CONFORMAL PREDICTION FOR MULTIVARIATE TIME SERIES FORECASTING UNDER HETEROGENEOUS DISTRIBUTION SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction on multivariate time series forecasting provides intervals that cover true values with high probability, supporting reliable decision-making under uncertainty. Existing conformal methods typically target marginal coverage across all components and rely on the assumption that calibration and future forecast data have a joint distribution invariant under reordering. However, distribution shifts are common in time series, violating this invariance assumption, and can furthermore be heterogeneous across components in multivariate settings, leading to component-specific coverage failures. In this paper, we propose *Boundary-aware Conformal Recalibration* (BaCR), which combines two complementary mechanisms: *a unified recalibrator* that periodically adapts the quantile estimates used to construct prediction intervals, and *component-wise trackers* that provide immediate online interval correction within the finite range supported by these quantile estimates and report a signal when additional expansion is needed. Theoretically, we derive an exact decomposition of component-wise empirical coverage error that characterizes tracker correction and the boundary limitations from the conditional quantile estimates. Experiments on a synthetic benchmark and four real-world datasets show that BaCR substantially reduces component-wise undercoverage, especially on large-shift components, and produces more efficient intervals compared to baselines.

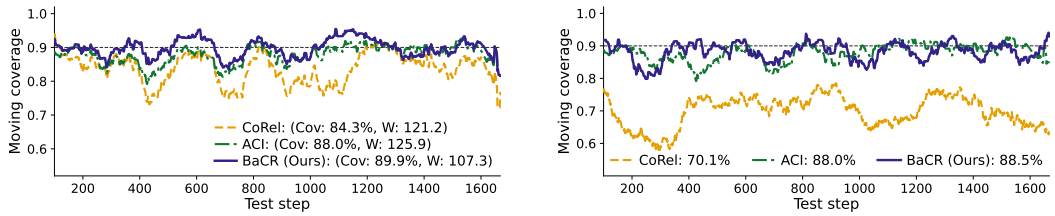
1 INTRODUCTION

Multivariate time series forecasting (MTSF) is a foundational task across science and engineering (Aburto & Weber, 2007; Liang et al., 2018; Lim & Zohren, 2021). Beyond point estimates, reliable uncertainty quantification is essential for deploying forecasting models in risk-sensitive settings, particularly when decisions depend on individual variables (Petropoulos et al., 2022).

Conformal prediction (CP), introduced by Vovk et al. (1999), is a nonparametric framework to quantify uncertainty, which provides a finite-sample marginal coverage guarantee under the exchangeability assumption. This framework offers a general methodology for converting the outputs of black-box models into prediction sets (Angelopoulos et al., 2023b).

However, exchangeability often fails to hold in sequential data and the validity of CP may be compromised because of distributional changes (Tibshirani et al., 2019). To address this challenge, conformal methods for non-exchangeable sequential data commonly model the conditional residual distribution or adapt calibration levels using online feedback. A representative method for MTSF is CoRel (Cini et al., 2025), which learns conditional residual quantiles from temporal and cross-component information. Despite its effectiveness, CoRel has an important practical *limitation*: repeatedly retraining the underlying models can be computationally prohibitive in high-dimensional multivariate settings. As a result, coverage may degrade when the estimated conditional quantiles fail to adapt sufficiently to changes in the test-time distribution. This issue is further exacerbated by the heterogeneous nature of distribution shifts across components. In particular, components subject to severe shifts may remain persistently undercovered, even when the aggregate coverage across all components is close to the nominal level. Such persistent component-wise undercoverage may conceal substantial risks, thereby undermining the reliability of downstream decision-making. Fig. 1

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.



(a) Largest-shift Component

(b) Worst-covered Component (per method)

Figure 1: Component-wise moving coverage on the AIR dataset (window of 100 test steps). Fig. 1(a) displays the results on the component with the largest shift (The shift is measured by the residuals, and how it is measured is given in the Appendix B.1). Legends report empirical coverage (Cov) and Winkler score (W, ↓). Fig. 1(b) shows the worst-covered component of each method with coverage reported in the legends (Winkler score is not reported because they are from different components). The black dashed line denotes the nominal level $1 - \alpha = 0.9$.

illustrates this limitation on the AIR dataset. CoRel experiences repeated coverage degradation on the component exhibiting the largest shift, while the worst-covered component reveals a substantially larger coverage gap. The reason is not the learned quantile estimates itself but the absence of test-time adaptation: once the distribution shifts, the quantile model lacks a mechanism to correct the resulting undercoverage.

Online conformal methods supply exactly this missing adaptation. Adaptive Conformal Inference (ACI) (Gibbs & Candes, 2021) adjusts, after each revealed observation, the effective level at which the empirical residual distribution is queried, and thereby controls the long-run empirical coverage rate. A component-wise extension to MTSF maintains an independent calibration state for each variable, so every component receives its own feedback. Under heterogeneous shifts, however, different components may require different learning rates, making a common learning rate difficult to select, while component-specific tuning introduces a substantial burden. Moreover, online methods may produce inflated intervals to achieve desired coverage, particularly when the component is under large shift. Notably, the largest-shift component in the illustrative example (Fig. 1(a)) is also ACI's worst-covered component (Fig. 1(b)), suggesting that direct online level adjustment remains challenged by heterogeneous shifts. Online methods are therefore not sufficient either: they adapt a scalar, symmetric radius that carries no conditional residual information, and there might be lags on the components whose shifts are larger than others.

Motivated by these limitations, we propose *Boundary-aware Conformal Recalibration* (**BaCR**), which combines a unified recalibrator with component-wise trackers without retraining the underlying models. The recalibrator updates the learned conditional quantile estimates periodically using recent residuals, while the trackers provide immediate online interval correction within the finite range supported by these estimates. When the trackers persistently use the available widening boundaries, BaCR feeds this observable saturation signal back to the recalibrator to request component-specific expansion. This signal directs additional expansion to components requiring greater tail reach without uniformly inflating all intervals.

Our contributions are summarized as follows:

- We propose **BaCR**, an online conformal method that combines a unified recalibrator with component-wise trackers. The recalibrator periodically adapts the conditional quantile estimates, while the trackers provide immediate online correction and report a signal when additional expansion of quantile estimates is needed.
- Theoretically, we derive an exact decomposition of component-wise empirical coverage error that characterizes tracker correction and the boundary limitations arising from the finite range of the quantile estimates.
- We conduct a comprehensive evaluation across a synthetic benchmark and four real-world datasets, which shows that BaCR achieve better coverage–efficiency tradeoff among baselines. Component ablations further demonstrate the complementary roles of the two mechanisms and the intended response to boundary limitations.

2 Related Work

Conformal Prediction. The task of uncertainty quantification for unobserved data has been widely studied (Abdar et al., 2021; Smith, 2024). Conformal inference is a significant branch of uncertainty quantification methods (Vovk et al., 2005). The power of conformal methods stems largely from its flexibility, simplicity and theoretical guarantee. Multiple variants of conformal methods and score functions have been developed to improve the adaptivity and efficiency (Vovk, 2015; Lei et al., 2018; Romano et al., 2019; Angelopoulos et al., 2020; Papadopoulos et al., 2002; Barber et al., 2021) and have been applied to many areas (Vovk et al., 2021; Stanton et al., 2023; Cresswell et al., 2024; Lekeufack et al., 2024). A detailed introduction of conformal inference is provided by Angelopoulos et al. (2023b).

Conformal Prediction under Non-exchangeability. Conformal methods highly rely on the exchangeability, but it is often violated when they are applied to practical problems (e.g., sequential data). Recent works of developing conformal methods beyond exchangeability can be summarized into three categories. The first strategy reweights past observations to reduce the influence of calibration samples that are less relevant to the current test distribution (Tibshirani et al., 2019; Barber et al., 2023). HopCPT (Auer et al., 2023) learns to retrieve and weight past residuals that are informative for the current prediction. The second strategy updates prediction intervals directly using revealed coverage errors. ACI (Gibbs & Candes, 2021) adjusts the effective level at which the empirical residual distribution is queried, while conformal PID (Angelopoulos et al., 2023a) updates an interval threshold through control theory. Related methods develop alternative online updates and learning-rate schedules (Zaffran et al., 2022; Bhatnagar et al., 2023; Gibbs & Candès, 2024; Angelopoulos et al., 2024). Since these approaches construct or update interval sizes from past residual scores, they have a relatively broad adjustment range. The third strategy estimates future residual quantiles from historical information. SPCI (Xu & Xie, 2023) predicts residual quantiles from past window. CoRel (Cini et al., 2025) extends this direction to MTSF by learning a shared relational quantile model from temporal and cross-component information. Nevertheless, under heterogeneous shifts, the learned quantile estimates may still provide insufficient adjustment for particular components, even when coverage aggregated across components remains close to the target.

3 Problem Setup

Online Multivariate Forecasting. Let $X_t = (X_{1,t}, \ldots, X_{N,t}) \in \mathbb{R}^N$ denote an observation of an N-component multivariate time series. Given forecast origin t and horizon $h \in \{1, \ldots, H\}$, the target is $Y_t^{(h)} := X_{t+h}$. Let $\mathcal{H}_t$ contain all information available when this forecast is issued, before $Y_t^{(h)}$ is revealed, including historical observations and previously observed residuals. A pretrained point forecaster produces $\widehat{Y}_t^{(h)} = \Theta(\mathcal{H}_t)$.

For notational clarity, all derivations below fix the forecast horizon as h and suppress the superscript (h); thus every component-wise statement is to be read for a fixed component–horizon pair (j, h). Feedback from a target is used only after that target has been observed. At every time t, we construct a collection of component-wise intervals $\{\widehat{C}_{j,t} \subset \mathbb{R}\}_{j=1}^N$. We do not seek simultaneous coverage of the full target vector; instead, each component is calibrated separately over time at the fixed horizon.

Residual Quantile Model. After Y_t is revealed, define the point-forecast residual $R_{j,t} = Y_{j,t} - \widehat{Y}_{j,t}$. Before observing Y_t, a pretrained quantile model Ψ produces an estimated conditional residual quantile $\widehat{Q}_{j,t}^{\tau} = [\Psi(\mathcal{H}_t, \tau)]_j$, $\tau \in (0, 1)$. Here, τ denotes the quantile index of the conditional residual distribution: smaller and larger values of τ correspond to the lower and upper residual tails, respectively. It is distinct from the target miscoverage level α. For a two-sided $(1-\alpha)$-interval, define the nominal endpoint indices as $\tau_{\mathrm{lo}} = \frac{\alpha}{2}, \tau_{\mathrm{hi}} = 1 - \frac{\alpha}{2}$.

We refer to the map $\tau \mapsto \widehat{Q}_{j,t}^{\tau}$, or equivalently to the arrangement of these estimates in residual value-space, as the *conditional residual quantile estimates*. The corresponding true conditional quantile is $Q_{j,t}^{\tau} = \inf\{z \in \mathbb{R} : \mathbb{P}(R_{j,t} \le z \mid \mathcal{H}_t) \ge \tau\}$. For a target miscoverage level $\alpha \in (0, 1)$, the pretrained residual quantile estimates give the base interval

$$\widehat{C}_{j,t}^{\mathrm{base}} = \left[\widehat{Y}_{j,t} + \widehat{Q}_{j,t}^{\tau_{\mathrm{lo}}},\ \widehat{Y}_{j,t} + \widehat{Q}_{j,t}^{\tau_{\mathrm{hi}}}\right] \tag{1}$$

These intervals provide component-specific uncertainty geometry, but they need not remain calibrated on a nonstationary test stream.

The quantile model is pretrained using the pinball loss (Koenker & Bassett Jr, 1978)

$$\ell_\tau(u) = u\left(\tau - \mathbf{1}\{u < 0\}\right), \qquad u = R_{j,t} - \widehat{Q}^{\tau}_{j,t}.$$

The point forecaster and quantile model remain frozen during online calibration.

Component-wise Online Calibration. For each component j, define the miscoverage indicator $e_{j,t} = \mathbf{1}\{Y_{j,t} \notin \widehat{C}_{j,t}\}$. Its empirical coverage over T steps is $\widehat{\mathrm{Cov}}_{j,T} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_{j,t} \in \widehat{C}_{j,t}\} = 1 - \frac{1}{T}\sum_{t=1}^{T} e_{j,t}$. Our online component-wise objective is $\left|\widehat{\mathrm{Cov}}_{j,T} - (1-\alpha)\right|$ should remain small for every component j.

4 METHODOLOGY

Overview. The proposed method BaCR consists of two modules: component-wise trackers and a unified recalibrator. Before online evaluation, the residual quantile model is trained on the calibration set and then frozen. At test time, the sequence is divided into fixed-length periods. Within each period, the trackers select component-specific endpoint levels from the available quantile range, construct prediction intervals, and update the levels after each revealed observation using coverage feedback. Tracking within learned conditional quantile estimates preserves component-specific and asymmetric residual information and can avoid the uniform interval inflation associated with directly updating a radius from past residual scores. The trade-off is that the trackers can query only the finite available quantile range. When the correction required by a component exceeds this range, an endpoint approaches an outer boundary and further updates are clipped. At the end of each period, BaCR summarizes this behavior through a saturation statistic. The unified recalibrator combines this signal with recent residuals to adjust the quantile estimates for the next period, providing targeted value-space expansion where additional correction is needed. The tracker states are then reset and replayed under the updated estimates.

Component-wise Trackers. At each time step, the tracker operates in two stages: it first uses its current states to select the endpoint levels and construct a prediction interval; after the observation is revealed, it updates these states using miscoverage indicator from each endpoint. Recall that the nominal endpoint levels are $\tau_{\mathrm{lo}} = \frac{\alpha}{2}$ and $\tau_{\mathrm{hi}} = 1 - \frac{\alpha}{2}$, and that quantile estimates are available over $[q_{\min}, q_{\max}]$, where $0 < q_{\min} < \tau_{\mathrm{lo}} < \frac{1}{2} < \tau_{\mathrm{hi}} < q_{\max} < 1$.

For each component j, the tracker maintains a lower- and an upper-endpoint adjustment state, $\delta^{\mathrm{lo}}_{j,t}$ and $\delta^{\mathrm{hi}}_{j,t}$. At the beginning of online evaluation, both states are initialized at zero, $\delta^{\mathrm{lo}}_{j,0} = \delta^{\mathrm{hi}}_{j,0} = 0$, so that the first prediction interval uses the nominal endpoint levels.

Before prediction at time t, the current states determine the effective endpoint levels $\tau^{\mathrm{lo}}_{j,t} = \tau_{\mathrm{lo}} + \delta^{\mathrm{lo}}_{j,t-1}, \tau^{\mathrm{hi}}_{j,t} = \tau_{\mathrm{hi}} + \delta^{\mathrm{hi}}_{j,t-1}$.

To keep these levels within the available quantile range and preserve the ordering of the interval endpoints, we require $\tau^{\mathrm{lo}}_{j,t} \in [q_{\min}, 1/2], \tau^{\mathrm{hi}}_{j,t} \in [1/2, q_{\max}]$.

Equivalently, the adjustment states are constrained to $I_{\mathrm{lo}} = [q_{\min} - \tau_{\mathrm{lo}},\ 1/2 - \tau_{\mathrm{lo}}], I_{\mathrm{hi}} = [1/2 - \tau_{\mathrm{hi}},\ q_{\max} - \tau_{\mathrm{hi}}]$.

Using the effective levels, the tracker constructs the component-wise prediction interval

$$\widehat{C}_{j,t} = \left[\widehat{Y}_{j,t} + \widetilde{Q}^{\tau^{\mathrm{lo}}_{j,t}}_{j,t},\ \widehat{Y}_{j,t} + \widetilde{Q}^{\tau^{\mathrm{hi}}_{j,t}}_{j,t}\right], \tag{2}$$

where $\widetilde{Q}^{\tau}_{j,t}$ denotes the recalibrated quantile estimates.

After $Y_{j,t}$ is revealed, the tracker determines whether the observed residual falls below the lower endpoint or above the upper endpoint and updates its states as

$$\begin{aligned}
\delta^{\mathrm{lo}}_{j,t} &:= \delta^{\mathrm{lo}}_{j,t-1} + \eta\left(\frac{\alpha}{2} - \mathbf{1}\left\{R_{j,t} < \widetilde{Q}^{\tau^{\mathrm{lo}}_{j,t}}_{j,t}\right\}\right), \\
\delta^{\mathrm{hi}}_{j,t} &:= \delta^{\mathrm{hi}}_{j,t-1} + \eta\left(\mathbf{1}\left\{R_{j,t} > \widetilde{Q}^{\tau^{\mathrm{hi}}_{j,t}}_{j,t}\right\} - \frac{\alpha}{2}\right).
\end{aligned} \tag{3}$$

Here, $\eta > 0$ is the tracker learning rate. After each update, the states are clipped when necessary so that $\delta_{j,t}^{\mathrm{lo}} \in I_{\mathrm{lo}}$ and $\delta_{j,t}^{\mathrm{hi}} \in I_{\mathrm{hi}}$.

A lower-side miscoverage decreases the lower endpoint level, while an upper-side miscoverage increases the upper endpoint level; both updates widen the prediction interval. Conversely, when no miscoverage occurs on a given side, the associated endpoint level moves inward: the lower level increases, whereas the upper level decreases. Since the available quantile estimates cover only the finite grid $[q_{\min}, q_{\max}]$, neither endpoint can move outward indefinitely.

When the conditional residual distribution changes in scale or shape, the current quantile estimates may no longer extend sufficiently far into one or both tails (Koenker & Xiao, 2006). Since the tracker operates within the finite grid $[q_{\min}, q_{\max}]$, repeated outward corrections may bring an endpoint close to its outer boundary. This behavior is directly observable from the tracker states.

We summarize the boundary use of component j during period $\mathcal{T}_k$ by the saturation signal

$$\mathrm{sat}_{j,k} = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \max \left\{ \frac{(-\delta_{j,t}^{\mathrm{lo}})_+}{\tau_{\mathrm{lo}} - q_{\min}},\ \frac{(\delta_{j,t}^{\mathrm{hi}})_+}{q_{\max} - \tau_{\mathrm{hi}}} \right\} \in [0, 1], \tag{4}$$

where $(x)_+ = \max\{x, 0\}$.

The two normalized terms within Eq. 4 measure the fractions of the available outward adjustment ranges used by the lower and upper endpoints, respectively. Taking their maximum captures the endpoint closer to its outer boundary, while averaging over the period distinguishes persistent boundary use from a temporary displacement. A value close to one indicates that at least one endpoint remained close to its outer boundary for much of the period, suggesting that the component may require additional value-space expansion. A value close to zero indicates little use of the available outward adjustment range. At the end of each period, this signal is passed to the unified recalibrator.

Algorithm 1 BaCR

Require: frozen models Θ and Ψ; quantile grid $\mathcal{Q}$; period length P; recalibration buffer size B; parameters η, λ, κ

For every component j, initialize $s_{j,1} \leftarrow 1$, $b_{j,1} \leftarrow 0$, and $\delta_{j,0}^{\mathrm{lo}}, \delta_{j,0}^{\mathrm{hi}} \leftarrow 0$
Initialize an empty rolling recalibration buffer $\mathcal{B}$
for period $k = 1, \ldots, K$ **do**
  for $t \in \mathcal{T}_k$ **do**
    Obtain $\widehat{Y}_{j,t}$ and $\{\widehat{Q}_{j,t}^{\tau} : \tau \in \mathcal{Q}\}$ from Θ and Ψ for every component j
    Apply $(s_{j,k}, b_{j,k})$ to obtain the adjusted quantile estimates using Eq. 5
    Select the effective endpoint levels and construct $\widehat{C}_{j,t}$ using Eq. 2
    After $Y_{j,t}$ is revealed, update the component-wise trackers using Eq. 3
    Append the revealed residuals and corresponding quantile estimates to $\mathcal{B}$, retaining its most recent B steps
  end for
  For every component j, compute $\mathrm{sat}_{j,k}$ from its tracker states using Eq. 4
  For every component j, update $(s_{j,k+1}, b_{j,k+1})$ by minimizing Eq. 6
  Reset the tracker states and replay the observations from $\mathcal{T}_k$ using the updated quantile estimates
end for

Unified Recalibrator. The unified recalibrator applies a component-specific affine transformation to all quantile estimates. During period $\mathcal{T}_k$, the adjusted estimate at level τ is

$$\widetilde{Q}_{j,t}^{\tau} = \widehat{Q}_{j,t}^{1/2} + s_{j,k} \left(\widehat{Q}_{j,t}^{\tau} - \widehat{Q}_{j,t}^{1/2} \right) + b_{j,k}, \quad s_{j,k} > 0, \tag{5}$$

where $s_{j,k}$ controls the scale around the estimated median and $b_{j,k}$ controls the location. The positive scale preserves the ordering of the pretrained quantile estimates. The recalibrator is initialized with $s_{j,1} = 1$ and $b_{j,1} = 0$, so the first period uses the pretrained estimates unchanged. Its parameters remain fixed within each period.

