0.005	[0.884, 0.898]	3.408 ± 0.083	7.233 ± 0.213
	$H = 7$	0.932 ± 0.004	[0.927, 0.937]	3.654 ± 0.054	6.587 ± 0.100

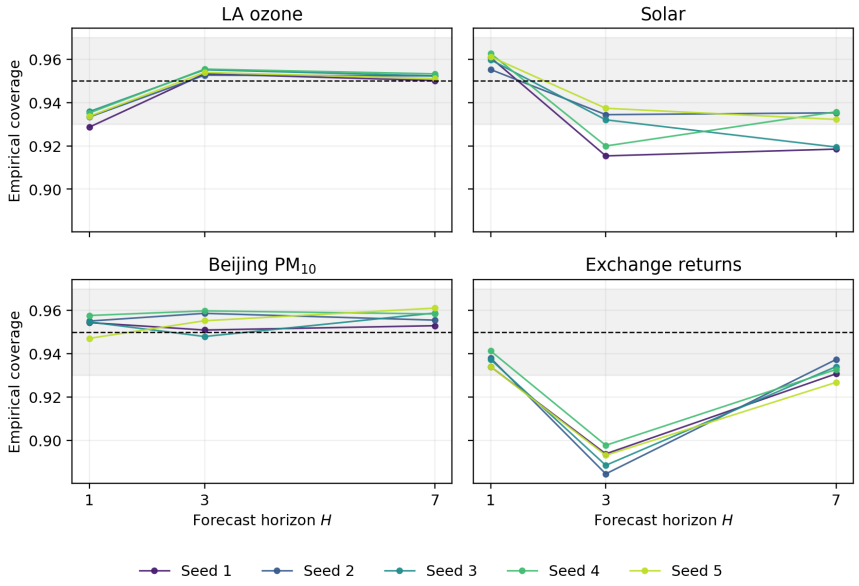


Figure 13: Empirical ResCP coverage across five stochastic seeds with all point forecasts, data splits, and selected ResCP hyperparameters held fixed. The dashed line marks the nominal 0.95 level, and the shaded band denotes coverage within 0.02 of nominal.

deterministic alternatives. The NGRC point forecasts, chronological fitting–calibration–test split, selected temperature, and selected history policy were held fixed in every run. Consequently, this check isolates sensitivity to ResCP's computational randomness rather than combining seed variation with hyperparameter reselection. The same availability-safe delayed-update protocol was used at every forecast horizon.

Across the 12 dataset–horizon configurations, the largest across-seed coverage standard deviation was 0.010 (Solar, $H = 3$). The largest relative standard deviations of mean interval width and Winkler score were 6.8% and 3.3%, respectively. Thus the qualitative comparisons are not driven by the particular stochastic seed used in the primary table.

Table 16: Comparison of point-forecasting architectures under the common official ResCP uncertainty procedure for one-step-ahead forecasting at 95% nominal coverage. Runtime includes preprocessing, point-model fitting, forecasting, ResCP tuning, and interval construction.

Dataset	Forecasting pipeline	RMSE	Coverage	PI width	Winkler	Total runtime (s)
LA ozone	ARIMA + ResCP	0.359	0.941	1.439	1.834	26.190
LA ozone	GRU + ResCP	0.365	0.956	1.524	1.810	60.960
LA ozone	Transformer + ResCP	0.364	0.935	1.391	1.841	141.942
LA ozone	NGRC + ResCP	0.360	0.929	1.369	1.797	19.390
Solar	ARIMA + ResCP	0.021	0.904	0.038	0.069	156.337
Solar	GRU + ResCP	0.020	0.894	0.041	0.070	442.007
Solar	Transformer + ResCP	0.029	0.907	0.041	0.067	1102.612
Solar	NGRC + ResCP	0.031	0.961	0.058	0.075	263.522
Beijing PM_{10}	ARIMA + ResCP	0.178	0.957	0.651	0.904	157.428
Beijing PM_{10}	GRU + ResCP	0.184	0.954	0.646	0.924	160.148
Beijing PM_{10}	Transformer + ResCP	0.204	0.956	0.671	0.919	351.799
Beijing PM_{10}	NGRC + ResCP	0.182	0.954	0.656	0.933	125.263
Exchange returns	ARIMA + ResCP	1.276	0.941	3.804	6.515	10.705
Exchange returns	GRU + ResCP	1.275	0.934	3.721	6.727	9.687
Exchange returns	Transformer + ResCP	1.276	0.921	3.625	6.684	24.113
Exchange returns	NGRC + ResCP	1.284	0.934	3.793	6.640	5.211

