Efficient conformal prediction intervals for time series: Online PID-Expert aggregation

Guodong Liu[a], Yanfei Kang[a], Ren Miao[b], Xiaoqian Wang[c,*]

[a]*School of Economics and Management, Beihang University, Beijing 100191, China*
[b]*Beijing Shuji Intelligent Technology Co., Ltd., Beijing 100176, China*
[c]*Center for Forecasting Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

For a given point forecaster, proportional–integral–derivative (PID) calibration configurations can attain similar overall coverage yet produce different interval widths. We introduce PID-Expert, an online aggregation method for improving the efficiency of conformal prediction intervals for time series. PID-Expert combines thresholds from a fixed library of PID calibrators, each evolving under its own coverage feedback. Aggregation weights depend on normalized interval width and miscoverage, while a shared multiplier adapts the miscoverage penalty using feedback from the reported interval. We establish local regret bounds for weighted expert losses under the realized multiplier sequence and, separately, a pathwise upper bound on time-averaged aggregate miscoverage, with control in expectation and almost surely under a stability condition. Across four simulation settings and two real-data applications, PID-Expert produces narrower mean intervals than a prespecified Conformal PID benchmark while keeping overall empirical coverage close to the nominal level. Compared with expert selection and equal averaging, aggregation generally provides a more balanced coverage–width trade-off across forecasting settings. Rolling analyses further reveal local coverage–efficiency trade-offs, particularly following abrupt distributional shifts. Overall, PID-Expert reduces reliance on a single PID configuration while improving interval efficiency.

Keywords: Online conformal prediction, Adaptive calibration, Expert aggregation, Prediction interval efficiency, Time series forecasting

1. Introduction

Online prediction intervals need to keep miscoverage near a target rate while avoiding unnecessary width. This balance between calibration and sharpness is central to probabilistic forecasting (Gneiting et al., 2007; Gneiting and Katzfuss, 2014) and interval forecasting practice (Chatfield, 1993; Hyndman and Athanasopoulos, 2021; Petropoulos et al., 2022).

*Corresponding author
Email addresses: guodongliu@buaa.edu.cn (Guodong Liu), yanfeikang@buaa.edu.cn (Yanfei Kang), miaoren@d-shuji.cn (Ren Miao), xiaoqian.wang@amss.ac.cn (Xiaoqian Wang)

For a given point forecaster, calibration rules with similar overall coverage can still produce intervals of different widths. Online conformal methods adjust calibration through realized coverage errors as new observations arrive (Gibbs and Candès, 2021; Angelopoulos et al., 2023). Even when these adjustments bring overall coverage close to the target, the choice of calibration rule can leave room to improve interval efficiency.

Conformal proportional–integral–derivative (PID) control makes this choice concrete (Angelopoulos et al., 2023). It updates score thresholds through recent coverage errors, accumulated coverage deviations, and forecasts of future nonconformity scores. The proportional gain, integral gain, and saturation constant determine how these feedback signals affect the threshold path. Faster responses can produce more variable intervals, while slower or more conservative responses can require wider intervals. Coverage feedback adjusts an individual controller but does not directly compare its interval efficiency with that of other configurations. Thus, even with a PID calibrator in place, a practical choice remains: which feedback configuration avoids unnecessary width while keeping coverage close to the target?

To assess how much interval efficiency depends on this choice, we compare PID configurations under a common forecasting setup. In the nonlinear simulation in Section 5, 343 PID controllers share the same point forecasts, nonconformity scores, and scorecaster (Figure 4). Their empirical coverage ranges from 0.901 to 0.909 at a target of 0.9, yet their mean interval widths differ. The prespecified Conformal PID controller has mean width 0.360, compared with 0.352 for PID-Expert and 0.348 for the narrowest library member selected retrospectively among those with empirical coverage of at least 0.9. These differences motivate combining PID configurations online, using their observed performance to guide the construction of prediction intervals.