At the end of period $\mathcal{T}_k$, the recalibrator updates these parameters using recent residuals from a rolling recalibration buffer and the saturation signals produced by the trackers. Let $\overline{\mathcal{L}}^{\mathrm{pin}}_{j,k}(s_j, b_j)$ denote the average multi-quantile pinball loss over this buffer. The recalibration objective is

$$\mathcal{L}_{j,k}(s_j, b_j) = \overline{\mathcal{L}}^{\mathrm{pin}}_{j,k}(s_j, b_j) + \lambda \left(\log s_j - \kappa \, \mathrm{sat}_{j,k}\right)^2, \tag{6}$$

and the parameters for the next period are obtained by minimizing this objective over $s_j > 0$ and $b_j \in \mathbb{R}$. The parameter $\lambda > 0$ controls the strength of this penalty, while $\kappa > 0$ determines how strongly the observed saturation is passed to scale expansion.

The pinball term adapts the location and scale of the quantile estimates to the recent residual distribution. The second term is a boundary-aware penalty that encourages the log-scale to follow the saturation-dependent target $\kappa \, \mathrm{sat}_{j,k}$. Since the affine transformation is centered at the estimated median, increasing s_j moves lower quantile estimates downward and upper quantile estimates upward, thereby expanding the estimates in value-space. A larger saturation value therefore encourages additional expansion for a component whose tracker persistently approaches an outer boundary. When $\mathrm{sat}_{j,k} = 0$, the scale target is $s_j = 1$, while the pinball term can still adjust both s_j and b_j according to the recent residuals.

After the recalibrator is updated, the tracker states are reset and replayed over the observations from the completed period using the new quantile estimates. The resulting states initialize the trackers for period $\mathcal{T}_{k+1}$. The complete procedure is summarized in Algorithm 1.

5 Theoretical Analysis

We analyze the component-wise coverage of BaCR. Since the endpoint levels are restricted to the finite conditional quantile levels $[q_{\min}, q_{\max}]$, some tracker updates may be clipped at the boundaries. Theorem 1 decomposes the resulting coverage gap into the tracker state changes and the clipping at the outer and inner boundaries. This decomposition holds for every realized sequence and requires neither exchangeability nor accurate quantile estimates. Formal definitions and proofs are provided in Appendix A.

Theorem 1 (Component-wise coverage decomposition)**.** *For every component j and every realized sequence of $T = KP$ online steps,*

$$\widehat{\mathrm{Cov}}_{j,T} - (1-\alpha) = -\mathcal{G}_{j,T} - \left(1 - \frac{\alpha}{2}\right) \bar{\pi}^{\mathrm{wid}}_{j,T} + \frac{\alpha}{2} \bar{\pi}^{\mathrm{nar}}_{j,T}. \tag{7}$$

Consequently,

$$-|\mathcal{G}_{j,T}| - \left(1 - \frac{\alpha}{2}\right) \bar{\pi}^{\mathrm{wid}}_{j,T} \le \widehat{\mathrm{Cov}}_{j,T} - (1-\alpha) \le |\mathcal{G}_{j,T}| + \frac{\alpha}{2} \bar{\pi}^{\mathrm{nar}}_{j,T}. \tag{8}$$

Here, $\mathcal{G}_{j,T}$ is the sum of the within-period upper-state changes minus the lower-state changes, divided by ηT, while $\bar{\pi}^{\mathrm{wid}}_{j,T}$ and $\bar{\pi}^{\mathrm{nar}}_{j,T}$ are the average normalized update amounts clipped at the outer grid boundaries and the median boundaries, respectively. Their complete definitions are given in Appendix A.

The proof follows the telescoping argument underlying the long-run coverage analysis of ACI (Gibbs & Candes, 2021). Theorem 1 extends this argument to separate lower- and upper-endpoint trackers. Restricting the trackers to a finite quantile grid produces the two explicit boundary clipping terms in Eq. 7. The first term in Eq. 7 records the difference between the accumulated changes of the tracker states. The second term represents outward updates clipped at $q_{\min}$ or $q_{\max}$ and therefore contributes directly to undercoverage. The third represents inward updates clipped at the median boundaries and contributes in the opposite direction. Accordingly, the bounds in Eq. 8 are asymmetric: widening clipping appears only in the lower bound with coefficient $1 - \alpha/2$, whereas narrowing clipping appears only in the upper bound with coefficient $\alpha/2$. Since widening clipping is the boundary term that lowers coverage, BaCR defines saturation using outward boundary use. Let $\overline{\mathrm{sat}}_{j,T} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{sat}_{j,k}$. Then $\bar{\pi}^{\mathrm{wid}}_{j,T} \le \overline{\mathrm{sat}}_{j,T}$. Thus, the observed period-averaged saturation provides an upper bound on the widening-clipping term. Narrowing clipping is retained in the exact decomposition but is not used as a recalibration signal. The corresponding definitions and proof are given in Appendix A.

A complementary telescoping argument bounds $|\mathcal{G}_{j,T}|$ by a bounded initial-to-final endpoint term plus the cumulative magnitude of the period-boundary state changes, divided by ηT. Consequently, continuous state carry-over yields an $O((\eta T)^{-1})$ term. Under reset and replay, the term still vanishes when the cumulative period-boundary variation is $o(\eta T)$. The proof is provided in Appendix A.

Remark 1 (Vanishing and boundary-limited cases)**.** *Suppose the conditional residual distribution is continuous and the adjusted quantile estimates equal the true conditional residual quantiles over the feasible levels. The nominal tracker state is then a common interior restoring point, and the expected clipping terms are of order $O((\eta P)^{-1} + \eta)$. Therefore, along a sequence of designs indexed by P, with $\eta = \eta_P$, $\eta_P \to 0$, and $\eta_P P \to \infty$, all three terms in Eq. 7 vanish and $\widehat{\mathrm{Cov}}_{j,T} \to 1 - \alpha$ in L^1. Exact quantile accuracy is sufficient but not necessary. The same conclusion extends to misspecified quantiles when, for each endpoint within each period, the conditional expected updates are zero at a common point separated from the feasible boundaries by a uniform positive margin and consistently direct the tracker state toward that point, and when the approximation and replay errors across periods vanish at the required rates. Requiring the same point at every step is stronger than requiring it only for the period-average update. If no interior restoring point exists, widening clipping may persist and quantile-level adjustment alone cannot attain the target tail frequency. BaCR therefore uses the saturation signal to encourage additional value-space expansion. The formal conditions, finite-time bounds, and limitations are given in Appendix A.*

6 Experiments

6.1 Setup

Underlying Algorithms. We use two point forecasters Θ: a recurrent network (RNN) (Cho et al., 2014) and a Transformer (Vaswani et al., 2017). On top of each, we train a quantile neural network Ψ (Cini et al., 2023) that outputs quantile estimates on a fixed grid $\tau \in \{0.01, \ldots, 0.99\}$ (step 0.01). Each series is split chronologically into training, calibration, and test sets of 48%, 32%, and 20%, respectively. Θ is trained on the training set and Ψ on Θ's residuals over the calibration set. Both Θ and Ψ are fit once and then frozen. The evaluation is performed on the test set.

Baselines. We compare against seven conformal baselines spanning methods from the Related Work. **Traditional**: Split Conformal Prediction (SCP) (Papadopoulos et al., 2002), a static calibration-quantile baseline under exchangeability. **Weighted**: NexCP (Barber et al., 2023), which reweights calibration scores toward recent observations, and HopCPT (Auer et al., 2023), which retrieves and weights relevant past residuals with a modern Hopfield network. **Quantile regression**: SPCI (Xu & Xie, 2023) and CoRel (Cini et al., 2025), which fit a quantile random forest and neural network on past residuals. **Online adaptive methods**: ACI (Gibbs & Candes, 2021) and conformal PID (Angelopoulos et al., 2023a), which adapt a scalar quantile level online from coverage feedback.

Implementation and Settings. Baseline configurations follow the recommended settings in their original papers and available public implementations. Repository sources and any implementation-specific adaptations are reported in Appendix B.1. BaCR uses a common configuration across all datasets and backbones. In the multi-step experiments, the realized targets of all H horizons of a forecast are fed back to every online method before the next forecast origin; this protocol is identical across methods (Appendix B.1). A complete sensitivity analysis are provided in Appendix B.4.

Evaluation Metrics. We use two metrics: the *coverage deviation* and the *Winkler score* (Winkler, 1972). At the target coverage $1 - \alpha = 0.9$, the *coverage deviation* is computed as $\Delta\mathrm{Cov} = 100(\overline{\mathrm{Cov}} - 0.9)$ percentage points (pp). Coverage is considered empirically acceptable if $|\Delta\mathrm{Cov}| \leq 1$ pp (this tolerance is an evaluation convention rather than a formal definition). For the main comparison, we report metrics over all components and over the S^+ slice (the top 10% components with the largest shifts). For component ablations, we additionally stratify by the *local tail pressure* χ of a component within a window (defined in Appendix B.3.1), which compares the outward reach that window requires for nominal coverage against the reference radius given by the calibration residual tails.

6.2 Synthetic Datasets

Description. We construct a multivariate time series by introducing fully controlled heterogeneous distribution shifts. For $N = 100$ components over $T = 12000$ steps, each component follows a het-

erogeneous AR(1) process $y_{i,t} = a_i\, y_{i,t-1} + \sigma_{i,t}\, \varepsilon_{i,t}, a_i \sim \mathcal{U}(0.3, 0.8)$. The innovation scale composes three independent per-component stresses on a base level $\sigma_0 = 0.4$, $\sigma_{i,t} = \sigma_0\, A_i\, B_i(t)\, C_{i,t}$, clipped to $[0.05, 10]$, where: (1) cross-component heteroscedasticity $A_i = e^{Z_i}$, $Z_i \sim \mathcal{N}(0, 1.3^2)$, so some components are an order of magnitude noisier than others; (2) per-component cyclic volatility $B_i(t) = 1 + 0.95 \sin(2\pi t/200 + \phi_i)$, $\phi_i \sim \mathcal{U}(0, 2\pi)$, each component peaking at a different phase; and (3) per-component volatility regimes $C_{i,t}$, in which each component's time axis is split into 8 contiguous segments at random change-points, each segment receiving an independent level e^{ζ}, $\zeta \sim \mathcal{N}(0, 0.96^2)$, yielding a non-stationary piecewise-constant scale. The innovations $\varepsilon_{i,t}$ are heavy-tailed, drawn from a Student-t with per-component degrees of freedom $\nu_i \in \{3, 4, 5\}$.

Results. Tab. 1 reports the results on the controlled synthetic dataset. Several baselines (NexCP, HopCPT, CoRel) are coverage-acceptable ($|\Delta\text{Cov}| \leq 1$) across all components, and BaCR stays close to the nominal coverage level. This confirms that marginal coverage alone does not reveal the hard-component failure targeted in this work. The difference becomes clear on S^+ slice. Static and conditional-quantile baselines lose coverage sharply on this slice. SCP and SPCI fail by a large margin, and CoRel is slightly undercovered, despite strong overall performance. In contrast, only NexCP, ACI, and BaCR remain within the coverage-acceptable band on both the full population and S^+ slice. The key difference among these methods is efficiency. NexCP and ACI maintain coverage by producing wider intervals, whereas BaCR remains substantially sharper. On S^+ slice, BaCR achieves a lower Winkler score than ACI and NexCP, and stays close to the oracle reference. Thus, among the methods lying within the coverage-acceptable band on both views, BaCR attains the lowest Winkler score.

Table 1: Synthetic results over 10 seeds (RNN, mean$_{\pm\text{std}}$). **Bold**: lowest value among methods that maintain acceptable coverage ($|\Delta\text{Cov}| \leq 1$). *Oracle* is not a competing method but the interval from the known generative scale.

Method	ΔCov(%)	Winkler ↓	ΔCov_{S^+}(%)	Winkler_{S^+} ↓
Oracle	$+0.0_{\pm 0.1}$	$3.91_{\pm 0.52}$	$+0.0_{\pm 0.2}$	$3.46_{\pm 0.84}$
SCP	$-4.9_{\pm 1.4}$	$6.21_{\pm 0.74}$	$-21.4_{\pm 4.4}$	$6.91_{\pm 1.62}$
NexCP	$+0.5_{\pm 0.0}$	$4.83_{\pm 0.61}$	$+0.3_{\pm 0.1}$	$4.41_{\pm 1.05}$
HopCPT	$-0.5_{\pm 1.1}$	$5.21_{\pm 0.62}$	$-5.7_{\pm 2.4}$	$5.26_{\pm 1.37}$
SPCI	$-4.6_{\pm 1.2}$	$6.02_{\pm 0.72}$	$-17.4_{\pm 3.3}$	$6.29_{\pm 1.44}$
CoRel	$-0.6_{\pm 0.8}$	$4.35_{\pm 0.55}$	$-2.1_{\pm 0.8}$	$3.81_{\pm 0.88}$
ACI	$-0.1_{\pm 0.0}$	$4.69_{\pm 0.61}$	$-0.1_{\pm 0.1}$	$4.23_{\pm 1.02}$
PID	$-2.4_{\pm 0.1}$	$4.46_{\pm 0.58}$	$-2.9_{\pm 0.3}$	$4.07_{\pm 0.96}$
BaCR	$-0.4_{\pm 0.2}$	$\mathbf{4.18_{\pm 0.54}}$	$-0.9_{\pm 0.3}$	$\mathbf{3.75_{\pm 0.89}}$

6.3 Real Datasets

Description. Four real MTSF datasets are used. **AIR** (Zheng et al., 2015) is hourly air-quality monitoring from 437 stations across different Chinese cities ($H = 3$). **LA** and **BAY** (Li et al., 2017) are traffic-speed readings in Los Angeles and the Bay Area ($H = 12$), respectively; **SD** (Liu et al., 2023) is the San Diego subset of LargeST ($H = 12$). We follow the preprocessing steps of works (Cini & Marisca, 2022).

Results. Tab. 2 first reports the primary all-component results. BaCR remains within the prescribed coverage band on all four datasets and attains the lowest Winkler score among coverage-acceptable methods in every case. ACI is coverage-acceptable on all four datasets, but has a higher Winkler score than BaCR on all datasets. The same pattern holds for the RNN backbone, as reported in the complete results Tab. 3 in Appendix B.2. The S^+ slice reveals failures on components experiencing the largest residual tail-scale growth. CoRel falls outside the coverage band on AIR, BAY, and SD despite competitive Winkler scores. ACI remains coverage-acceptable on all four datasets, while BaCR obtains a lower Winkler score in every comparison. On AIR, LA, and SD, BaCR is both coverage-acceptable and has the lowest Winkler score. On the S^+ slice, BAY is the only case where BaCR falls just outside the band (-1.2%, lowest raw Winkler 13.5; ACI -0.3% at 14.0).

6.4 Ablation Study

The S^+ slice asks which components are most affected by shift; stratifying by χ asks how each mechanism's contribution changes with the local pressure it meets, which a component-level average over the whole test segment cannot show. Fig. 2 reports the ablation across the five strata. In the highest stratum the complete model shows less undercoverage than either single-component variant on all four datasets and both backbones. Adding the tracker improves coverage and the Winkler score together in all eight highest-stratum comparisons. Adding the recalibrator instead widens the intervals, by 10% to 29% in the highest stratum, and that cost is largely offset by the fall in the miss

Table 2: Performance of models (on Transformer) on real datasets, mean$_{\pm \text{std}}$ over 10 seeds.

		SCP	NexCP	HopCPT	SPCI	CoRel	ACI	PID	**BaCR**
AIR	ΔCov(%)	$+3.2_{\pm 0.1}$	$+1.0_{\pm 0.0}$	$+2.4_{\pm 0.3}$	$+1.8_{\pm 0.1}$	$-1.9_{\pm 0.4}$	$-0.5_{\pm 0.0}$	$-0.6_{\pm 0.0}$	$+0.6_{\pm 0.1}$
	Winkler↓	$111.8_{\pm 1.0}$	$104.0_{\pm 0.8}$	$112.5_{\pm 4.9}$	$102.8_{\pm 0.9}$	$88.8_{\pm 0.7}$	$99.0_{\pm 0.6}$	$98.7_{\pm 0.6}$	$\mathbf{89.5_{\pm 0.8}}$
	ΔCov_{S^+}(%)	$-1.4_{\pm 0.2}$	$+0.9_{\pm 0.0}$	$-0.0_{\pm 0.4}$	$-1.7_{\pm 0.2}$	$-5.4_{\pm 0.6}$	$-0.8_{\pm 0.0}$	$-0.9_{\pm 0.0}$	$+0.4_{\pm 0.2}$
	Winkler_{S^+} ↓	$110.4_{\pm 2.2}$	$104.1_{\pm 2.3}$	$152.2_{\pm 29.2}$	$104.1_{\pm 1.9}$	$90.3_{\pm 2.3}$	$97.1_{\pm 2.2}$	$97.1_{\pm 2.3}$	$\mathbf{86.6_{\pm 2.1}}$
LA	ΔCov(%)	$-2.3_{\pm 0.1}$	$+0.7_{\pm 0.0}$	$-0.1_{\pm 0.6}$	$-2.3_{\pm 0.0}$	$-0.6_{\pm 0.1}$	$-0.3_{\pm 0.0}$	$-0.5_{\pm 0.1}$	$-0.3_{\pm 0.1}$
	Winkler↓	$32.3_{\pm 0.4}$	$29.5_{\pm 0.3}$	$27.1_{\pm 1.1}$	$24.1_{\pm 0.3}$	$21.9_{\pm 0.2}$	$23.5_{\pm 0.2}$	$31.3_{\pm 0.3}$	$\mathbf{20.2_{\pm 0.3}}$
	ΔCov_{S^+}(%)	$-6.0_{\pm 0.4}$	$+0.6_{\pm 0.1}$	$-1.4_{\pm 0.7}$	$-2.5_{\pm 0.1}$	$-1.0_{\pm 0.2}$	$-0.4_{\pm 0.0}$	$-1.4_{\pm 0.3}$	$-0.3_{\pm 0.1}$
	Winkler_{S^+} ↓	$41.7_{\pm 1.2}$	$36.2_{\pm 1.0}$	$33.6_{\pm 1.8}$	$27.2_{\pm 0.6}$	$25.5_{\pm 0.7}$	$27.0_{\pm 0.7}$	$44.5_{\pm 2.8}$	$\mathbf{22.9_{\pm 0.7}}$
BAY	ΔCov(%)	$-0.6_{\pm 0.1}$	$+0.7_{\pm 0.0}$	$+0.4_{\pm 0.3}$	$+2.0_{\pm 0.1}$	$-0.3_{\pm 0.1}$	$-0.4_{\pm 0.0}$	$-2.1_{\pm 0.1}$	$-0.5_{\pm 0.1}$
	Winkler↓	$17.1_{\pm 0.1}$	$14.9_{\pm 0.1}$	$11.7_{\pm 0.8}$	$10.9_{\pm 0.1}$	$10.8_{\pm 0.1}$	$10.2_{\pm 0.1}$	$30.5_{\pm 1.1}$	$\mathbf{9.6_{\pm 0.1}}$
	ΔCov_{S^+}(%)	$-2.9_{\pm 0.3}$	$+0.7_{\pm 0.1}$	$-0.6_{\pm 0.4}$	$+0.5_{\pm 0.3}$	$-4.3_{\pm 0.2}$	$-0.3_{\pm 0.0}$	$-1.5_{\pm 0.3}$	$-1.2_{\pm 0.1}$
	Winkler_{S^+} ↓	$24.6_{\pm 0.2}$	$20.5_{\pm 0.1}$	$16.8_{\pm 1.4}$	$15.3_{\pm 0.2}$	$15.6_{\pm 0.2}$	$\mathbf{14.0_{\pm 0.2}}$	$34.0_{\pm 3.3}$	$13.5_{\pm 0.2}$
SD	ΔCov(%)	$-2.4_{\pm 0.3}$	$+0.8_{\pm 0.0}$	$-0.7_{\pm 0.2}$	$-2.1_{\pm 0.2}$	$-1.7_{\pm 0.5}$	$-0.4_{\pm 0.0}$	$-0.2_{\pm 0.0}$	$-0.6_{\pm 0.1}$
	Winkler↓	$174.9_{\pm 2.8}$	$164.1_{\pm 2.8}$	$139.5_{\pm 2.6}$	$146.4_{\pm 2.5}$	$143.2_{\pm 2.5}$	$149.0_{\pm 2.8}$	$151.7_{\pm 2.5}$	$\mathbf{127.5_{\pm 2.6}}$
	ΔCov_{S^+}(%)	$-4.7_{\pm 0.5}$	$+0.9_{\pm 0.1}$	$-1.0_{\pm 0.2}$	$-3.3_{\pm 0.2}$	$-4.0_{\pm 0.5}$	$-0.4_{\pm 0.0}$	$-0.2_{\pm 0.0}$	$-0.7_{\pm 0.1}$
	Winkler_{S^+} ↓	$220.6_{\pm 5.8}$	$194.6_{\pm 5.9}$	$162.6_{\pm 4.5}$	$177.2_{\pm 4.9}$	$174.2_{\pm 4.4}$	$171.6_{\pm 5.5}$	$175.5_{\pm 5.0}$	$\mathbf{148.0_{\pm 4.1}}$