F.9 Architecture comparison + ResCP

We finally compare NGRC, ARIMA, a GRU, and a causal Transformer under a common ResCP uncertainty layer. Models use the same target transform, chronological blocks, and lag/history budget where their architectures permit it. The GRU and Transformer use the original ResCP benchmark's MAE training objective and the same chronological early-stopping budget. We therefore report both point error and interval metrics and treat RMSE as a descriptive comparison rather than an objective optimized equally by every model. End-to-end runtimes are reported with the interval metrics in Table 16. Because model classes and optimization problems cannot be made identical, this is an end-to-end pipeline comparison, not an isolated test of architecture.

Across these datasets, more computationally expensive forecasting architectures do not uniformly improve interval quality under the common ResCP uncertainty procedure. We do not interpret the results as showing that NGRC is generally superior to recurrent or Transformer models. Rather, point RMSE, residual shape, local calibration, and computational cost interact, and the relative performance of the complete pipelines depends on both the dataset and the evaluation metric.

Overall, residual shape and changes in residual scale explain much of the Bayesian–conformal ordering across these applications, adaptive recalibration is useful under temporal shift, and narrow localized intervals must be judged together with their coverage and interval score. These findings are consistent with the fixed-dimensional mechanisms developed above and illustrate their practical relevance in forecasting applications.

F.10 Runtime and Computational Details

We distinguish two runtime quantities throughout the real-data analysis. UQ-only runtime measures method-specific calibration, hyperparameter selection, sequential updating, and interval construction after the common NGRC point predictor has been fitted. End-to-end runtime additionally includes preprocessing, point-model fitting, and forecasting.

The runtimes reported for the primary UQ comparison are UQ-only measurements. The architecture comparison instead reports end-to-end pipeline runtime. These quantities should therefore not be compared directly.

For the multi-horizon experiments, wall-clock time is recorded separately at each forecast horizon. Bayesian ridge and ordinary SCP require little additional computation after fitting the common point

Table 17: Uncertainty-quantification runtime in seconds for forecast horizons $H = 3$ and $H = 7$. Times include pre-test uncertainty-quantification hyperparameter selection where applicable and exclude fitting of the common NGRC point predictor.

Dataset	Method	$H = 3$	$H = 7$
LA ozone	Bayesian ridge	0.0125	0.0107
	SCP	0.0003	0.0007
	Adaptive CP	6.4812	6.4685
	Time-weighted CP	9.5753	9.2434
	ResCP	22.2484	31.4085
Solar	Bayesian ridge	0.1008	0.0988
	SCP	0.0012	0.0012
	Adaptive CP	71.8993	73.2512
	Time-weighted CP	99.6658	100.6878
	ResCP	126.1974	122.0124
Beijing PM_{10}	Bayesian ridge	0.0267	0.0249
	SCP	0.0008	0.0007
	Adaptive CP	32.0766	31.9235
	Time-weighted CP	44.4136	45.1172
	ResCP	131.2923	131.9915
Exchange returns	Bayesian ridge	0.0041	0.0015
	SCP	0.0002	0.0001
	Adaptive CP	1.4273	1.4068
	Time-weighted CP	1.8807	1.8806
	ResCP	5.3240	5.4499

predictor, whereas adaptive CP, time-weighted CP, and ResCP require sequential processing and, where applicable, pre-test hyperparameter selection.