The closest existing approaches address related choices at different stages of forecasting and calibration. Adaptive conformal inference (ACI), introduced by Gibbs and Candès (2021), adjusts the calibration level in response to observed coverage errors. AgACI, as originally proposed by Zaffran et al. (2022), aggregates interval bounds from ACI predictors with different step sizes using Bernstein online aggregation (Wintenberger, 2017), while DtACI (Gibbs and Candès, 2024) adapts their influence through their calibration levels. These methods reduce reliance on a single ACI learning rate. Forecast combinations address uncertainty at the predictive-model stage (Petropoulos et al., 2018; Wang et al., 2023). Model-selection approaches also adapt across model-specific conformal predictors (Hajihashemi and Shen, 2024), and Bellman conformal inference optimizes interval length using families of nominal prediction intervals (Yang et al., 2024). Our focus is the efficiency left by PID calibration configurations for a shared point forecaster. We combine the threshold paths generated by distinct feedback histories and use coverage of the final reported interval to adjust how these paths are evaluated.

We propose *PID-Expert aggregation* with two connected feedback layers. At the expert layer, each PID controller maintains its own coverage history and generates a candidate threshold. Keeping these states separate preserves the different calibration paths available

for comparison. At the aggregation layer, weights combine the candidate thresholds and are updated using normalized interval width and expert miscoverage. Coverage of the reported interval updates a common multiplier that sets the relative penalty on miscoverage in these expert losses. This additional feedback is needed because coverage of an averaged threshold is not the weighted average of the experts' coverage indicators. The multiplier therefore brings the reported interval's coverage into the comparison of calibration paths.

The paper makes three contributions. First, PID-Expert aggregation turns reliance on one prespecified PID configuration into an online comparison of independently evolving calibration paths, with an explicit width–miscoverage loss informed by aggregate coverage. Second, the analysis separates the two questions created by this design: local regret bounds compare weighted expert losses with a fixed library member under the realized multiplier sequence, while a separate pathwise relation links aggregate miscoverage to multiplier growth. A stability condition then gives time-averaged upper control in expectation and almost surely; the regret bound also motivates reference aggregation parameters. Third, four simulation settings and two real-data applications evaluate PID-Expert aggregation against the prespecified Conformal PID benchmark, DtACI, and AgACI, together with expert selection and simple averaging as structured benchmarks for isolating the role of online aggregation. Across the reported forecasting configurations, PID-Expert aggregation produces narrower mean intervals than the prespecified PID benchmark while maintaining empirical coverage close to the target. The comparison with selection and averaging shows that the gains are not explained simply by choosing a favorable expert or averaging expert outputs, while rolling-window analyses further characterize the associated local coverage–efficiency tradeoffs.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature and introduces the online prediction setup and conformal PID calibration. Section 3 presents PID-Expert aggregation. Section 4 develops the regret and coverage analysis. Section 5 evaluates the method through simulations and real-data applications. Section 6 discusses the findings and concludes. Proofs and additional implementation details are provided in the appendices.

2. Background and Problem Setup

This section places the calibration problem in the time-series literature, defines the online prediction setting, and introduces the PID states used by the expert library.

2.1. Conformal prediction for time series

Conformal prediction constructs prediction sets by calibrating nonconformity scores, which measure disagreement between predictions and observations (Vovk et al., 2005; Shafer and Vovk, 2008). Split conformal prediction separates model fitting from calibration and selects a score quantile from a calibration sample, obtaining finite-sample marginal coverage under exchangeability (Papadopoulos et al., 2002; Lei et al., 2018; Angelopoulos and Bates, 2023).

Temporal dependence and structural change motivate approaches beyond exchangeable calibration. In the forecasting literature, Koo and Seo (2015) study structural-break models under misspecification and their implications for forecasting, while Koo et al. (2026) distinguish changes in factor variances from changes in loadings in macroeconomic factor models. In conformal prediction, weighted calibration addresses covariate shift and other departures from exchangeability (Tibshirani et al., 2019; Barber et al., 2023), while permutation-based methods account for dependent data (Chernozhukov et al., 2018). Xu and Xie (2021) combine bootstrap ensembles with sequential residual calibration, and Xu and Xie (2023) use temporal dependence in scores to estimate conditional score quantiles. Their guarantees depend on the respective assumptions about the score process and estimation accuracy.