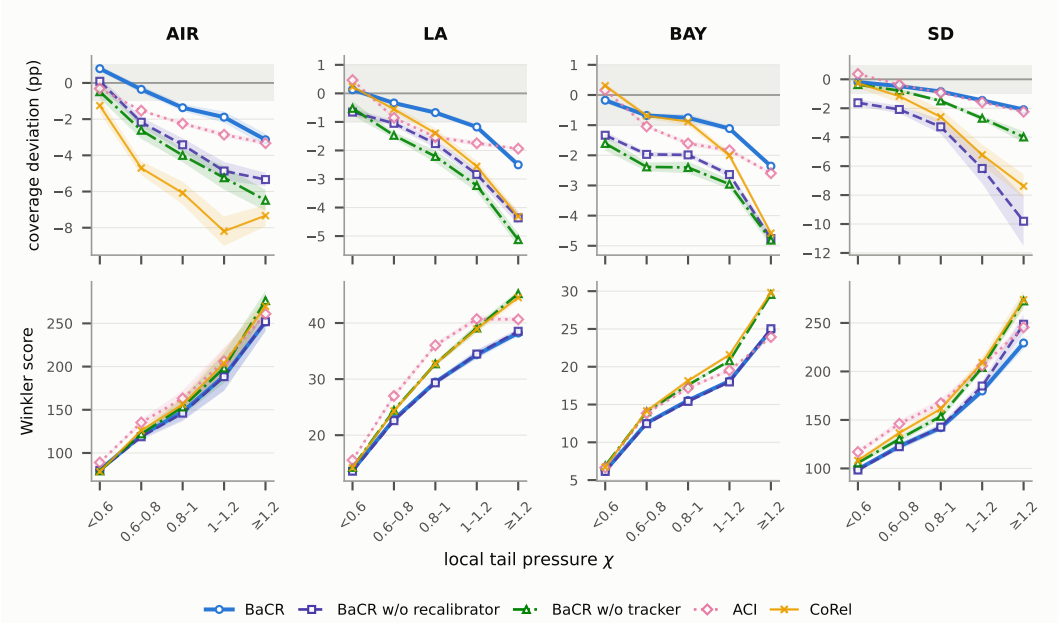


Figure 2: Component ablation stratified by local tail pressure χ (Transformer, 10 seeds). Top: coverage deviation, shaded band $|\Delta\text{Cov}| \leq 1$. Bottom: Winkler score.

penalty. Removing the recalibrator costs 1.9 to 7.7 pp of coverage there and removing the tracker costs 1.9 to 3.4; the recalibrator matters by far the most on SD, where the tracker alone falls to -9.8 pp while the complete model holds -2.1 pp. The coverage advantage of the complete model is larger in the highest stratum than in the lowest in all comparisons. Per-stratum tables for both backbones and the window-length sensitivity analysis are in Appendix B.3.1; the corresponding ablation on the S^+ slice is in Fig. 3 in Appendix B.3.2. We further examine whether saturation signals reflect the boundary limitation they are designed to identify in Appendix B.3.3.

7 Conclusion

This work introduced **BaCR**, an online method that couples component-wise trackers with a unified recalibrator while keeping the pretrained models frozen. An observable saturation statistic converts a tracker's persistent widening against its adjustment limit into targeted expansion of component-specific quantile estimates. Experiments on four real datasets show that BaCR achieves a coverage–efficiency trade-off while substantially reducing undercoverage on large-shift components.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing, including improving grammar, clarity, readability, and phrasing. We did not use generative AI tools to generate research ideas, formulate hypotheses, design the methodology, conduct experiments, analyze data, produce results, or draw scientific conclusions. All AI-assisted edits were reviewed by the authors to ensure that they accurately reflected the intended meaning and did not alter the technical content or scientific claims. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The anonymous github link is: https://anonymous.4open.science/r/BaCR.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Luis Aburto and Richard Weber. Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1):136–144, 2007.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023b.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*, 2024.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in neural information processing systems*, 36: 56027–56074, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation*, pp. 103–111, 2014.

Andrea Cini and Ivan Marisca. Torch Spatiotemporal, 3 2022. URL `https://github.com/TorchSpatiotemporal/tsl`.

Andrea Cini, Ivan Marisca, Daniele Zambon, and Cesare Alippi. Taming local effects in graph-based spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 36:55375–55393, 2023.

Andrea Cini, Alexander Jenkins, Danilo Mandic, Cesare Alippi, and Filippo Maria Bianchi. Relational conformal prediction for correlated time series. *arXiv preprint arXiv:2502.09443*, 2025.

Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

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

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Roger Koenker and Zhijie Xiao. Quantile autoregression. *Journal of the American statistical association*, 101(475):980–990, 2006.

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.

Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11668–11675. IEEE, 2024.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *Ijcai*, volume 2018, pp. 3428–3434, 2018.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems*, 36:75354–75371, 2023.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J Bessa, Jakub Bijak, John E Boylan, et al. Forecasting: theory and practice. *International Journal of forecasting*, 38(3):705–871, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Ralph C Smith. *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2024.

Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 959–986. PMLR, 2023.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74 (1):9–28, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pp. 191–210. PMLR, 2021.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727. PMLR, 2023.

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.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2267–2276, 2015.

A THEORETICAL ANALYSIS

Fix a component j and suppress the fixed forecast horizon. Let $0 < \alpha \le 1/2$ and

$$\tau_{\mathrm{lo}} = \frac{\alpha}{2}, \qquad \tau_{\mathrm{hi}} = 1 - \frac{\alpha}{2}.$$

We consider $T = KP$ online steps divided into periods $\mathcal{T}_k = \{t_k + 1, \dots, t_k + P\}$, where $t_k = (k-1)P$, and write $R_{j,t} = Y_{j,t} - \widehat{Y}_{j,t}$.

Assume

$$0 < q_{\min} < \tau_{\mathrm{lo}} < \frac{1}{2} < \tau_{\mathrm{hi}} < q_{\max} < 1,$$

that $p \mapsto \widehat{Q}^{p}_{j,t}$ is nondecreasing, and that $s_{j,k} > 0$. The tracker states satisfy

$$\tau_{\mathrm{lo}} + \delta^{\mathrm{lo}}_{j,t} \in [q_{\min}, 1/2], \qquad \tau_{\mathrm{hi}} + \delta^{\mathrm{hi}}_{j,t} \in [1/2, q_{\max}].$$

Hence the adjusted quantile estimates remain nondecreasing and the two interval endpoints remain ordered. The strict lower- and upper-side miscoverage events used below are therefore disjoint and exactly describe miscoverage of the closed prediction interval. All coverage statements are component-wise and averaged over time; they do not imply conditional coverage or simultaneous coverage of the multivariate vector.

Endpoint-wise Tracker Identities. We first derive exact identities for the lower- and upper-endpoint tracker updates within a period. During $\mathcal{T}_k$, the recalibrator parameters $(s_{j,k}, b_{j,k})$ remain fixed, while the two tracker states are updated after each observation.

Define the lower- and upper-side miscoverage indicators as

$$e^{\mathrm{lo}}_{j,t} = \mathbf{1}\left\{ R_{j,t} < \widetilde{Q}^{\tau_{\mathrm{lo}} + \delta^{\mathrm{lo}}_{j,t-1}}_{j,t} \right\}, e^{\mathrm{hi}}_{j,t} = \mathbf{1}\left\{ R_{j,t} > \widetilde{Q}^{\tau_{\mathrm{hi}} + \delta^{\mathrm{hi}}_{j,t-1}}_{j,t} \right\}.$$

The clipped tracker updates are

$$\begin{aligned} \delta^{\mathrm{lo}}_{j,t} &= \Pi_{I_{\mathrm{lo}}}\left[\delta^{\mathrm{lo}}_{j,t-1} + \eta\left(\frac{\alpha}{2} - e^{\mathrm{lo}}_{j,t} \right) \right], \\ \delta^{\mathrm{hi}}_{j,t} &= \Pi_{I_{\mathrm{hi}}}\left[\delta^{\mathrm{hi}}_{j,t-1} + \eta\left(e^{\mathrm{hi}}_{j,t} - \frac{\alpha}{2} \right) \right], \end{aligned} \tag{9}$$

where

$$I_{\mathrm{lo}} = \left[q_{\min} - \tau_{\mathrm{lo}}, \frac{1}{2} - \tau_{\mathrm{lo}} \right], I_{\mathrm{hi}} = \left[\frac{1}{2} - \tau_{\mathrm{hi}}, q_{\max} - \tau_{\mathrm{hi}} \right],$$

and Π_I denotes clipping to I.

For each endpoint, define the amount removed by clipping as

$$\begin{aligned} \upsilon^{\mathrm{lo}}_{j,t} &= \delta^{\mathrm{lo}}_{j,t-1} + \eta\left(\frac{\alpha}{2} - e^{\mathrm{lo}}_{j,t} \right) - \delta^{\mathrm{lo}}_{j,t}, \\ \upsilon^{\mathrm{hi}}_{j,t} &= \delta^{\mathrm{hi}}_{j,t-1} + \eta\left(e^{\mathrm{hi}}_{j,t} - \frac{\alpha}{2} \right) - \delta^{\mathrm{hi}}_{j,t}. \end{aligned} \tag{10}$$

For $x \in \mathbb{R}$, write $(x)_+ = \max\{x, 0\}$ and $(x)_- = \max\{-x, 0\}$. The normalized endpoint-specific clipping masses are

$$\begin{aligned} \pi^{\mathrm{lo},+}_{j,k} &= \frac{\sum_{t \in \mathcal{T}_k} (\upsilon^{\mathrm{lo}}_{j,t})_+}{\eta P(\alpha/2)}, & \pi^{\mathrm{lo},-}_{j,k} &= \frac{\sum_{t \in \mathcal{T}_k} (\upsilon^{\mathrm{lo}}_{j,t})_-}{\eta P(1-\alpha/2)}, \\ \pi^{\mathrm{hi},+}_{j,k} &= \frac{\sum_{t \in \mathcal{T}_k} (\upsilon^{\mathrm{hi}}_{j,t})_+}{\eta P(1-\alpha/2)}, & \pi^{\mathrm{hi},-}_{j,k} &= \frac{\sum_{t \in \mathcal{T}_k} (\upsilon^{\mathrm{hi}}_{j,t})_-}{\eta P(\alpha/2)}. \end{aligned} \tag{11}$$

For the lower endpoint, the positive mass records inward clipping at the median boundary, while the negative mass records outward clipping at $q_{\min}$. For the upper endpoint, the positive mass records outward clipping at $q_{\max}$, while the negative mass records inward clipping at the median boundary.

Lemma 1 (Endpoint-wise clipped-update identities)**.** *For every realized sequence and every period $\mathcal{T}_k$,*

$$\frac{1}{P}\sum_{t\in\mathcal{T}_k} e_{j,t}^{\mathrm{hi}} = \frac{\alpha}{2} + \frac{\delta_{j,t_k+P}^{\mathrm{hi}} - \delta_{j,t_k}^{\mathrm{hi}}}{\eta P} + \left(1-\frac{\alpha}{2}\right)\pi_{j,k}^{\mathrm{hi},+} - \frac{\alpha}{2}\pi_{j,k}^{\mathrm{hi},-},$$
$$\frac{1}{P}\sum_{t\in\mathcal{T}_k} e_{j,t}^{\mathrm{lo}} = \frac{\alpha}{2} - \frac{\delta_{j,t_k+P}^{\mathrm{lo}} - \delta_{j,t_k}^{\mathrm{lo}}}{\eta P} - \frac{\alpha}{2}\pi_{j,k}^{\mathrm{lo},+} + \left(1-\frac{\alpha}{2}\right)\pi_{j,k}^{\mathrm{lo},-}. \tag{12}$$

All four endpoint-specific clipping masses lie in $[0,1]$.

Proof. Equations 9 and 10 imply

$$\delta_{j,t}^{\mathrm{hi}} - \delta_{j,t-1}^{\mathrm{hi}} = \eta\left(e_{j,t}^{\mathrm{hi}} - \frac{\alpha}{2}\right) - \upsilon_{j,t}^{\mathrm{hi}},$$
$$\delta_{j,t}^{\mathrm{lo}} - \delta_{j,t-1}^{\mathrm{lo}} = \eta\left(\frac{\alpha}{2} - e_{j,t}^{\mathrm{lo}}\right) - \upsilon_{j,t}^{\mathrm{lo}}.$$

Summing these identities over $\mathcal{T}_k$, using $\upsilon = (\upsilon)_+ - (\upsilon)_-$, and applying Eq. 11 gives Eq. 12.

For the upper endpoint, an outward clip requires $e_{j,t}^{\mathrm{hi}} = 1$ and removes at most $\eta(1-\alpha/2)$, while an inward clip requires $e_{j,t}^{\mathrm{hi}} = 0$ and removes at most $\eta\alpha/2$. For the lower endpoint, an inward clip requires $e_{j,t}^{\mathrm{lo}} = 0$ and removes at most $\eta\alpha/2$, while an outward clip requires $e_{j,t}^{\mathrm{lo}} = 1$ and removes at most $\eta(1-\alpha/2)$. The normalizations in Eq. 11 therefore place all four clipping masses in $[0,1]$. □

Coverage Decomposition Across Periods. For each endpoint $e \in \{\mathrm{lo},\mathrm{hi}\}$, define its within-period state change as

$$\Delta_{j,k}^{e} = \delta_{j,t_k+P}^{e} - \delta_{j,t_k}^{e},$$

and define the empirical miscoverage in period $\mathcal{T}_k$ as

$$\mathrm{err}_{j,k} = \frac{1}{P}\sum_{t\in\mathcal{T}_k} \mathbf{1}\{Y_{j,t} \notin \widehat{C}_{j,t}\}.$$

The endpoint-specific clipping masses are grouped according to their directions:

$$\pi_{j,k}^{\mathrm{wid}} = \pi_{j,k}^{\mathrm{hi},+} + \pi_{j,k}^{\mathrm{lo},-}, \qquad \pi_{j,k}^{\mathrm{nar}} = \pi_{j,k}^{\mathrm{hi},-} + \pi_{j,k}^{\mathrm{lo},+}.$$

Proposition 1 (Period-wise miscoverage decomposition)**.** *For every component j and period $\mathcal{T}_k$,*

$$\mathrm{err}_{j,k} - \alpha = \frac{\Delta_{j,k}^{\mathrm{hi}} - \Delta_{j,k}^{\mathrm{lo}}}{\eta P} + \left(1-\frac{\alpha}{2}\right)\pi_{j,k}^{\mathrm{wid}} - \frac{\alpha}{2}\pi_{j,k}^{\mathrm{nar}}. \tag{13}$$

Proof. Because the adjusted quantile estimates are nondecreasing and the two endpoint levels are separated by the median boundaries,

$$\mathbf{1}\{Y_{j,t} \notin \widehat{C}_{j,t}\} = e_{j,t}^{\mathrm{lo}} + e_{j,t}^{\mathrm{hi}}.$$

Adding the two identities in Lemma 1, the two target terms sum to α, the tracker-state terms give $(\Delta_{j,k}^{\mathrm{hi}} - \Delta_{j,k}^{\mathrm{lo}})/(\eta P)$, and the endpoint-specific clipping masses combine according to their widening and narrowing directions. This gives Eq. 13. □

The decomposition is exact for every realized sequence and does not require exchangeability, stationarity, or accurate quantile estimates. Averaging over the K periods, define

$$\bar{\pi}_{j,T}^{d} = \frac{1}{K}\sum_{k=1}^{K} \pi_{j,k}^{d}, \qquad d \in \{\mathrm{wid},\mathrm{nar}\},$$

and

$$\mathcal{G}_{j,T} = \frac{1}{\eta T}\sum_{k=1}^{K}\left(\Delta_{j,k}^{\mathrm{hi}} - \Delta_{j,k}^{\mathrm{lo}}\right).$$

Theorem 2 (Component-wise coverage decomposition (restatement of Theorem 1))**.** *For every component j and every realized sequence of $T = KP$ online steps,*

$$\widehat{\mathrm{Cov}}_{j,T} - (1-\alpha) = -\mathcal{G}_{j,T} - \left(1 - \frac{\alpha}{2}\right)\bar{\pi}^{\mathrm{wid}}_{j,T} + \frac{\alpha}{2}\bar{\pi}^{\mathrm{nar}}_{j,T}. \tag{14}$$

Consequently,

$$-|\mathcal{G}_{j,T}| - \left(1 - \frac{\alpha}{2}\right)\bar{\pi}^{\mathrm{wid}}_{j,T} \le \widehat{\mathrm{Cov}}_{j,T} - (1-\alpha) \le |\mathcal{G}_{j,T}| + \frac{\alpha}{2}\bar{\pi}^{\mathrm{nar}}_{j,T}, \tag{15}$$

where

$$\widehat{\mathrm{Cov}}_{j,T} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_{j,t} \in \widehat{C}_{j,t}\}.$$

Proof. Averaging Eq. 13 over the K periods gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_{j,t} \notin \widehat{C}_{j,t}\} - \alpha = \mathcal{G}_{j,T} + \left(1 - \frac{\alpha}{2}\right)\bar{\pi}^{\mathrm{wid}}_{j,T} - \frac{\alpha}{2}\bar{\pi}^{\mathrm{nar}}_{j,T}.$$

Taking the complement gives Eq. 14. Since the two clipping masses are nonnegative and

$$-|\mathcal{G}_{j,T}| \le \mathcal{G}_{j,T} \le |\mathcal{G}_{j,T}|,$$

dropping the favorable clipping term on each side gives Eq. 15. □

We next express $\mathcal{G}_{j,T}$ in terms of the tracker states at the period boundaries. Define

$$\delta^{e,\mathrm{start}}_{j,k} = \delta^{e}_{j,t_k}, \qquad \delta^{e,\mathrm{end}}_{j,k} = \delta^{e}_{j,t_k+P},$$

and let

$$J^{e}_{j,k} = \delta^{e,\mathrm{start}}_{j,k+1} - \delta^{e,\mathrm{end}}_{j,k}$$

be the state change between periods k and $k+1$. Telescoping the within-period state changes gives

$$\eta T\, \mathcal{G}_{j,T} = \left(\delta^{\mathrm{hi,end}}_{j,K} - \delta^{\mathrm{hi,start}}_{j,1}\right) - \left(\delta^{\mathrm{lo,end}}_{j,K} - \delta^{\mathrm{lo,start}}_{j,1}\right) - \sum_{k=1}^{K-1}\left(J^{\mathrm{hi}}_{j,k} - J^{\mathrm{lo}}_{j,k}\right). \tag{16}$$

Let $|I_e| = \mathrm{diam}(I_e)$ and define

$$\mathcal{D}_{j,T} = \frac{1}{K}\sum_{k=1}^{K}\left(|\Delta^{\mathrm{lo}}_{j,k}| + |\Delta^{\mathrm{hi}}_{j,k}|\right).$$

Then

$$|\mathcal{G}_{j,T}| \le \min\left\{\frac{\mathcal{D}_{j,T}}{\eta P},\ \frac{|I_{\mathrm{hi}}| + |I_{\mathrm{lo}}| + \sum_{k=1}^{K-1}\left(|J^{\mathrm{hi}}_{j,k}| + |J^{\mathrm{lo}}_{j,k}|\right)}{\eta T}\right\}. \tag{17}$$

With continuous state carry-over, $J^{e}_{j,k} = 0$, so $|\mathcal{G}_{j,T}| = O((\eta T)^{-1})$. Under reset and replay, $\mathcal{G}_{j,T}$ still vanishes when $\eta T \to \infty$ and the cumulative period-boundary state changes are $o(\eta T)$. Fixed-frequency resets alone do not guarantee this condition. This completes the pathwise analysis of the coverage decomposition. Equation 17 characterizes when the tracker-state term is small. The remaining question is when the widening and narrowing clipping terms become small. We first study this question under accurate quantile estimates and then extend the analysis to misspecified estimates.

Accurate Quantile Estimates. For the stochastic results below, let $\mathcal{H}_t$ denote the information available before $Y_{j,t}$ is revealed. We assume that the adjusted quantile estimates and the pre-update tracker states used at time t are $\mathcal{H}_t$-measurable. Define the conditional expected tracker updates as

$$g^{\mathrm{hi}}_{j,t}(\delta) = \Pr\Big(R_{j,t} > \widetilde{Q}^{\tau_{\mathrm{hi}}+\delta}_{j,t} \mid \mathcal{H}_t\Big) - \frac{\alpha}{2},$$
$$g^{\mathrm{lo}}_{j,t}(\delta) = \frac{\alpha}{2} - \Pr\Big(R_{j,t} < \widetilde{Q}^{\tau_{\mathrm{lo}}+\delta}_{j,t} \mid \mathcal{H}_t\Big). \tag{18}$$

Suppose the conditional residual distribution given $\mathcal{H}_t$ is continuous and the adjusted quantile estimates equal the true conditional residual quantiles over the feasible levels:

$$\widetilde{Q}^{p}_{j,t} = Q_{R_{j,t}|\mathcal{H}_t}(p), \qquad p \in [q_{\min}, q_{\max}].$$

Then, for every feasible adjustment δ,

$$g^{\mathrm{hi}}_{j,t}(\delta) = g^{\mathrm{lo}}_{j,t}(\delta) = -\delta.$$

Thus, $\delta = 0$ is a common interior restoring point for the expected updates at every step. Conditional on the deployed tracker states,

$$\Pr\Big(Y_{j,t} \in \widehat{C}_{j,t} \mid \mathcal{H}_t\Big) = 1 - \alpha + \delta^{\mathrm{hi}}_{j,t-1} - \delta^{\mathrm{lo}}_{j,t-1}.$$

Therefore, the interval is conditionally calibrated when both tracker states are zero. Accurate quantile estimates do not make an interval with arbitrary nonzero tracker states conditionally calibrated, nor do they prevent random updates from reaching a boundary.

Define the minimum distance from the zero state to a tracker boundary as

$$m_0 = \min\left\{\tau_{\mathrm{lo}} - q_{\min}, \frac{1}{2} - \tau_{\mathrm{lo}}, q_{\max} - \tau_{\mathrm{hi}}, \tau_{\mathrm{hi}} - \frac{1}{2}\right\} > 0,$$

and let

$$D_I = \max\{|I_{\mathrm{lo}}|, |I_{\mathrm{hi}}|\}.$$

Lemma 2 (Clipping under accurate quantile estimates)**.** *Suppose the accurate-quantile condition above holds. If*

$$2\eta \le 1, \qquad \eta\left(1 - \frac{\alpha}{2}\right) \le \frac{m_0}{2},$$

then, for every period k and $d \in \{\mathrm{wid}, \mathrm{nar}\}$,

$$\mathbb{E}\big[\pi^{d}_{j,k}\big] \le \frac{8}{m_0^2}\left(\frac{D_I^2}{2\eta P} + \frac{\eta}{2}\right). \tag{19}$$

The same bound holds for $\mathbb{E}[\bar{\pi}^{d}_{j,T}]$.

Proof. For an endpoint $e \in \{\mathrm{lo}, \mathrm{hi}\}$, let

$$V^e_t = \mathbb{E}\Big[\big(\delta^e_{j,t}\big)^2\Big].$$

The clipping operation is non-expansive, and $g^e_{j,t}(\delta) = -\delta$, so

$$V^e_t \le (1 - 2\eta)V^e_{t-1} + \eta^2.$$

Summing this recursion over a period and using $V^e_{t_k} \le D_I^2$ gives

$$\frac{1}{P}\sum_{t \in \mathcal{T}_k} V^e_{t-1} \le \frac{D_I^2}{2\eta P} + \frac{\eta}{2}.$$

A clipped update requires the pre-update state to lie within $\eta(1 - \alpha/2)$ of a boundary. Under the stated step-size condition, the state must therefore be at least $m_0/2$ away from the restoring point. Markov's inequality gives

$$\Pr(\text{endpoint } e \text{ is clipped at time } t) \le \frac{4V^e_{t-1}}{m_0^2}.$$

Each normalized clip contributes at most $1/P$ to its directional clipping mass. Summing over the two endpoints and the period gives Eq. 19. Averaging the period-wise bound over $k = 1, \ldots, K$ gives the result for $\bar{\pi}^{d}_{j,T}$. □

Remark 2 (Long-run component-wise coverage)**.** *Theorem 2 first gives the model-free implication*

$$\mathcal{G}_{j,T} \to 0,\ \ \bar{\pi}^{\mathrm{wid}}_{j,T} \to 0,\ \ \bar{\pi}^{\mathrm{nar}}_{j,T} \to 0 \quad \Longrightarrow \quad \widehat{\mathrm{Cov}}_{j,T} \to 1-\alpha.$$

Under accurate quantile estimates, the clipping conditions need not be assumed separately. Consider a sequence of designs in which η is constant within each design. Combining Lemma 2 with

$$|\mathcal{G}_{j,T}| \le \frac{|I_{\mathrm{hi}}| + |I_{\mathrm{lo}}|}{\eta P}$$

gives

$$\mathbb{E}\left|\widehat{\mathrm{Cov}}_{j,T} - (1-\alpha)\right| \le \frac{|I_{\mathrm{hi}}| + |I_{\mathrm{lo}}|}{\eta P} + \frac{8}{m_0^2}\left(\frac{D_I^2}{2\eta P} + \frac{\eta}{2}\right).$$

Hence,

$$\eta \to 0, \qquad \eta P \to \infty$$

imply convergence to component-wise nominal coverage in L^1. For example, $\eta = P^{-1/2}$ satisfies both conditions. Increasing P with fixed η removes the $1/(\eta P)$ terms but leaves the $O(\eta)$ fluctuation term.

Misspecified Quantile Estimates. Exact conditional quantile accuracy is sufficient but not necessary. We now give conditions under which the tracker can compensate for misspecified quantile estimates.

For each endpoint $e \in \{\mathrm{lo}, \mathrm{hi}\}$, define the period-average conditional expected update as

$$g^e_{j,k}(\delta) = \frac{1}{P}\sum_{t \in \mathcal{T}_k} g^e_{j,t}(\delta).$$

For $k \ge 2$, replay the observations from $\mathcal{T}_{k-1}$ using the recalibrator parameters $(s_{j,k}, b_{j,k})$ deployed in period k, and denote the resulting empirical signed update by $\widehat{g}^e_{\mathcal{T}_{k-1}}(\delta)$.

Assumption 1 (Within-period tracker stability)**.** *There exist constants $c, m > 0$ such that, for every period k and endpoint $e \in \{\mathrm{lo}, \mathrm{hi}\}$, the following conditions hold.*

(i) Common restoring point. *There is an $\mathcal{H}_{t_k}$-measurable point $\delta^{e,\star}_{j,k} \in I_e$ satisfying*

$$\mathrm{dist}\left(\delta^{e,\star}_{j,k}, \partial I_e\right) \ge m$$

and

$$\left(\delta - \delta^{e,\star}_{j,k}\right) g^e_{j,t}(\delta) \le -c\left(\delta - \delta^{e,\star}_{j,k}\right)^2 \tag{20}$$

for every $t \in \mathcal{T}_k$ and $\delta \in I_e$.

(ii) Approximation across periods. *For $k \ge 2$, there is a point $\delta^{e,\star}_{\mathcal{T}_{k-1}} \in I_e$ satisfying*

$$\left|\widehat{g}^e_{\mathcal{T}_{k-1}}\left(\delta^{e,\star}_{\mathcal{T}_{k-1}}\right)\right| \le \varepsilon^{\mathrm{emp}}_{e,k}.$$

The discrepancy between the replayed and current expected updates is

$$\Delta_{P,k} = \max_{e \in \{\mathrm{lo},\mathrm{hi}\}} \sup_{\delta \in I_e} \left|\widehat{g}^e_{\mathcal{T}_{k-1}}(\delta) - g^e_{j,k}(\delta)\right|.$$

(iii) Replay error. *After reset and replay, the period-start tracker state satisfies*

$$\varepsilon^{\mathrm{rep}}_{e,k} = \mathbb{E}\left|\delta^{e,\star}_{\mathcal{T}_{k-1}} - \delta^{e,\mathrm{start}}_{j,k}\right|.$$

For the first period, define

$$\varepsilon^{\mathrm{init}}_{e} = \mathbb{E}\left|\delta^{e,\star}_{j,1} - \delta^{e,\mathrm{start}}_{j,1}\right| \le |I_e|.$$

No vanishing rate is imposed on $\Delta_{P,k}$, $\varepsilon^{\mathrm{emp}}_{e,k}$, or $\varepsilon^{\mathrm{rep}}_{e,k}$.

Condition (i) requires the same restoring point at every step within a period and is therefore stronger than requiring a zero only for the period-average expected update. Under accurate quantile estimates, it holds with $\delta_{j,k}^{e,\star} = 0$, $c = 1$, and $m = m_0$. The tolerance $\varepsilon_{e,k}^{\mathrm{emp}}$ is included because $\widehat{g}_{\mathcal{T}_{k-1}}^{e}$ is a step function and need not have an exact zero.

For later use, define the distances between the period-boundary states and the restoring point:

$$E_{j,k}^{e,\mathrm{start}} = \left|\delta_{j,k}^{e,\mathrm{start}} - \delta_{j,k}^{e,\star}\right|, \qquad E_{j,k}^{e,\mathrm{end}} = \left|\delta_{j,k}^{e,\mathrm{end}} - \delta_{j,k}^{e,\star}\right|.$$

They satisfy

$$|\Delta_{j,k}^{e}| \le E_{j,k}^{e,\mathrm{start}} + E_{j,k}^{e,\mathrm{end}}.$$

Lemma 3 (One-step contraction of the clipped tracker)**.** *Suppose Assumption 1(i) holds and* $2c\eta \le 1$. *For* $t \in \mathcal{T}_k$, *let*

$$V_t^e = \mathbb{E}\left[\left(\delta_{j,t}^{e} - \delta_{j,k}^{e,\star}\right)^2\right].$$

Then

$$V_t^e \le (1 - 2c\eta)V_{t-1}^e + \eta^2, \tag{21}$$

and consequently

$$V_t^e \le (1 - 2c\eta)^{t-t_k} V_{t_k}^e + \frac{\eta}{2c}.$$

Proof. For the two endpoints, define the signed feedback

$$Z_{j,t}^{\mathrm{lo}} = \frac{\alpha}{2} - e_{j,t}^{\mathrm{lo}}, \qquad Z_{j,t}^{\mathrm{hi}} = e_{j,t}^{\mathrm{hi}} - \frac{\alpha}{2}.$$

Then

$$\mathbb{E}\left[Z_{j,t}^{e} \mid \mathcal{H}_t\right] = g_{j,t}^{e}\left(\delta_{j,t-1}^{e}\right), \qquad |Z_{j,t}^{e}| \le 1.$$

Because clipping is non-expansive and $\delta_{j,k}^{e,\star} \in I_e$,

$$\mathbb{E}\left[\left(\delta_{j,t}^{e} - \delta_{j,k}^{e,\star}\right)^2 \mid \mathcal{H}_t\right] \le \left(\delta_{j,t-1}^{e} - \delta_{j,k}^{e,\star}\right)^2 + 2\eta\left(\delta_{j,t-1}^{e} - \delta_{j,k}^{e,\star}\right) g_{j,t}^{e}\left(\delta_{j,t-1}^{e}\right) + \eta^2.$$

Applying Eq. 20 and taking expectations gives Eq. 21. Unrolling the recursion gives the second bound. □

Lemma 3 gives

$$\mathbb{E}\left[E_{j,k}^{e,\mathrm{end}}\right] \le \sqrt{V_{t_k+P}^{e}} \le (1 - c\eta)^P \left(\mathbb{E}\left[\left|\delta_{j,k}^{e,\mathrm{start}} - \delta_{j,k}^{e,\star}\right|^2\right]\right)^{1/2} + \sqrt{\frac{\eta}{2c}}.$$

For $k \ge 2$, Eq. 20 and Assumption 1(ii) imply

$$\left|\delta_{j,k}^{e,\star} - \delta_{\mathcal{T}_{k-1}}^{e,\star}\right| \le \frac{\Delta_{P,k} + \varepsilon_{e,k}^{\mathrm{emp}}}{c}.$$

Combining this inequality with the replay error gives

$$\mathbb{E}\left[E_{j,k}^{e,\mathrm{start}}\right] \le \frac{\mathbb{E}[\Delta_{P,k}] + \mathbb{E}[\varepsilon_{e,k}^{\mathrm{emp}}]}{c} + \varepsilon_{e,k}^{\mathrm{rep}}.$$

Therefore, for $k \ge 2$,

$$\begin{aligned}\mathbb{E}\left[|\Delta_{j,k}^{e}|\right] \le (1 - c\eta)^P |I_e| + \sqrt{\frac{\eta}{2c}} \\ + \frac{\mathbb{E}[\Delta_{P,k}] + \mathbb{E}[\varepsilon_{e,k}^{\mathrm{emp}}]}{c} + \varepsilon_{e,k}^{\mathrm{rep}}.\end{aligned} \tag{22}$$

For the first period,

$$\mathbb{E}\left[|\Delta_{j,1}^{e}|\right] \le (1 - c\eta)^P |I_e| + \sqrt{\frac{\eta}{2c}} + \varepsilon_{e}^{\mathrm{init}}. \tag{23}$$

These bounds retain the approximation and replay errors explicitly. Even with an accurate period start, a constant tracker learning rate leaves an $O(\sqrt{\eta})$ fluctuation term.

Saturation and Boundary Clipping. We first connect the observable saturation statistic to the boundary clipping terms in Theorem 2. Recall that $\mathrm{sat}_{j,k}$ measures the use of the available outward adjustment range during period k. For notational clarity, write

$$\mathrm{sat}^{\mathrm{wid}}_{j,k} := \mathrm{sat}_{j,k}.$$

For the inward direction, define the auxiliary quantity

$$\mathrm{sat}^{\mathrm{nar}}_{j,k} = \frac{1}{P}\sum_{t\in\mathcal{T}_k} \max\left\{\frac{(\delta^{\mathrm{lo}}_{j,t})_+}{1/2-\tau_{\mathrm{lo}}}, \frac{(-\delta^{\mathrm{hi}}_{j,t})_+}{\tau_{\mathrm{hi}}-1/2}\right\}. \tag{24}$$

Their averages over the evaluation horizon are

$$\overline{\mathrm{sat}}^{d}_{j,T} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{sat}^{d}_{j,k}, \qquad d\in\{\mathrm{wid},\mathrm{nar}\}.$$

We abbreviate $\overline{\mathrm{sat}}_{j,T} := \overline{\mathrm{sat}}^{\mathrm{wid}}_{j,T}$. The narrowing quantity is introduced only for the theoretical upper coverage bound; BaCR does not use it as a recalibration signal.

Lemma 4 (Saturation bounds boundary clipping)**.** *For every component j and period k,*

$$\pi^{\mathrm{wid}}_{j,k} \le \mathrm{sat}^{\mathrm{wid}}_{j,k}, \qquad \pi^{\mathrm{nar}}_{j,k} \le 2\,\mathrm{sat}^{\mathrm{nar}}_{j,k}.$$

Consequently,

$$\bar{\pi}^{\mathrm{wid}}_{j,T} \le \overline{\mathrm{sat}}_{j,T}, \qquad \bar{\pi}^{\mathrm{nar}}_{j,T} \le 2\,\overline{\mathrm{sat}}^{\mathrm{nar}}_{j,T}.$$

Proof. Whenever an endpoint update is clipped in a given direction, the post-update state lies on the corresponding boundary, so the associated normalized displacement equals one. Widening clips at the two endpoints cannot occur simultaneously, which gives the first inequality. Narrowing clips may occur at both endpoints at the same time, producing the factor two in the second inequality. Averaging over periods gives the remaining statements. □

Lemma 4 shows that the saturation statistic directly upper-bounds the widening clipping term responsible for undercoverage in Theorem 2. In particular,

$$\widehat{\mathrm{Cov}}_{j,T} \ge 1-\alpha-|\mathcal{G}_{j,T}| - \left(1-\frac{\alpha}{2}\right)\overline{\mathrm{sat}}_{j,T}.$$

Thus, once the tracker-state term is small, low widening saturation implies a small component-wise coverage deficit.

We next give the conditional finite-time specialization used in the second case of Remark 1. For each endpoint $e\in\{\mathrm{lo},\mathrm{hi}\}$, define

$$\bar{\Delta}_K = \frac{1}{K}\sum_{k=2}^{K}\mathbb{E}[\Delta_{P,k}],$$

$$\bar{\varepsilon}^{\mathrm{emp}}_{e,K} = \frac{1}{K}\sum_{k=2}^{K}\mathbb{E}\left[\varepsilon^{\mathrm{emp}}_{e,k}\right],$$

$$\bar{\varepsilon}^{\mathrm{rep}}_{e,K} = \frac{1}{K}\sum_{k=2}^{K}\varepsilon^{\mathrm{rep}}_{e,k}.$$

Under Assumption 1, define

$$r_K(\eta,P) = \sum_{e\in\{\mathrm{lo},\mathrm{hi}\}}\left[\frac{(1-c\eta)^P|I_e|}{\eta P} + \frac{1}{\sqrt{2c\eta}\,P} + \frac{\bar{\Delta}_K+\bar{\varepsilon}^{\mathrm{emp}}_{e,K}}{c\eta P} + \frac{\bar{\varepsilon}^{\mathrm{rep}}_{e,K}}{\eta P} + \frac{\varepsilon^{\mathrm{init}}_{e}}{\eta T}\right].$$

Proposition 2 (Finite-time tracker-state bound)**.** *Under Assumption 1,*

$$\mathbb{E}[|\mathcal{G}_{j,T}|] \le r_K(\eta,P).$$

Proof. The first inequality in Eq. 17 gives

$$\mathbb{E}\left[|\mathcal{G}_{j,T}|\right] \le \frac{1}{\eta P K} \sum_{e\in\{\mathrm{lo},\mathrm{hi}\}} \sum_{k=1}^{K} \mathbb{E}\left[|\Delta_{j,k}^{e}|\right].$$

Apply Eq. 23 to the first period and Eq. 22 to periods $k = 2, \ldots, K$. Summing the resulting bounds and using the definitions of $\bar{\Delta}_K$, $\bar{\varepsilon}_{e,K}^{\mathrm{emp}}$, and $\bar{\varepsilon}_{e,K}^{\mathrm{rep}}$ gives $r_K(\eta, P)$. □

The same interior margin also limits how often projection can alter an endpoint update.