Other extensions broaden the prediction target. Conformal methods have been developed for multiple forecast horizons (Stankeviciute et al., 2021; Wang and Hyndman, 2024), multistep multivariate prediction (Schlembach et al., 2025), and functional time series with applications to age-specific mortality (Shang, 2026). These studies extend calibration to prediction tasks with different temporal and response structures.

Online feedback provides a further approach: ACI uses realized coverage errors to update subsequent calibration levels (Gibbs and Candès, 2021). Building on ACI, Zaffran et al. (2022) aggregate intervals from predictors with different step sizes, while Gibbs and Candès (2024) dynamically adapt the choice of step size. Such updates can also exploit predictable structure. Sun and Yu (2025) combine latent-state prediction with adaptive calibration to address change points. Areces et al. (2025) formulate online conformal prediction through online optimization, and Hu et al. (2026) incorporate estimated score distributions into optimistic threshold updates. We study one-step symmetric intervals, holding the point forecasts and score information fixed to compare the efficiency of different PID calibration paths.

2.2. Online prediction setup

Consider responses $y_t \in \mathbb{R}$ observed sequentially, with any covariates available at the forecast origin. Before observing y_t, we construct a one-step-ahead point forecast $\widehat{y}_{t|t-1}$ and choose a nonnegative threshold q_t. For a fixed target miscoverage rate $0 < \alpha < 1$, the reported interval is

$$\widehat{C}_{t,1-\alpha} = [\widehat{y}_{t|t-1} - q_t, \widehat{y}_{t|t-1} + q_t].$$

After observing the response, we obtain the nonconformity score and miscoverage indicator

$$s_t = |y_t - \widehat{y}_{t|t-1}|, \qquad err_t = \mathbf{1}\{y_t \notin \widehat{C}_{t,1-\alpha}\} = \mathbf{1}\{s_t > q_t\}.$$

The interval width is $2q_t$. Increasing the threshold can reduce miscoverage at the cost of width, so interval efficiency must be assessed alongside coverage.

We allow temporal dependence and distributional change and study calibration through sequential feedback. The objective is to keep the time-averaged miscoverage $T^{-1}\sum_{t=1}^{T} err_t$

near α while avoiding unnecessary width. This criterion differs from coverage at each forecast origin. The theory in Section 4 gives upper control of time-averaged miscoverage under its stated conditions.

For the quantiles used below, write

$$Q_\tau(P) = \inf\{r \in \mathbb{R} : P((-\infty, r]) \geq \tau\},$$

where P is a probability measure, and let δ_s denote a unit point mass at s.

2.3. Conformal PID calibration

Conformal PID updates the threshold directly on the score scale (Angelopoulos et al., 2023; Wang and Hyndman, 2024). Its components respond to recent miscoverage, accumulated coverage deviations, and predictable changes in future scores. These components make a controller's configuration relevant to both its coverage response and its interval width.

After observing the score at time t, the controller updates the nonnegative threshold as

$$q_{t+1} = \left[q^P_{t+1} + K_I \tan\left(\frac{I_t \log t}{t C_{\text{sat}}} \right) + \widehat{s}_{t+1|t} \right]_+, \tag{1}$$

where $[z]_+ = \max\{z, 0\}$.

The first component is the proportional, or quantile-tracking, state

$$q^P_{t+1} = q^P_t + \eta R_t (err_t - \alpha),$$

where $\eta > 0$ is the proportional learning rate that controls the responsiveness to the latest coverage error. For $t \geq 2$, we define the recent score scale as

$$R_t = \max_{r \in \mathcal{W}^P_t} s_r - \min_{r \in \mathcal{W}^P_t} s_r, \qquad \mathcal{W}^P_t = \{\max(1, t - L_P), \ldots, t-1\}.$$

This range is based on a lookback window of length L_P. In our implementation, we set $R_1 = 1$ when no previous score is available.