Lemma 5 (Boundary-clipping bound under interior stability)**.** *Suppose Assumption 1 holds. If*

$$2\eta\left(1 - \frac{\alpha}{2}\right) \le m,$$

then, for every period k and every clipping direction d,

$$\mathbb{E}[\pi_{j,k}^{d}] \le \frac{8}{m^2}\left(\frac{D_I^2}{2c\eta P} + \frac{\eta}{2c}\right), \tag{25}$$

where

$$D_I = \max_{e\in\{\mathrm{lo},\mathrm{hi}\}} |I_e|,$$

Proof. A clipped update requires the pre-update state to lie within one update step of the corresponding boundary. Under the stated margin condition, this implies that the state is at least $m/2$ from the restoring point. Markov's inequality, followed by the within-period contraction bound, then yields Eq. 25. □

Remark 3 (Conditional finite-time specialization)**.** *Combining Theorem 2, Proposition 2, and Lemma 5 gives*

$$\mathbb{E}\left[\left|\widehat{\mathrm{Cov}}_{j,T} - (1-\alpha)\right|\right] \le r_K(\eta, P) + \frac{8}{m^2}\left(\frac{D_I^2}{2c\eta P} + \frac{\eta}{2c}\right).$$

Hence the expected component-wise coverage error vanishes when the right-hand side tends to zero. For example, the stochastic terms vanish under $\eta \to 0$ and $\eta P \to \infty$, while the averaged approximation, replay, and initialization errors must vanish at the corresponding rates.

The margin and step-size conditions above are additional requirements for this conditional finite-time specialization. They are not required for the exact pathwise decomposition in Theorem 2 or the saturation bounds in Lemma 4, and they are not used to justify the fixed experimental hyperparameters.

A.1 Conditional Role of the Unified Recalibrator.

The preceding results show that widening clipping is the boundary term that lowers the component-wise coverage bound, while Lemma 4 shows that the observable saturation statistic upper-bounds this term. BaCR passes the saturation signal to the unified recalibrator through Eq. 6. The following result gives a conditional interpretation of how the resulting scale expansion can make an interior restoring point attainable after one recalibration step.

Lemma 6 (Conditional reachability of an interior restoring point)**.** *Fix component j and period $k{+}1$, and abbreviate the returned bias $b_{j,k+1}$ by b. For $s > 0$, define*

$$\widetilde{Q}_{j,t}^{p}(s, b) = \widehat{Q}_{j,t}^{1/2} + s\left(\widehat{Q}_{j,t}^{p} - \widehat{Q}_{j,t}^{1/2}\right) + b.$$

Suppose the pretrained quantile estimates are nondecreasing in p and, almost surely, the period-average conditional exceedance probability is continuous in the effective level. Assume that

$$\frac{1}{P}\sum_{t\in\mathcal{T}_{k+1}} \Pr\Big(R_{j,t} > \widetilde{Q}_{j,t}^{1/2}(s, b) \mid \mathcal{H}_t\Big) > \frac{\alpha}{2}. \tag{26}$$

The left-hand side is independent of s. For $\varepsilon_0 \in (0, \alpha/2)$, define the ex-post required scale by

$$s^{\mathrm{req}}_{j,k+1}(b) = \inf\left\{ s > 0 : \frac{1}{P} \sum_{t \in \mathcal{T}_{k+1}} \Pr\Big(R_{j,t} > \widetilde{Q}^{q_{\max}}_{j,t}(s, b) \mid \mathcal{H}_t \Big) \le \frac{\alpha}{2} - \varepsilon_0 \right\}. \tag{27}$$

Suppose this value is finite and the defining inequality holds at $s = s^{\mathrm{req}}_{j,k+1}(b)$. If the recalibrator returns

$$s_{j,k+1} \ge s^{\mathrm{req}}_{j,k+1}(b),$$

then, with $g^{\mathrm{hi}}_{j,k+1}$ evaluated using the returned parameters,

$$g^{\mathrm{hi}}_{j,k+1}\left(\frac{1}{2} - \tau_{\mathrm{hi}}\right) > 0, \qquad g^{\mathrm{hi}}_{j,k+1}\left(q_{\max} - \tau_{\mathrm{hi}}\right) \le -\varepsilon_0.$$

Consequently, there exists

$$\delta^{\mathrm{hi},\star}_{j,k+1} \in \left(\frac{1}{2} - \tau_{\mathrm{hi}}, q_{\max} - \tau_{\mathrm{hi}}\right)$$

such that

$$g^{\mathrm{hi}}_{j,k+1}\left(\delta^{\mathrm{hi},\star}_{j,k+1}\right) = 0.$$

Moreover, let $x = \log s$ and suppose the pinball term in Eq. 6, viewed as a function of x at the returned bias, has subgradients bounded in magnitude by G_{pin}. At any minimizer,

$$|\log s_{j,k+1} - \kappa \,\mathrm{sat}_{j,k}| \le \frac{G_{\mathrm{pin}}}{2\lambda}. \tag{28}$$

Therefore, the sufficient condition

$$\kappa \,\mathrm{sat}_{j,k} - \frac{G_{\mathrm{pin}}}{2\lambda} \ge \log s^{\mathrm{req}}_{j,k+1}(b)$$

ensures that the returned scale reaches the required value. The lower endpoint follows under analogous conditions.

Proof. Because $\widehat{Q}^{q_{\max}}_{j,t} - \widehat{Q}^{1/2}_{j,t} \ge 0$, the adjusted estimate at $q_{\max}$ is nondecreasing in s, and its conditional exceedance probability is therefore nonincreasing. The required-scale condition gives a nonpositive expected update at the outer boundary, whereas Eq. 26 gives a positive expected update at the inner boundary. Continuity and the intermediate value theorem then give an interior zero.

For the optimization statement, first-order subgradient optimality in $x = \log s$ gives

$$0 \in \partial_x \overline{\mathcal{L}}^{\mathrm{pin}}_{j,k}(e^x, b) + 2\lambda\left(x - \kappa \,\mathrm{sat}_{j,k}\right).$$

The assumed subgradient bound yields Eq. 28. □

Lemma 6 is a conditional single-refit result. The quantity $s^{\mathrm{req}}_{j,k+1}(b)$ depends on the next-period conditional distribution and is not available when the recalibrator is fitted. The lemma also establishes only a zero of the period-average expected update. It does not imply the common restoring point, quantitative restoring condition, or positive margin required by Assumption 1(i), and it does not establish convergence of the complete tracker–recalibrator procedure. Its role is to explain why saturation-dependent scale expansion is directed at the boundary limitation identified by Theorem 2; its realized effect is examined in the mechanism diagnostics.

Conditional distribution-shift interpretation. Because the threshold varies with time and covariates, total variation must compare conditional residual laws rather than residual marginals alone. Let $P^0_{j,t}(\cdot \mid \mathcal{H}_t)$ be a reference law and $P_{j,t}(\cdot \mid \mathcal{H}_t)$ the test law. Suppose, almost surely,

$$d_{\mathrm{TV}}\big(P_{j,t}(\cdot \mid \mathcal{H}_t), P^0_{j,t}(\cdot \mid \mathcal{H}_t)\big) \le \rho_{j,t},$$

and the predictable outer threshold is reference-calibrated:

$$P^0_{j,t}\Big(R_{j,t} > \widetilde{Q}^{q_{\max}}_{j,t} \mid \mathcal{H}_t\Big) \le 1 - q_{\max}.$$

With $\bar{\rho}_{j,k} = P^{-1} \sum_{t \in \mathcal{T}_k} \rho_{j,t}$, averaging the event-wise TV inequality gives

$$\left[\frac{1}{P} \sum_{t \in \mathcal{T}_k} P_{j,t}\Big(R_{j,t} > \widetilde{Q}_{j,t}^{q_{\max}} \mid \mathcal{H}_t \Big) - \frac{\alpha}{2} \right]_+ \leq \left[\bar{\rho}_{j,k} - (q_{\max} - \tau_{\mathrm{hi}}) \right]_+ .$$

The lower-tail statement is analogous.

Remark 4 (Shift regime for the two clipping floors)**.** *If both reference outer quantiles are conditionally calibrated and $\bar{\rho}_{j,k} \leq \delta_{\max} - \varepsilon_0$, the period-average outer tail maps have the required outer-wall signs. Establishing value-space reach additionally requires the corresponding opposite signs at the inner walls and continuity of the update maps. These conditions do not imply the common-root or restoring-slope condition used in the stochastic bound. Conversely, observed clipping does not identify or lower-bound total variation: initialization, replay, finite-step noise, and quantile sensitivity can also produce clips. Total variation alone does not determine the value-space scale required by the affine head without additional density or quantile-sensitivity assumptions.*

B Additional Experimental Details

B.1 Experimental implementation

The methodology and theory are presented for a single fixed forecast horizon. In the multi-step experiments, a separate endpoint tracker is maintained for each component–horizon pair, while one affine head is shared across horizons for each component. The implementation anchors this shared head using

$$\mathrm{sat}_{j,k}^{\mathrm{agg}} = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \max_{1 \leq h \leq H} \max\left\{ \frac{(-\delta_{j,h,t}^{\mathrm{lo}})_+}{\mathrm{hr}_{\mathrm{lo}}^{\mathrm{wid}}}, \frac{(\delta_{j,h,t}^{\mathrm{hi}})_+}{\mathrm{hr}_{\mathrm{hi}}^{\mathrm{wid}}} \right\}.$$

The head-fitting loss pools all valid residual–horizon pairs in the rolling recalibration buffer. The reported all-step metrics likewise pool valid predictions over forecast horizons, as defined in Appendix B.1.

Feedback timing. All online methods, BaCR and the online baselines alike, are evaluated with the same block-synchronous feedback protocol: after the forecasts issued at origin t are scored, the realised targets $Y_t^{(h)} = X_{t+h}$ of all horizons $h = 1, \ldots, H$ are fed back before the forecasts at origin $t + 1$ are issued. This covers the tracker updates and the recalibration buffer of BaCR as well as the score pools and state updates of the online baselines. For $h = 1$ this coincides with deployment. For $h \geq 2$ the target is used $h - 1$ steps before it would be observed in deployment ($H = 3$ on AIR and $H = 12$ on LA, BAY and SD), so the reported results correspond to an idealised feedback schedule that is shared by all methods rather than to strictly causal deployment.

Real Datasets. We evaluate on four standard MTSF datasets. For each, we follow the preprocessing steps of a public library (Cini & Marisca, 2022) and previous works (Cini et al., 2023). We adopt a chronological $48\%/32\%/20\%$ split for training, calibration, and testing.

- **AIR** (Zheng et al., 2015): This dataset consists of 8,760 hourly measurements of pollutant PM2.5 from 437 monitoring stations in China. We use window of 24 time steps as inputs and predict 3 steps observations.
- **LA** (METR-LA) and **BAY** (PEMS-BAY) (Li et al., 2017): They are 5-minute traffic-*speed* readings from 207 loop detectors on Los Angeles highways and 325 sensors in the San Francisco Bay Area respectively. We use window of 12 steps as inputs and predict 12 steps.
- **SD** (Liu et al., 2023): This dataset records the 716-sensor San Diego subset of LargeST and, being traffic flow rather than speed, has a markedly larger residual scale; it is the highest- dimensional benchmark here and exhibits the most pronounced shifted-tail (S^+) regime. We use window of 12 steps as inputs and predict 12 steps.

Underlying Algorithms. We consider two point forecasters: RNN with GRU cells (1 layer with hidden size 32) and a decoder-only Transformer (hidden size 32, feed-forward 64, 2 heads, 3 layers, dropout 0.1). Both are trained with masked MAE.

Residual Quantile Model. We construct the signed residual sequence on the calibration split and train a residual quantile model Ψ using the relational architecture of CoRel. The model encodes recent residual histories and cross-component relations and outputs conditional residual quantile estimates on the grid $\mathcal{Q} = \{0.01, 0.02, \ldots, 0.99\}$. We use a hidden size of 64, a node-embedding size of 16, two graph layers, and 20 neighbors in the learned relational graph. The model is trained with the masked multi-quantile pinball loss on the calibration set. We use Adam with learning rate 0.003, batch size 64, and at most 100 epochs. The selected model is then frozen throughout online evaluation. CoRel and BaCR start from the same pretrained residual quantile model and its raw quantile estimates. They then differ in what is allowed to change at test time: BaCR keeps Ψ frozen and adapts only the tracker states and the component-wise affine head, whereas CoRel continues to update the regressor head of Ψ online.

Recalibrator. The quantile-estimate recalibrator is implemented as the component-wise affine head defined in Section Methodology. Each head contains a positive scale and a bias, is centered at the estimated residual median, and is initialized as the identity transformation. Every $P = 36$ forecast origins, only the affine-head parameters are updated for one pass over a rolling buffer containing the most recent $B = 128$ forecast origins. We use Adam with learning rate 0.01 and the saturation-anchored multi-quantile objective with $\lambda = 1$ and $\kappa = 0.5$. After each update, the tracker states are reset and replayed over the just-completed period. The point forecaster and residual quantile model remain frozen.

Baselines. We compare against seven conformal methods spanning the Related Work, according to its original paper and public code. Unless noted, all baselines share the same frozen backbone, the same calibration split, and the same per-node, per-horizon protocol as our method. We report the design of each below.

Traditional. **SCP** (Papadopoulos et al., 2002) is the static floor. Using the nonconformity score (absolute value), it calibrates a per-node, per-horizon radius equal to the finite-sample-corrected empirical quantile of the calibration scores (the $\lceil (m+1)(1-\alpha) \rceil$-th smallest of the calibration scores for that node) and applies it unchanged on the test stream. It performs no online adaptation.

Weighted. **NexCP** (Barber et al., 2023) keeps SCP's score machinery but replaces the equal-weighted calibration quantile by a fixed, recency-weighted one: at step t each past score R_i receives a data-independent geometric weight $w_i = \rho^{t-i}$ ($\rho = 0.99$) over a recency window of 300 steps; the radius is the weighted $(1-\alpha)$ quantile of that pool. It reweights the score distribution toward recent observations. **HopCPT** (Auer et al., 2023) is a learned method: a modern Hopfield network associates the current step with past steps in the same error regime, where a two-layer MLP encoder is employed (hidden 32, embedding 32). We train for 300 epochs (Adam, learning rate 10^{-3}).

Quantile regression. **SPCI** (Xu & Xie, 2023) fits a quantile random forest (10 trees, maximum depth 2) that predicts the conditional quantiles of the next residual from the last $w = 100$ residuals as autoregressive features. For **CoRel** (Cini et al., 2025), we use the relational quantile network architecture with hidden size 64, embedding size 16, two graph layers, and $K = 20$ neighbors in the graph-learning module. The model is trained on the calibration split with 10% of the calibration data held out for validation. We train for at most 100 epochs with batch size 64 using Adam with initial learning rate 0.003. The best checkpoint is selected by validation Winkler score.

Online adaptive. Both **ACI** (Gibbs & Candes, 2021) and conformal **PID** (Angelopoulos et al., 2023a) use the symmetric absolute-residual score relative to the frozen median forecast and maintain a separate state for every component–horizon pair. For ACI, the rolling score window is initialized by the corresponding calibration scores and retains that initial window length, appending each newly observed score and evicting the oldest. At round t, its radius is the empirical $(1-\alpha_{j,h,t})$ quantile of this window and $\alpha_{j,h,t+1} = \Pi_{[0.001,0.5]}\{\alpha_{j,h,t} + \gamma(\alpha - \mathrm{err}_{j,h,t})\}$, where $\alpha = 0.1$ is the nominal miscoverage level. For PID, the score-quantile state is initialized at the empirical calibration $(1-\alpha)$ quantile. Let $\widehat{B}_{j,h,t}$ be the largest observed score in a trailing window of 100 updates. The proportional step size is $\eta_{j,h,t} = \beta \widehat{B}_{j,h,t}$, and we use the PI controller con-

sisting of the quantile-tracking update $q_{j,h,t+1} = q_{j,h,t} + \eta_{j,h,t}(\mathrm{err}_{j,h,t} - \alpha)$ and the tangent integrator $\widehat{B}_{j,h,t} \tan\{E_{j,h,t} \log(\max\{t,2\})/[\max\{t,2\}C_{\mathrm{sat}}]\}$, where $E_{j,h,t} = \sum_{r<t}(\mathrm{err}_{j,h,r} - \alpha)$ and $C_{\mathrm{sat}} = \max\{(2/\pi)[\lceil 0.1 \log T \rceil - 1/\log T], 10^{-3}\}$. The resulting radius is truncated below at zero. We leave the scorecaster (D term) off, so the reported baseline is the PI rather than the full PID configuration. For ACI and PID, we select dataset-specific learning rates from $\{0.005, 0.01, 0.03, 0.05, 0.1\}$; in the order AIR/LA/BAY/SD, ACI uses $(0.1, 0.05, 0.05, 0.05)$ and PID uses $(0.1, 0.03, 0.03, 0.1)$

Evaluation Metrics. We evaluate every method using an empirical coverage metric and an efficiency metric. The primary results average these metrics over all components. To diagnose heterogeneous performance under distribution shift, we report two evaluation metrics from three views: over all components, over S^+ slice (the top 10% components with the largest shifts), and over S^- slice (the top 10% components with the smallest shifts).

Coverage deviation. The reported all-step metrics pool all valid forecast horizons. For component j, let

$$\mathcal{O}_j^{\mathrm{test}} = \{(t,h) : Y_{j,t}^{(h)} \text{ is observed in the test split}\}.$$

Its empirical all-step coverage is

$$\mathrm{Cov}_j = \frac{1}{|\mathcal{O}_j^{\mathrm{test}}|} \sum_{(t,h)\in\mathcal{O}_j^{\mathrm{test}}} \mathbf{1}\{Y_{j,t}^{(h)} \in \widehat{C}_{j,t}^{(h)}\}.$$

We report the *coverage deviation* $\Delta\mathrm{Cov} = 100 \cdot (\overline{\mathrm{Cov}} - (1-\alpha))$ in percentage points, where $\overline{\mathrm{Cov}}$ is the mean of Cov_j over the component set in view (all components, S^+ and S^-).

Winkler score. For a nominal $(1-\alpha)$ interval $\widehat{C}_{j,t}^{(h)} = [\ell_{j,t}^{(h)}, u_{j,t}^{(h)}]$ and observation $Y_{j,t}^{(h)}$, the Winkler score (Winkler, 1972) is

$$\begin{aligned}\mathrm{W}_{j,t}^{(h)} = u_{j,t}^{(h)} - \ell_{j,t}^{(h)} &+ \frac{2}{\alpha}(\ell_{j,t}^{(h)} - Y_{j,t}^{(h)})\mathbf{1}\{Y_{j,t}^{(h)} < \ell_{j,t}^{(h)}\} \\ &+ \frac{2}{\alpha}(Y_{j,t}^{(h)} - u_{j,t}^{(h)})\mathbf{1}\{Y_{j,t}^{(h)} > u_{j,t}^{(h)}\}.\end{aligned}$$

The Winkler score jointly penalizes interval width and the magnitude of coverage misses, with lower values indicating better interval quality. Because empirical validity is our primary criterion, we use the Winkler score to compare efficiency after checking whether a method satisfies the coverage-acceptable band.

The S^+ shift slice. The S^+ slice is a secondary, method-independent stress diagnostic that identifies components exhibiting the largest increase in residual tail scale from calibration to test. Such components are expected to challenge a frozen quantile geometry. The slice is constructed solely from the frozen point-forecaster residuals $R_{j,t}^{(h)} = Y_{j,t}^{(h)} - \widehat{Y}_{j,t}^{(h)}$ and is therefore independent of the subsequent calibration method. Let $\mathcal{O}_j^{\mathrm{calib}}$ be the calibration-split analogue of $\mathcal{O}_j^{\mathrm{test}}$.

For each eligible component j, we define the all-step tail-growth ratio

$$\rho_j = \frac{q_{0.9}\Big(\{|R_{j,t}^{(h)}| : (t,h) \in \mathcal{O}_j^{\mathrm{test}}\}\Big)}{q_{0.9}\Big(\{|R_{j,t}^{(h)}| : (t,h) \in \mathcal{O}_j^{\mathrm{calib}}\}\Big) + \varepsilon},$$

where $\varepsilon > 0$ is a small numerical stabilizer. A component is eligible if it has at least 20 observed residual–horizon pairs in both the calibration and test segments and if its number of valid test pairs is at least half of the maximum across components. These conditions prevent unstable tail estimates for sparsely observed components.

We define S^+ as the top decile of eligible components ranked by ρ_j. The bottom decile defines S^- and represents components with the smallest relative tail growth, which typically corresponds to a reduction in residual scale. Because the slices are defined using the realized test residuals, they are used only as descriptive, ex-post stress diagnostics and never for model selection or hyperparameter tuning. The all-component results remain our primary evaluation.

B.2 MAIN RESULTS.

Tab. 3 reports the complete results under three views: all components, the S^+ slice (the top 10% components with the largest shifts), and the S^- slice (the top 10% components with the smallest shifts).

All components. BaCR provides the strongest overall validity–efficiency trade-off. It remains within the prescribed coverage band in all eight dataset–backbone combinations and attains the lowest Winkler score among coverage-acceptable methods in every case. ACI also provides highly consistent calibration, but BaCR obtains a lower Winkler score than ACI in all eight comparisons. The conclusions are nearly unchanged between the Transformer and RNN backbones, suggesting that the improvement primarily arises from the calibration layer.

Largest-shift components (S^+). The S^+ slice exposes failures that are hidden by the all-component averages. Static and conditional-quantile methods exhibit substantial undercoverage on several datasets, most notably SCP on LA and CoRel on AIR and BAY. ACI remains within the coverage band in all eight dataset–backbone cells, demonstrating the effectiveness of direct online coverage feedback. BaCR obtains a lower raw Winkler score than ACI in every cell and is coverage-acceptable on AIR, LA, and SD. BAY is the only borderline case: BaCR attains the lowest raw Winkler score but falls slightly outside the band (-1.2%), making ACI the best coverage-acceptable method on this slice.

Smallest-shift components (S^-). BaCR remains within the coverage band in all eight S^- cells and attains the lowest coverage-acceptable Winkler score in six of them. The two exceptions are the AIR results, where CoRel remains coverage-acceptable and obtains a lower Winkler score. The behavior of the frozen baselines is dataset-dependent: several methods overcover on AIR and BAY, whereas SD remains broadly undercovered even in its lowest-growth decile. Therefore, S^- should be interpreted as a relative lowest-growth diagnostic rather than as a uniformly stationary or contracting regime.

Overall, BaCR attains the lowest coverage-acceptable Winkler score in 20 of the 24 dataset–view–backbone cells. It also obtains a lower Winkler score than ACI in all 24 cells, while remaining within the coverage band in 22 of them.

B.3 ABLATION STUDY

B.3.1 ABLATION STRATIFIED BY LOCAL TAIL PRESSURE

Definition. Let R_j denote the frozen point-forecaster residuals of Appendix B.1, and let $\mathcal{W}_1, \dots, \mathcal{W}_K$ partition the test segment along the target-time axis into consecutive windows of W steps, anchored at the first and last test target. Using the calibration segment only, set $c_j = \mathrm{med}(R_j^{\mathrm{calib}})$ and define the outward reference radii at the two grid levels, $a_j^{\mathrm{hi}} = q_{q_{\max}}(R_j^{\mathrm{calib}}) - c_j$ and $a_j^{\mathrm{lo}} = c_j - q_{q_{\min}}(R_j^{\mathrm{calib}})$. The *local tail pressure* of component j in window k is

$$\chi_{j,k} = \max\left\{ \big(q_{1-\alpha/2}(R_j^{\mathcal{W}_k}) - c_j\big)/a_j^{\mathrm{hi}},\ \big(c_j - q_{\alpha/2}(R_j^{\mathcal{W}_k})\big)/a_j^{\mathrm{lo}} \right\}. \tag{29}$$

Each forecast–horizon pair is assigned to a window by its own target time, so the five strata $[0, 0.6)$, $[0.6, 0.8)$, $[0.8, 1)$, $[1, 1.2)$, $[1.2, \infty)$ partition the evaluated pairs and every method is compared on an identical support. Metrics are pooled within a seed over valid pairs and averaged with equal weight across seeds.

Eligibility and support. A component is included for a given window when its calibration segment has at least 200 valid residuals, both reference radii are strictly positive, and the window itself has at least $\max(40, \lceil W/5 \rceil)$ valid residuals. Window quantiles use one base-forecast residual per target time, whereas the reported metrics retain every valid forecast–horizon pair, so all horizons are evaluated. Between 98.9% and 100% of valid forecast pairs fall in an eligible component and window. The main analysis uses $W = 3P$; $W = 6P$ and $W = 12P$ are reported below as a sensitivity analysis. The bin edges are the ones used during exploration and are not claimed to be preregistered.

Interpretation. χ compares a window's realised tail against a fixed calibration reference radius. A value above one means that at least one side of the window quantile lies outside that radius; it

Table 3: Complete real-data results (both backbones, mean$_{\pm\text{std}}$ over 10 seeds), including the S^- slice (least-shifted decile). This is the complete version of Tab. 2.

			SCP	NexCP	HopCPT	SPCI	CoRel	ACI	PID	BaCR
AIR	Tr	ΔCov	$+3.2_{\pm0.1}$	$+1.0_{\pm0.0}$	$+2.4_{\pm0.3}$	$+1.8_{\pm0.1}$	$-1.9_{\pm0.4}$	$-0.5_{\pm0.0}$	$-0.6_{\pm0.0}$	$+0.6_{\pm0.1}$
		Winkler↓	$111.8_{\pm1.0}$	$104.0_{\pm0.8}$	$112.5_{\pm4.9}$	$102.8_{\pm0.9}$	$88.8_{\pm0.7}$	$99.0_{\pm0.6}$	$98.7_{\pm0.6}$	$\mathbf{89.5_{\pm0.8}}$
		ΔCov_{S+}	$-1.4_{\pm0.2}$	$+0.9_{\pm0.0}$	$-0.0_{\pm0.4}$	$-1.7_{\pm0.2}$	$-5.4_{\pm0.6}$	$-0.8_{\pm0.0}$	$-0.9_{\pm0.0}$	$+0.4_{\pm0.2}$
		Winkler_{S+} ↓	$110.4_{\pm2.2}$	$104.1_{\pm2.3}$	$152.2_{\pm29.2}$	$104.1_{\pm1.9}$	$90.3_{\pm2.3}$	$97.1_{\pm2.2}$	$97.1_{\pm2.3}$	$\mathbf{86.6_{\pm2.1}}$
		ΔCov_{S-}	$+5.1_{\pm0.2}$	$+1.1_{\pm0.0}$	$+4.0_{\pm0.2}$	$+3.0_{\pm0.8}$	$-0.3_{\pm0.5}$	$-0.3_{\pm0.0}$	$-0.5_{\pm0.0}$	$+0.9_{\pm0.1}$
		Winkler_{S-} ↓	$147.9_{\pm2.7}$	$131.6_{\pm2.6}$	$136.2_{\pm3.4}$	$134.4_{\pm2.0}$	$\mathbf{110.1_{\pm2.4}}$	$127.4_{\pm2.6}$	$126.6_{\pm2.7}$	$115.2_{\pm2.5}$
	RNN	ΔCov	$+3.2_{\pm0.1}$	$+1.0_{\pm0.0}$	$+2.5_{\pm0.1}$	$+1.7_{\pm0.1}$	$-1.8_{\pm0.2}$	$-0.5_{\pm0.0}$	$-0.6_{\pm0.0}$	$+0.6_{\pm0.1}$
		Winkler↓	$111.7_{\pm1.0}$	$104.0_{\pm0.7}$	$110.9_{\pm3.7}$	$102.7_{\pm0.7}$	$88.6_{\pm0.3}$	$99.1_{\pm0.6}$	$98.7_{\pm0.5}$	$\mathbf{89.8_{\pm0.7}}$
		ΔCov_{S+}	$-1.2_{\pm0.2}$	$+0.9_{\pm0.0}$	$+0.2_{\pm0.4}$	$-1.7_{\pm0.2}$	$-5.2_{\pm0.7}$	$-0.8_{\pm0.0}$	$-0.9_{\pm0.0}$	$+0.4_{\pm0.2}$
		Winkler_{S+} ↓	$105.3_{\pm1.1}$	$99.1_{\pm1.0}$	$133.4_{\pm19.3}$	$99.0_{\pm1.2}$	$85.8_{\pm1.1}$	$92.3_{\pm1.1}$	$92.4_{\pm0.9}$	$\mathbf{82.4_{\pm0.9}}$
		ΔCov_{S-}	$+5.1_{\pm0.1}$	$+1.2_{\pm0.0}$	$+4.3_{\pm0.2}$	$+3.7_{\pm0.2}$	$-0.6_{\pm0.3}$	$-0.3_{\pm0.0}$	$-0.4_{\pm0.0}$	$+1.0_{\pm0.1}$
		Winkler_{S-} ↓	$159.1_{\pm3.5}$	$141.7_{\pm3.5}$	$146.6_{\pm3.9}$	$143.6_{\pm3.2}$	$\mathbf{118.2_{\pm2.7}}$	$137.9_{\pm3.6}$	$136.8_{\pm3.6}$	$125.0_{\pm3.6}$
LA	Tr	ΔCov	$-2.3_{\pm0.1}$	$+0.7_{\pm0.0}$	$-0.1_{\pm0.6}$	$-2.3_{\pm0.0}$	$-0.6_{\pm0.1}$	$-0.3_{\pm0.0}$	$-0.5_{\pm0.1}$	$-0.3_{\pm0.1}$
		Winkler↓	$32.3_{\pm0.4}$	$29.5_{\pm0.3}$	$27.1_{\pm1.1}$	$24.1_{\pm0.3}$	$21.9_{\pm0.2}$	$23.5_{\pm0.2}$	$31.3_{\pm0.3}$	$\mathbf{20.2_{\pm0.3}}$
		ΔCov_{S+}	$-6.0_{\pm0.4}$	$+0.6_{\pm0.1}$	$-1.4_{\pm0.7}$	$-2.5_{\pm0.1}$	$-1.0_{\pm0.2}$	$-0.4_{\pm0.0}$	$-1.4_{\pm0.3}$	$-0.3_{\pm0.1}$
		Winkler_{S+} ↓	$41.7_{\pm1.2}$	$36.2_{\pm1.0}$	$33.6_{\pm1.8}$	$27.2_{\pm0.6}$	$25.5_{\pm0.7}$	$27.0_{\pm0.7}$	$44.5_{\pm2.8}$	$\mathbf{22.9_{\pm0.7}}$
		ΔCov_{S-}	$+0.5_{\pm0.3}$	$+0.7_{\pm0.1}$	$+1.0_{\pm0.8}$	$-2.0_{\pm0.1}$	$+0.2_{\pm0.2}$	$-0.3_{\pm0.0}$	$-0.6_{\pm0.4}$	$-0.2_{\pm0.1}$
		Winkler_{S-} ↓	$25.9_{\pm0.8}$	$24.1_{\pm0.7}$	$22.0_{\pm0.9}$	$20.7_{\pm0.5}$	$19.0_{\pm0.5}$	$20.0_{\pm0.5}$	$26.6_{\pm1.6}$	$\mathbf{17.8_{\pm0.4}}$
	RNN	ΔCov	$-2.4_{\pm0.1}$	$+0.7_{\pm0.0}$	$-0.2_{\pm0.5}$	$-2.3_{\pm0.0}$	$-0.6_{\pm0.1}$	$-0.3_{\pm0.0}$	$-0.4_{\pm0.1}$	$-0.3_{\pm0.1}$
		Winkler↓	$32.3_{\pm0.2}$	$29.5_{\pm0.2}$	$26.9_{\pm1.2}$	$23.9_{\pm0.1}$	$21.9_{\pm0.2}$	$23.5_{\pm0.1}$	$31.3_{\pm0.3}$	$\mathbf{20.2_{\pm0.1}}$
		ΔCov_{S+}	$-6.0_{\pm0.5}$	$+0.7_{\pm0.1}$	$-1.3_{\pm0.5}$	$-2.5_{\pm0.1}$	$-1.0_{\pm0.2}$	$-0.4_{\pm0.0}$	$-1.2_{\pm0.5}$	$-0.4_{\pm0.1}$
		Winkler_{S+} ↓	$42.3_{\pm1.1}$	$36.8_{\pm0.8}$	$33.4_{\pm1.9}$	$27.2_{\pm0.5}$	$26.1_{\pm0.7}$	$27.5_{\pm0.7}$	$46.1_{\pm3.0}$	$\mathbf{23.4_{\pm0.6}}$
		ΔCov_{S-}	$+0.2_{\pm0.4}$	$+0.7_{\pm0.0}$	$+0.7_{\pm0.7}$	$-2.1_{\pm0.1}$	$-0.1_{\pm0.2}$	$-0.3_{\pm0.0}$	$-0.2_{\pm0.1}$	$-0.2_{\pm0.1}$
		Winkler_{S-} ↓	$26.0_{\pm0.9}$	$24.3_{\pm0.8}$	$22.1_{\pm1.2}$	$20.8_{\pm0.6}$	$19.3_{\pm0.5}$	$20.1_{\pm0.6}$	$24.4_{\pm1.6}$	$\mathbf{18.0_{\pm0.5}}$
BAY	Tr	ΔCov	$-0.6_{\pm0.1}$	$+0.7_{\pm0.0}$	$+0.4_{\pm0.3}$	$+2.0_{\pm0.1}$	$-0.3_{\pm0.1}$	$-0.4_{\pm0.0}$	$-2.1_{\pm0.1}$	$-0.5_{\pm0.1}$
		Winkler↓	$17.1_{\pm0.1}$	$14.9_{\pm0.1}$	$11.7_{\pm0.8}$	$10.9_{\pm0.1}$	$10.8_{\pm0.1}$	$10.2_{\pm0.1}$	$30.5_{\pm1.1}$	$\mathbf{9.6_{\pm0.1}}$
		ΔCov_{S+}	$-2.9_{\pm0.3}$	$+0.7_{\pm0.1}$	$-0.6_{\pm0.4}$	$+0.5_{\pm0.3}$	$-4.3_{\pm0.2}$	$-0.3_{\pm0.0}$	$-1.5_{\pm0.3}$	$-1.2_{\pm0.1}$
		Winkler_{S+} ↓	$24.6_{\pm0.2}$	$20.5_{\pm0.1}$	$16.8_{\pm1.4}$	$15.3_{\pm0.2}$	$15.6_{\pm0.2}$	$\mathbf{14.0_{\pm0.2}}$	$34.0_{\pm3.3}$	$13.5_{\pm0.2}$
		ΔCov_{S-}	$+1.3_{\pm0.2}$	$+0.7_{\pm0.0}$	$+1.3_{\pm0.4}$	$+3.0_{\pm0.2}$	$+1.0_{\pm0.1}$	$-0.4_{\pm0.0}$	$-0.8_{\pm0.4}$	$-0.2_{\pm0.1}$
		Winkler_{S-} ↓	$18.6_{\pm0.5}$	$14.8_{\pm0.4}$	$11.5_{\pm0.8}$	$10.7_{\pm0.3}$	$10.2_{\pm0.2}$	$9.7_{\pm0.2}$	$26.6_{\pm2.3}$	$\mathbf{9.0_{\pm0.2}}$
	RNN	ΔCov	$-0.7_{\pm0.1}$	$+0.7_{\pm0.0}$	$+0.5_{\pm0.6}$	$+2.1_{\pm0.1}$	$-0.3_{\pm0.1}$	$-0.4_{\pm0.0}$	$-2.2_{\pm0.1}$	$-0.4_{\pm0.1}$
		Winkler↓	$17.2_{\pm0.1}$	$15.0_{\pm0.1}$	$11.8_{\pm0.7}$	$11.1_{\pm0.1}$	$10.9_{\pm0.0}$	$10.2_{\pm0.0}$	$31.8_{\pm0.6}$	$\mathbf{9.6_{\pm0.0}}$
		ΔCov_{S+}	$-3.0_{\pm0.3}$	$+0.7_{\pm0.0}$	$-0.4_{\pm0.6}$	$+0.4_{\pm0.2}$	$-4.5_{\pm0.3}$	$-0.3_{\pm0.0}$	$-1.5_{\pm0.3}$	$-1.2_{\pm0.1}$
		Winkler_{S+} ↓	$24.9_{\pm0.3}$	$20.6_{\pm0.2}$	$16.8_{\pm1.1}$	$15.7_{\pm0.3}$	$15.6_{\pm0.1}$	$\mathbf{14.1_{\pm0.1}}$	$34.7_{\pm1.9}$	$13.6_{\pm0.1}$
		ΔCov_{S-}	$+1.3_{\pm0.2}$	$+0.7_{\pm0.0}$	$+1.6_{\pm0.6}$	$+3.2_{\pm0.2}$	$+1.0_{\pm0.2}$	$-0.4_{\pm0.0}$	$-0.6_{\pm0.4}$	$-0.2_{\pm0.1}$
		Winkler_{S-} ↓	$19.5_{\pm0.7}$	$15.2_{\pm0.3}$	$11.9_{\pm0.7}$	$11.1_{\pm0.2}$	$10.4_{\pm0.2}$	$10.0_{\pm0.2}$	$29.2_{\pm2.1}$	$\mathbf{9.2_{\pm0.1}}$
SD	Tr	ΔCov	$-2.4_{\pm0.3}$	$+0.8_{\pm0.0}$	$-0.7_{\pm0.2}$	$-2.1_{\pm0.2}$	$-1.7_{\pm0.5}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.6_{\pm0.1}$
		Winkler↓	$174.9_{\pm2.8}$	$164.1_{\pm2.8}$	$139.5_{\pm2.6}$	$146.4_{\pm2.5}$	$143.2_{\pm2.5}$	$149.0_{\pm2.8}$	$151.7_{\pm2.5}$	$\mathbf{127.5_{\pm2.6}}$
		ΔCov_{S+}	$-4.7_{\pm0.5}$	$+0.9_{\pm0.1}$	$-1.0_{\pm0.2}$	$-3.3_{\pm0.2}$	$-4.0_{\pm0.5}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.7_{\pm0.1}$
		Winkler_{S+} ↓	$220.6_{\pm5.8}$	$194.6_{\pm5.9}$	$162.6_{\pm4.5}$	$177.2_{\pm4.9}$	$174.2_{\pm4.4}$	$171.6_{\pm5.5}$	$175.5_{\pm5.0}$	$\mathbf{148.0_{\pm4.1}}$
		ΔCov_{S-}	$-2.0_{\pm0.5}$	$+0.9_{\pm0.0}$	$-1.2_{\pm0.2}$	$-2.7_{\pm0.4}$	$-1.3_{\pm0.4}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.5_{\pm0.1}$
		Winkler_{S-} ↓	$180.4_{\pm8.2}$	$169.1_{\pm7.2}$	$160.4_{\pm6.4}$	$152.2_{\pm6.8}$	$146.2_{\pm6.9}$	$152.9_{\pm6.6}$	$156.0_{\pm6.8}$	$\mathbf{130.5_{\pm5.7}}$
	RNN	ΔCov	$-2.5_{\pm0.3}$	$+0.8_{\pm0.0}$	$-0.9_{\pm0.2}$	$-2.0_{\pm0.3}$	$-1.6_{\pm0.4}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.6_{\pm0.1}$
		Winkler↓	$177.4_{\pm2.9}$	$166.2_{\pm3.3}$	$142.0_{\pm3.3}$	$145.9_{\pm1.6}$	$143.7_{\pm2.5}$	$150.5_{\pm3.1}$	$153.3_{\pm2.9}$	$\mathbf{128.8_{\pm2.0}}$
		ΔCov_{S+}	$-4.9_{\pm0.3}$	$+0.9_{\pm0.0}$	$-1.3_{\pm0.3}$	$-3.2_{\pm0.4}$	$-4.1_{\pm0.4}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.7_{\pm0.1}$
		Winkler_{S+} ↓	$227.1_{\pm5.5}$	$200.5_{\pm6.0}$	$168.3_{\pm5.6}$	$179.3_{\pm4.4}$	$177.1_{\pm4.0}$	$176.2_{\pm5.4}$	$180.5_{\pm5.1}$	$\mathbf{152.2_{\pm4.0}}$
		ΔCov_{S-}	$-2.1_{\pm0.4}$	$+0.9_{\pm0.0}$	$-1.4_{\pm0.2}$	$-2.6_{\pm0.3}$	$-1.5_{\pm0.5}$	$-0.4_{\pm0.0}$	$-0.2_{\pm0.0}$	$-0.6_{\pm0.2}$
		Winkler_{S-} ↓	$181.6_{\pm6.4}$	$170.0_{\pm6.4}$	$163.7_{\pm7.8}$	$151.8_{\pm5.2}$	$147.0_{\pm6.2}$	$154.3_{\pm5.7}$	$157.0_{\pm5.8}$	$\mathbf{132.1_{\pm4.6}}$

does not certify that a tracker reached its grid boundary, and a value below one does not certify that a component is calibrated. Like S^+, χ is computed from realised test residuals and is used only as an ex-post descriptive stratification, never for model selection or hyperparameter tuning.

Results across strata. Tab. 4 and Tab. 5 give the full five-stratum breakdown for both backbones. The two backbones agree closely: across the 100 dataset–stratum–method cells the coverage deviations differ by a median of 0.03 pp and by more than 0.3 points in only five cells, and the Winkler scores differ by a median of 0.7%. The conclusions below therefore hold for both and we quote the Transformer numbers.

Every method degrades as χ increases, which is expected: the stratification is built from the realised tails of the same windows on which coverage is measured. What separates the methods is the rate. The spread of coverage deviations across the five configurations widens from 2.0, 1.1, 1.9 and 2.0 points in the lowest stratum, on AIR, LA, BAY and SD respectively, to 4.2, 3.2, 2.5 and

7.7 points in the highest. Compared with either single-component ablation, the complete model exhibits a smaller coverage decline from the lowest to the highest stratum on every dataset, moving from $+0.79$, $+0.13$, -0.18 and -0.19 in the lowest stratum to -3.14, -2.50, -2.36 and -2.10 in the highest, while the variant without the recalibrator reaches -5.34, -4.36, -4.76 and -9.81 and the variant without the tracker reaches -6.48, -5.13, -4.81 and -3.99. The two ablations fail in different places: removing the recalibrator is far more damaging on SD, removing the tracker is more damaging on AIR, and on LA and BAY the two are comparable.

Interval scores. The Winkler entries show that the components pay for coverage differently. Adding the tracker lowers the score in every highest-stratum comparison, so it improves both criteria at once. Adding the recalibrator instead widens the intervals, by 10.0%, 11.0%, 14.3% and 29.2% on AIR, LA, BAY and SD in the highest stratum, and that cost is offset by a fall in the miss penalty of 5.8%, 12.0%, 14.2% and 42.7%. The net effect on the Winkler score is therefore below 1% on AIR, LA and BAY and a 7.8% reduction on SD. A near-constant Winkler score across a recalibrator ablation should not be read as a near-constant interval width.

Qualifications. Three entries qualify the picture. First, the complete model still undercovers by 2.1 to 3.3 points in the highest stratum on every dataset, so the two components mitigate rather than remove the limitation. Second, the reference baselines are not dominated: in LA's highest stratum ACI is closer to nominal coverage than the complete model (-1.94 against -2.50), and in BAY's it attains a lower Winkler score (23.9 against 24.8). Third, in AIR's lowest stratum the complete model overcovers by 0.79 points while the variant without the recalibrator sits at $+0.09$, so there the additional coverage increases rather than decreases the absolute deviation; a signed coverage gain is not automatically a calibration gain.

Window length. Tab. 6 repeats the highest-minus-lowest comparison at $W = 3P$, $6P$ and $12P$. All 48 entries are positive, so the larger contribution of both components under higher local pressure is not an artefact of one window length. The exception is the sample support on AIR, whose test segment is the shortest: its highest stratum holds 63 to 79 units per seed at $W = 3P$ but only 7 to 13 at $6P$ and 0 to 2 at $12P$, with one Transformer seed empty at $12P$. The corresponding AIR entries for the tracker ablation, $+0.47 \pm 0.54$ and $+0.15 \pm 0.67$, have standard deviations that cross zero. We therefore read the window-length check as supported on LA, BAY and SD, and as uninformative on AIR beyond $W = 3P$.

B.3.2 Ablation study on S^+ slice

Fig. 3 extends the component ablation to all four real-world datasets. The pattern is consistent across datasets: both the tracker-only and recalibrator-only variants fall outside the prescribed coverage band, whereas the full model moves coverage substantially closer to the target and obtains the lowest Winkler score in every panel. The comparison also shows that the two components provide complementary improvements. Adding the quantile-estimate recalibrator to the tracker-only variant improves ΔCov from -2.0% to $+0.4\%$ on AIR, from -1.8% to -0.3% on LA, from -4.3% to -1.2% on BAY, and from -4.4% to -0.7% on SD. Conversely, adding the tracker to the recalibrator-only variant improves the corresponding deviations from -1.5%, -1.6%, -3.9%, and -3.9%. In every dataset, the full model also achieves a lower Winkler score than either single-component variant. The full model enters the prescribed coverage band on AIR, LA, and SD. BAY remains the only borderline case: the full model reaches $\Delta Cov = -1.2\%$, compared with -4.3% for tracker only and -3.9% for calibrator only, while reducing the Winkler score to 13.5 from 15.8 and 15.6, respectively. Thus, even where the full model does not completely close the coverage gap, both mechanisms make substantial and complementary contributions.

B.3.3 Saturation Signal Analysis

We use two experiments to examine whether saturation behaves as intended. The first uses synthetic data, where the true conditional residual quantiles are known, to isolate the effect of insufficient tail reach. The second uses BAY to test whether changing only the anchor gain produces the intended response in the real-data pipeline.

Table 4: Component ablation by local tail pressure χ (Transformer backbone). Each cell is ΔCov (pp) / Winkler score. *Cells* is the per-seed number of component–window units in the stratum, reported as a range over seeds.

		<0.6	0.6–0.8	0.8–1	1–1.2	≥ 1.2
AIR	Cells	5679–5743	562–625	179–221	63–73	63–72
	CoRel	-1.25 / 78.4	-4.70 / 126.0	-6.08 / 156.2	-8.19 / 203.7	-7.33 / 269.8
	ACI	-0.31 / 89.0	-1.53 / 135.3	-2.25 / 163.1	-2.86 / 206.2	-3.33 / 261.1
	BaCR w/o tracker	-0.49 / 79.3	-2.61 / 122.1	-4.00 / 153.3	-5.24 / 198.4	-6.48 / 276.3
	BaCR w/o recalibrator	$+0.09$ / 79.8	-2.17 / 118.7	-3.42 / 146.1	-4.87 / 188.3	-5.34 / 252.0
	BaCR	$+0.79$ / 80.4	-0.36 / 120.3	-1.37 / 147.7	-1.89 / 189.0	-3.14 / 251.8
LA	Cells	6816–6919	1831–1965	1815–1894	849–908	454–513
	CoRel	$+0.23$ / 14.2	-0.55 / 24.3	-1.40 / 32.7	-2.56 / 39.0	-4.31 / 44.5
	ACI	$+0.47$ / 15.6	-0.86 / 27.0	-1.53 / 36.0	-1.75 / 40.7	-1.94 / 40.6
	BaCR w/o tracker	-0.52 / 14.3	-1.47 / 24.4	-2.21 / 32.7	-3.23 / 39.1	-5.13 / 45.2
	BaCR w/o recalibrator	-0.66 / 13.6	-1.05 / 22.6	-1.76 / 29.3	-2.84 / 34.5	-4.36 / 38.5
	BaCR	$+0.13$ / 13.7	-0.34 / 22.8	-0.67 / 29.5	-1.18 / 34.3	-2.50 / 38.2
BAY	Cells	21319–21804	3264–3551	2838–3192	1639–1702	1887–1970
	CoRel	$+0.31$ / 6.7	-0.69 / 14.2	-0.88 / 18.1	-2.01 / 21.6	-4.58 / 29.8
	ACI	$+0.16$ / 6.7	-1.04 / 13.8	-1.61 / 17.2	-1.83 / 19.5	-2.59 / 23.9
	BaCR w/o tracker	-1.61 / 6.9	-2.38 / 14.0	-2.40 / 17.6	-2.96 / 20.8	-4.81 / 29.6
	BaCR w/o recalibrator	-1.33 / 6.1	-1.97 / 12.5	-1.99 / 15.4	-2.64 / 18.0	-4.76 / 25.0
	BaCR	-0.18 / 6.2	-0.68 / 12.6	-0.76 / 15.5	-1.11 / 18.1	-2.36 / 24.8
SD	Cells	13813–15677	16094–16903	8265–9285	2493–2704	2149–2637
	CoRel	-0.27 / 108.2	-1.17 / 136.6	-2.60 / 161.4	-5.22 / 209.1	-7.38 / 273.8
	ACI	$+0.37$ / 116.9	-0.37 / 146.1	-0.93 / 167.2	-1.59 / 205.8	-2.24 / 245.5
	BaCR w/o tracker	-0.37 / 105.7	-0.78 / 130.3	-1.49 / 153.7	-2.69 / 204.0	-3.99 / 272.7
	BaCR w/o recalibrator	-1.62 / 98.3	-2.08 / 122.4	-3.28 / 142.7	-6.16 / 184.9	-9.81 / 248.7
	BaCR	-0.19 / 99.1	-0.47 / 122.9	-0.85 / 142.0	-1.45 / 179.9	-2.10 / 229.3

Table 5: Component ablation by local tail pressure χ (RNN backbone, $W = 3P$, 10 seeds). Each cell is ΔCov (pp) / Winkler score. *Cells* is the per-seed number of component–window units in the stratum, reported as a range over seeds.

		<0.6	0.6–0.8	0.8–1	1–1.2	≥ 1.2
AIR	Cells	5674–5725	584–621	181–205	61–81	68–79
	CoRel	-1.24 / 78.4	-4.62 / 123.8	-6.10 / 158.6	-7.69 / 185.2	-7.18 / 268.0
	ACI	-0.31 / 89.3	-1.50 / 133.1	-2.23 / 166.0	-2.69 / 185.8	-3.30 / 257.5
	BaCR w/o tracker	-0.66 / 79.7	-2.60 / 119.9	-4.12 / 155.0	-5.02 / 178.2	-6.57 / 276.3
	BaCR w/o recalibrator	$+0.01$ / 80.0	-2.14 / 116.4	-3.43 / 147.5	-4.51 / 168.5	-5.44 / 250.8
	BaCR	$+0.79$ / 80.9	-0.32 / 118.2	-1.33 / 149.3	-1.60 / 169.3	-3.34 / 251.2
LA	Cells	6792–6865	1859–1941	1833–1952	872–922	468–521
	CoRel	$+0.16$ / 14.2	-0.56 / 24.4	-1.35 / 32.6	-2.53 / 39.4	-4.31 / 44.4
	ACI	$+0.47$ / 15.4	-0.86 / 27.0	-1.51 / 35.7	-1.83 / 40.9	-1.92 / 40.5
	BaCR w/o tracker	-0.60 / 14.2	-1.49 / 24.4	-2.21 / 32.6	-3.45 / 39.6	-4.98 / 45.2
	BaCR w/o recalibrator	-0.67 / 13.5	-1.04 / 22.6	-1.70 / 29.2	-2.87 / 34.7	-4.42 / 38.4
	BaCR	$+0.12$ / 13.6	-0.32 / 22.8	-0.68 / 29.3	-1.29 / 34.6	-2.37 / 38.1
BAY	Cells	21470–21708	3267–3466	2869–3066	1663–1736	1901–1957
	CoRel	$+0.31$ / 6.7	-0.69 / 14.2	-0.84 / 18.0	-1.87 / 21.6	-4.69 / 30.1
	ACI	$+0.17$ / 6.7	-1.06 / 14.0	-1.62 / 17.2	-1.88 / 19.5	-2.58 / 24.2
	BaCR w/o tracker	-1.49 / 6.9	-2.27 / 14.1	-2.35 / 17.4	-2.95 / 20.7	-4.74 / 29.8
	BaCR w/o recalibrator	-1.30 / 6.2	-1.96 / 12.5	-1.99 / 15.3	-2.62 / 17.9	-4.83 / 25.2
	BaCR	-0.14 / 6.2	-0.66 / 12.6	-0.75 / 15.4	-1.10 / 18.0	-2.32 / 24.9
SD	Cells	14715–16175	15644–16513	8000–8675	2458–2731	2275–2576
	CoRel	-0.24 / 108.8	-1.10 / 137.5	-2.62 / 164.2	-4.87 / 209.8	-6.84 / 271.6
	ACI	$+0.40$ / 117.7	-0.40 / 147.8	-0.99 / 172.2	-1.63 / 208.8	-2.41 / 247.7
	BaCR w/o tracker	-0.34 / 106.7	-0.81 / 132.0	-1.51 / 157.9	-2.66 / 206.7	-3.93 / 273.1
	BaCR w/o recalibrator	-1.57 / 99.4	-2.08 / 123.9	-3.33 / 146.3	-5.92 / 186.7	-9.24 / 246.4
	BaCR	-0.21 / 100.0	-0.52 / 124.4	-0.88 / 145.9	-1.45 / 182.6	-2.08 / 230.6

For the synthetic diagnostic, we run the quantile-level tracker with the recalibrator disabled. Exact conditional quantiles provide a well-specified reference. We then deliberately narrow the oracle quantile geometry while leaving the residual stream unchanged, creating a controlled shortage of value-space tail reach.

Tab. 7 shows a clear pattern. Even with exact quantiles, binary feedback produces temporary tracker displacements and therefore a nonzero saturation baseline. Nevertheless, boundary clipping remains

Table 6: Window-length sensitivity. Each entry is the coverage advantage of the complete model over the named ablation in the highest χ stratum minus the same quantity in the lowest, in percentage points (10 seeds, mean±std). AIR at $W = 6P$ and $W = 12P$ has too few units in the highest stratum to support the comparison (7–13 and 0–2 per seed; one Transformer seed is empty).

		$W = 3P$	$W = 6P$	$W = 12P$
Transformer	w/o tracker (AIR)	$+2.07 \pm 0.22$	$+0.70 \pm 0.39$	$+0.47 \pm 0.54$
	w/o recalibrator (AIR)	$+1.50 \pm 0.28$	$+4.15 \pm 0.58$	$+1.88 \pm 0.30$
	w/o tracker (LA)	$+1.97 \pm 0.17$	$+1.24 \pm 0.15$	$+1.22 \pm 0.23$
	w/o recalibrator (LA)	$+1.08 \pm 0.20$	$+1.97 \pm 0.42$	$+3.73 \pm 0.38$
	w/o tracker (BAY)	$+1.02 \pm 0.18$	$+1.04 \pm 0.13$	$+1.75 \pm 0.25$
	w/o recalibrator (BAY)	$+1.24 \pm 0.15$	$+2.22 \pm 0.28$	$+3.54 \pm 0.43$
	w/o tracker (SD)	$+1.72 \pm 0.23$	$+1.19 \pm 0.27$	$+1.26 \pm 0.23$
	w/o recalibrator (SD)	$+6.29 \pm 1.40$	$+7.09 \pm 1.73$	$+5.80 \pm 1.43$
RNN	w/o tracker (AIR)	$+1.78 \pm 0.29$	$+0.50 \pm 0.48$	$+0.15 \pm 0.67$
	w/o recalibrator (AIR)	$+1.32 \pm 0.19$	$+3.83 \pm 0.52$	$+2.05 \pm 0.73$
	w/o tracker (LA)	$+1.90 \pm 0.22$	$+1.22 \pm 0.15$	$+1.16 \pm 0.24$
	w/o recalibrator (LA)	$+1.27 \pm 0.17$	$+2.20 \pm 0.26$	$+3.78 \pm 0.33$
	w/o tracker (BAY)	$+1.07 \pm 0.18$	$+1.05 \pm 0.18$	$+2.06 \pm 0.42$
	w/o recalibrator (BAY)	$+1.34 \pm 0.15$	$+2.31 \pm 0.27$	$+3.75 \pm 0.46$
	w/o tracker (SD)	$+1.72 \pm 0.21$	$+1.19 \pm 0.26$	$+1.17 \pm 0.26$
	w/o recalibrator (SD)	$+5.79 \pm 0.75$	$+6.46 \pm 0.99$	$+5.26 \pm 0.87$

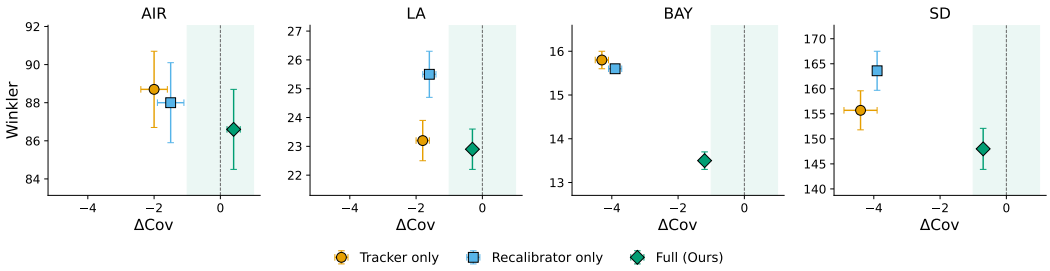


Figure 3: Component ablation on the S^+ slice (Transformer, 10 seeds, mean±std). x is ΔCov (shaded = $|\Delta\text{Cov}| \leq 1$), y is the Winkler score; removing either the recalibrator or the trackers drops coverage out of the band.

rare and coverage stays close to the target. As the quantile estimates is narrowed, the tracker spends more time near its grid boundaries, widening clips become more frequent, and coverage deteriorates. In the most severely narrowed setting, the widening clip term accounts for nearly all of the observed coverage deficit, while the decomposition in Theorem 1 holds to numerical precision. This supports the intended distinction: realized clip mass is the direct ex-post coverage certificate, whereas saturation is the operational signal indicating that the available quantile geometry is under strain.

For the real-data diagnostic, we run the BAY dataset three times while changing only κ. Setting $\kappa = 0$ removes the saturation-dependent anchor but leaves the pinball-based recalibration active. The same high-tail-excess component and the same top-decile tail-excess slice are used in all three runs. Fig. 4 shows a consistent directional response. A stronger anchor gain produces greater value-space expansion, reduces saturation, and improves coverage on the fixed high-tail-excess slice. The improvement is modest and is accompanied by slight interval widening. Thus, the pinball refit performs the primary adaptation, while the saturation anchor provides an additional targeted correction when the tracker persistently uses its widening headroom. As this is a single-seed diagnostic, it is intended to illustrate the mechanism rather than serve as an additional performance comparison.

B.4 Sensitivity Analysis

We examine the key online-calibration hyperparameters of BaCR around the common default configuration $(\eta, P, B, \lambda, \kappa) = (0.03, 36, 128, 1, 0.5)$. We sweep $\eta \in \{0.01, 0.03, 0.05, 0.1, 0.5\}$, $P \in \{12, 24, 36, 72\}$, $B \in \{32, 64, 128, 256\}$, $\lambda \in \{0, 1, 2\}$, $\kappa \in \{0, 0.25, 0.5, 1, 2\}$, and target coverage in $\{0.80, 0.90, 0.95\}$. All experiments use the Transformer backbone on the four real datasets and reuse the same frozen point forecaster and residual quantile model as the main exper-

Table 7: Controlled diagnostics using the exact conditional residual quantiles of the synthetic generator (seed 0). Tail-reach deficiencies are introduced by multiplying the oracle quantile geometry by the reported factor. The adverse clip term is $100(1-\alpha/2)\bar{\pi}^{\mathrm{wid}}$ in percentage points.

Quantile geometry	Mean saturation	90th pct.	Coverage	$\bar{\pi}^{\mathrm{wid}}$	Adverse clip term
Exact	0.359	0.658	89.60%	0.0057	0.54 pp
0.7× oracle	0.709	0.848	86.16%	0.038	3.61 pp
0.5× oracle	0.857	0.918	79.52%	0.109	10.36 pp

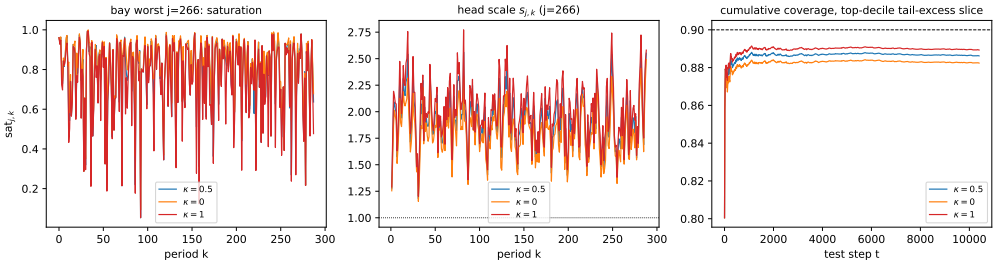


Figure 4: Paired saturation-gain diagnostic on BAY using the Transformer backbone. The first two panels show period-wise saturation and affine scale for the component with the largest frozen-geometry tail excess ($j = 266$); the third shows cumulative coverage on the fixed top-decile tail-excess slice. Only κ changes across the three runs.

iments; only the online BaCR stage is rerun. We vary one parameter at a time, except for the full $\eta \times P$ cross, and report both coverage deviation and Winkler score over all components and the S^+ slice. Results are averaged over the same ten seeds as the main experiments.

Step size η and period P. Fig. 5 and Fig. 6 reveal a broad coverage-stable region. For $\eta \in \{0.01, 0.03\}$, coverage changes only moderately across the six-fold range $P \in [12, 72]$. Larger learning rates produce increasingly negative coverage deviations, whereas the effect of P remains comparatively small at a fixed η. The Winkler panels Fig. 7 and Fig. 8 show the corresponding trade-off. Increasing η from 0.03 to 0.1 reduces the Winkler score by 2.4%–5.6% at $P = 36$, but costs approximately 3 pp of coverage. Conversely, $\eta = 0.01$ produces more conservative coverage with a 2.4%–5.4% increase in Winkler score. The default $\eta = 0.03$ lies near the coverage-acceptable end of this empirical frontier. At $\eta = 0.5$, coverage decreases by approximately 12–13 pp and depends only weakly on P. This suggests that the learning rate of tracker, rather than the recalibration frequency, dominates behavior in this regime.

Recalibration buffer size $|\mathcal{B}|$. Fig. 9 examines how the recalibration set size B could influence the results. Across $B \in \{32, 64, 128, 256\}$, Fig. 9 shows that marginal coverage deviation varies by at most 0.5 pp on every dataset, while the variation on the S^+ slice is at most 0.6 pp. Increasing B produces a small upward coverage adjustment on all four datasets without materially changing the overall coverage–efficiency trade-off. The default $B = 128$ lies in the middle of the tested range and is used unchanged across datasets. These results indicate that BaCR is not sensitive to the recalibration-buffer size over the tested range.

Anchor strength λ. Fig. 10 examines the strength λ of the quadratic saturation-anchor penalty. Setting $\lambda = 0$ removes this penalty while retaining the pinball-based refitting of the affine recalibrator; increasing λ places greater weight on pulling the component-specific log-scale toward the saturation-dependent target $\kappa\,\mathrm{sat}_{j,k}$. Across $\lambda \in \{0, 1, 2\}$, marginal coverage remains within 1 pp of the target on every dataset. Increasing λ produces a modest upward coverage adjustment on LA and BAY, moves AIR slightly further into overcoverage, and has almost no effect on SD. The S^+ results exhibit the same general pattern: the larger anchor strength reduces undercoverage on LA and BAY, although BAY remains slightly outside the empirical coverage-acceptability band. Thus, the empirical coverage of BaCR is not highly sensitive to λ over the tested range, while the direc-

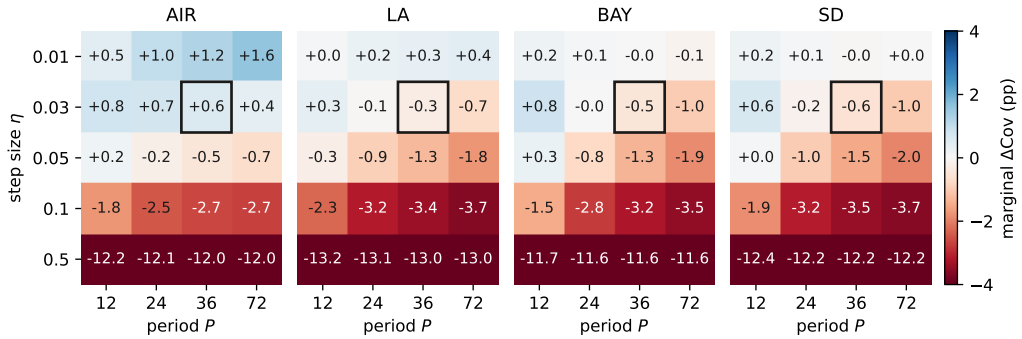


Figure 5: Marginal ΔCov (points, target 90%) over the $\eta \times P$ cross (Transformer, 10 seeds), one panel per dataset. Cell text gives exact values; the color scale diverges around 0 and is clipped at ± 4; the boxed cell is the default configuration. Coverage is stable over $\eta \in \{0.01, 0.03\}$ and the full tested range of P.

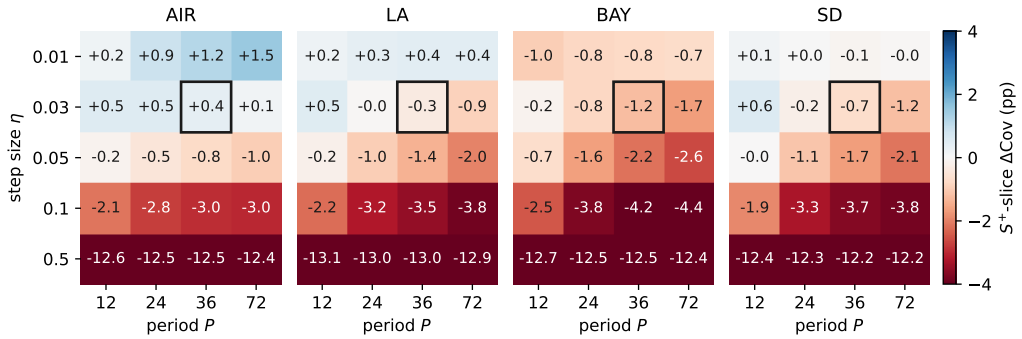


Figure 6: As Figure 5, for the S^+ shift slice. The plateau and the $\eta = 0.5$ collapse have the same geometry as the marginal view; within the plateau the S^+ cells sit mildly below their marginal counterparts (by up to 1.2%, BAY the most), consistent with the residual S^+ gaps of the main results.

tion of the adjustment depends on the component population's initial coverage error. We use the intermediate value $\lambda = 1$ as the common default across datasets.

Anchor gain κ. Fig. 11 examines the gain κ that maps the observed saturation statistic to the component-specific anchor target $\log s_{j,k}^{\text{anchor}} = \kappa \, \text{sat}_{j,k}$. When $\kappa = 0$, the affine recalibrator is still refit using the pinball loss, while the anchor penalty is centered at the identity scale $\log s_j = 0$. Increasing κ therefore strengthens the saturation-dependent expansion target rather than changing the overall penalty weight. Across $\kappa \in \{0, 0.25, 0.5, 1, 2\}$, Fig. 11 shows a consistent upward coverage adjustment as κ increases. The effect is most pronounced on the S^+ slices of LA and BAY, whose coverage increases by approximately 1.2 and 1.1 pp, respectively, between $\kappa = 0$ and $\kappa = 2$. LA crosses the nominal target, while BAY moves into the empirical coverage-acceptability band. AIR, which is already slightly overcovered at $\kappa = 0$, becomes more conservative, whereas SD changes only marginally. These coverage trends are consistent with the intended directional role of the saturation coupling: increasing κ produces a stronger upward coverage adjustment on the components for which additional expansion is needed, but can also increase overcoverage when the intervals are already sufficiently wide. We use the intermediate value $\kappa = 0.5$ as the common default across datasets.

Target level. Fig. 12 examines BaCR at target coverage levels $1 - \alpha \in \{0.80, 0.90, 0.95\}$, with ΔCov measured relative to the corresponding target and all other hyperparameters held fixed. At the 80% target, all four datasets remain at or close to their nominal level: marginal coverage deviation ranges from $+0.06$ pp on SD to $+0.8$ pp on AIR, while the S^+ deviation ranges from -0.1 to $+0.7$

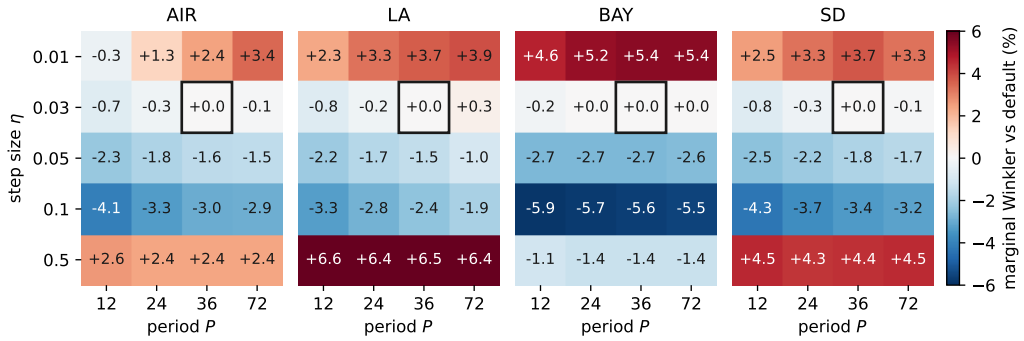


Figure 7: Efficiency companion to Fig. 5: marginal Winkler score over the $\eta \times P$ cross, reported as the percentage change relative to each dataset's default cell (Transformer, 10 seeds; red denotes a higher, worse score). Moderate increases in η lower the interval score but lead to undercoverage. At $\eta = 0.5$, severe miscoverage reverses this trend on AIR, LA, and SD.

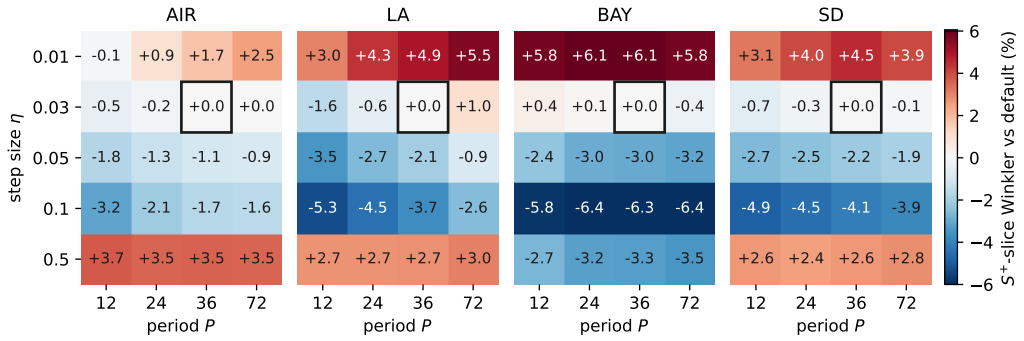


Figure 8: As Fig. 7, for the S^+ slice (% change relative to each dataset's default-cell S^+ Winkler score).

pp. The 90% target reproduces the coverage pattern of the main experiments. At the more extreme 95% target, all four datasets exhibit undercoverage, with marginal deviations between -1.5 and -2.1 pp and S^+ deviations between -1.8 and -3.0 pp; BAY exhibits the largest S^+ deficit. This deterioration is consistent with the reduced outward level-space headroom at more extreme target levels. On the fixed quantile grid, the widening budget

$$\delta_{\max} = \min\{\tau_{\text{lo}} - q_{\min}, q_{\max} - \tau_{\text{hi}}\}$$

decreases from 0.09 at the 80% target to 0.04 at 90% and 0.015 at 95%. Consequently, the tracker reaches its outer boundaries after a smaller outward displacement, making the boundary-limited regime more likely. The target-level sweep therefore shows that BaCR remains stable at moderate coverage levels, while very high target coverage remains more challenging under a fixed finite quantile grid.

Overall, the default configuration lies in a broad region of stable marginal coverage, while the BAY S^+ slice remains close to the boundary of the prescribed coverage criterion. The results identify the step size as the dominant tuning parameter: moderate values produce a stable coverage–Winkler trade-off. Recalibration buffer size and anchor strength have comparatively mild marginal effects, while the saturation gain primarily affects the high-growth components of LA and BAY. Performance degrades at the 95% target as the available quantile-level headroom shrinks. These patterns are qualitatively consistent with the roles of fixed-step fluctuation and boundary clipping in the analysis.

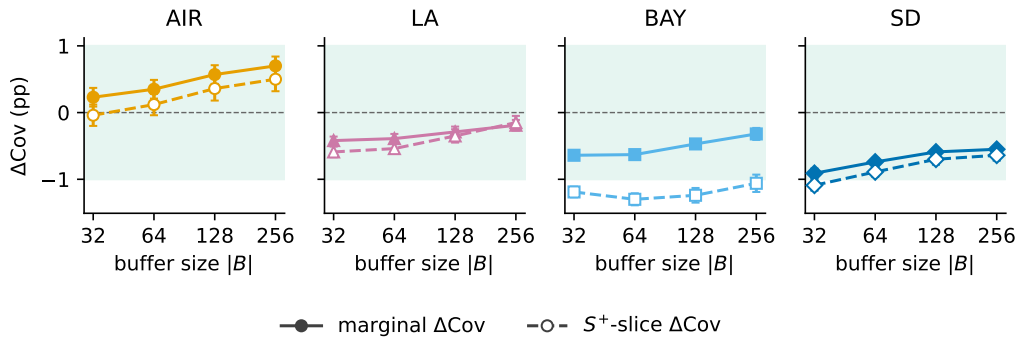


Figure 9: Sensitivity to the rolling recalibration buffer size B (Transformer, 10 seeds): marginal $\Delta\mathrm{Cov}$ (solid, filled markers) and S^+-slice $\Delta\mathrm{Cov}$ (dashed, open markers) for each dataset. Error bars denote the standard deviation across seeds, and the shaded region denotes the empirical coverage-acceptability band $|\Delta\mathrm{Cov}| \leq 1$ pp.

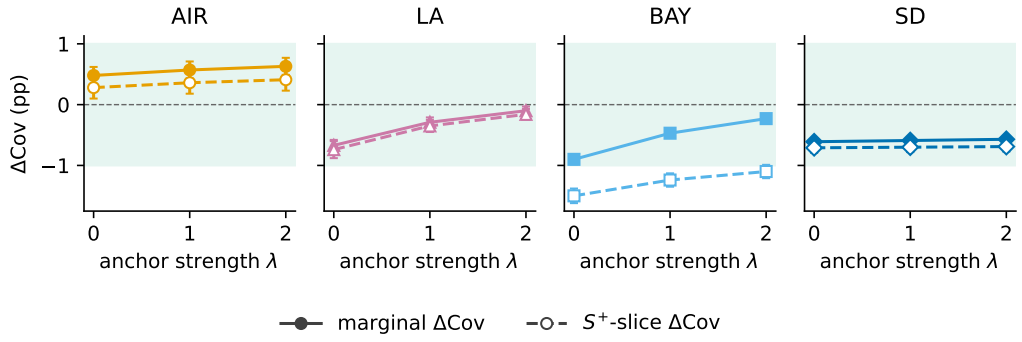


Figure 10: Sensitivity to the anchor strength λ (Transformer, 10 seeds); markers and band as in Fig. 9.

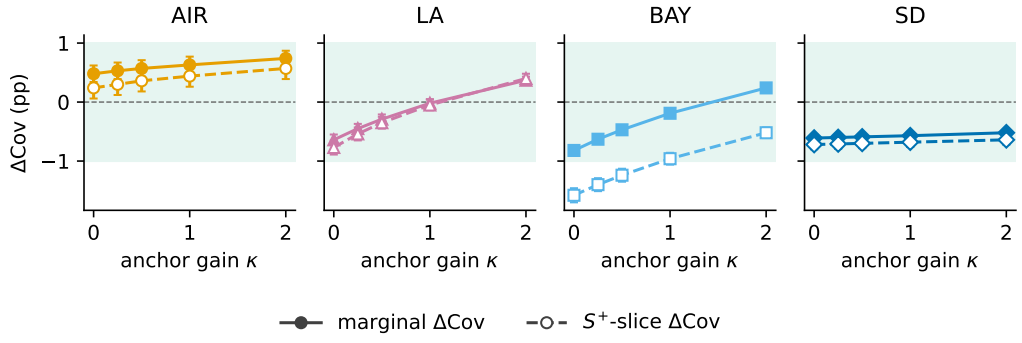


Figure 11: Sensitivity to the anchor gain κ (Transformer, 10 seeds); markers and band as in Fig. 9.

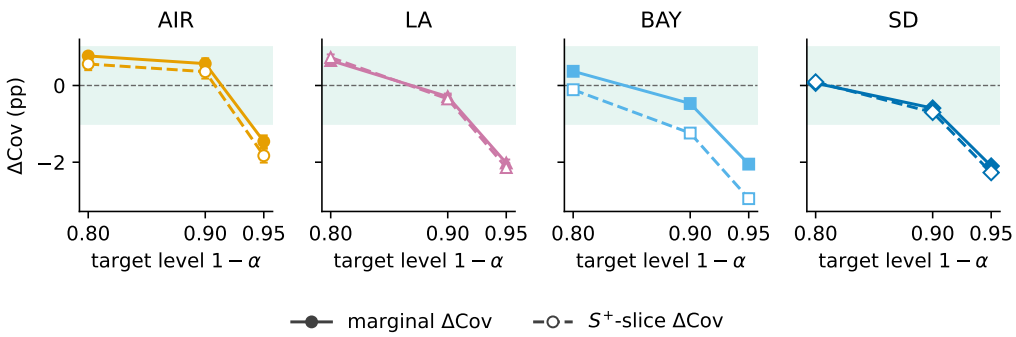


Figure 12: Sensitivity to the target level $1-\alpha$ (Transformer, 10 seeds); markers and band as in Fig. 9. ΔCov is measured against the respective target $(80/90/95\%)$.