The integral component responds to accumulated deviations from the target miscoverage rate through

$$I_t = \sum_{r=1}^{t} (err_r - \alpha).$$

Positive values of I_t indicate cumulative undercoverage, whereas negative values indicate cumulative overcoverage. The integral gain $K_I \geq 0$ determines the magnitude of the correction on the score scale, while the saturation constant $C_{\text{sat}} > 0$ governs how rapidly the correction steepens as the cumulative error grows. Within the unsaturated range, smaller values produce stronger corrections for the same error at a given time. In our implementation, the tangent argument is clipped inside $(-\pi/2, \pi/2)$ to keep the update finite.

The final component, $\widehat{s}_{t+1|t}$, forecasts the next nonconformity score using information available through round t. This scorecaster provides the anticipatory component of the controller by exploiting predictable temporal structure in the score sequence beyond that captured by coverage feedback alone.

Together, the three components allow the controller to respond to recent coverage errors, correct persistent deviations from the target, and anticipate changes in the scale of forecast uncertainty. Holding the scorecaster and lookback rule fixed, its behavior depends on the controller configuration

$$\theta = (\eta, K_I, C_{\text{sat}}).$$

Different controller configurations can generate markedly different adjustment dynamics following a regime change. As illustrated schematically in Figure 1, an aggressive configuration may restore coverage rapidly but induce oscillations in both coverage and interval width, whereas a slower configuration can exhibit prolonged undercoverage as interval width adjusts only gradually. A more conservative configuration may maintain coverage above the target at the cost of persistently wider intervals. By contrast, a balanced configuration can deliver a more stable adjustment in both coverage and interval width. These contrasting responses motivate dynamically combining heterogeneous PID controllers rather than relying on a single prespecified configuration.

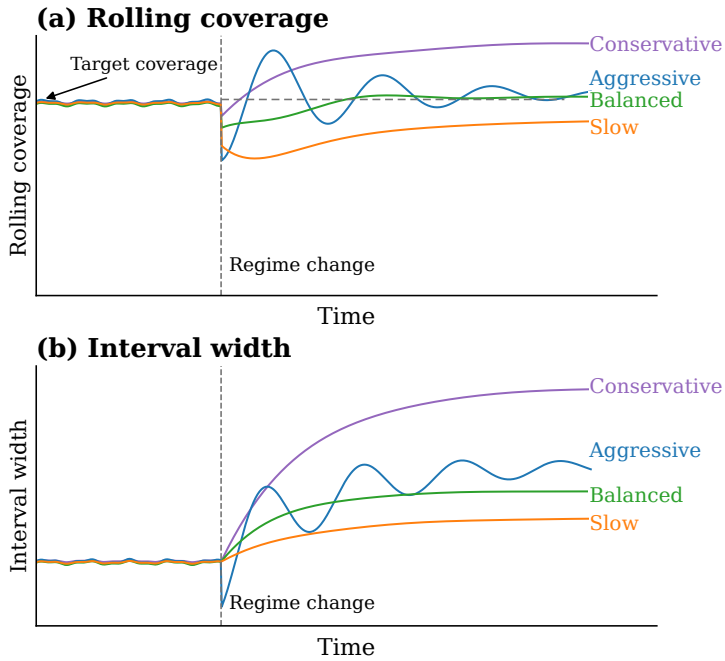


Figure 1: Schematic illustration of coverage and interval width dynamics under different PID controller configurations following a regime change.

3. PID-Expert aggregation

3.1. Overview

Our objective is to achieve a favorable coverage–efficiency tradeoff under evolving data distributions by constructing narrow prediction intervals while maintaining coverage close to the target level. To this end, we propose *PID-Expert aggregation*, an online method that adaptively combines candidate thresholds from a fixed library of PID controllers with heterogeneous parameter configurations. All experts share the same point forecasts and sequence of nonconformity scores, but each evolves according to its own history of coverage feedback. The aggregation mechanism is designed to account jointly for coverage performance and interval efficiency when adapting the relative influence of the experts.

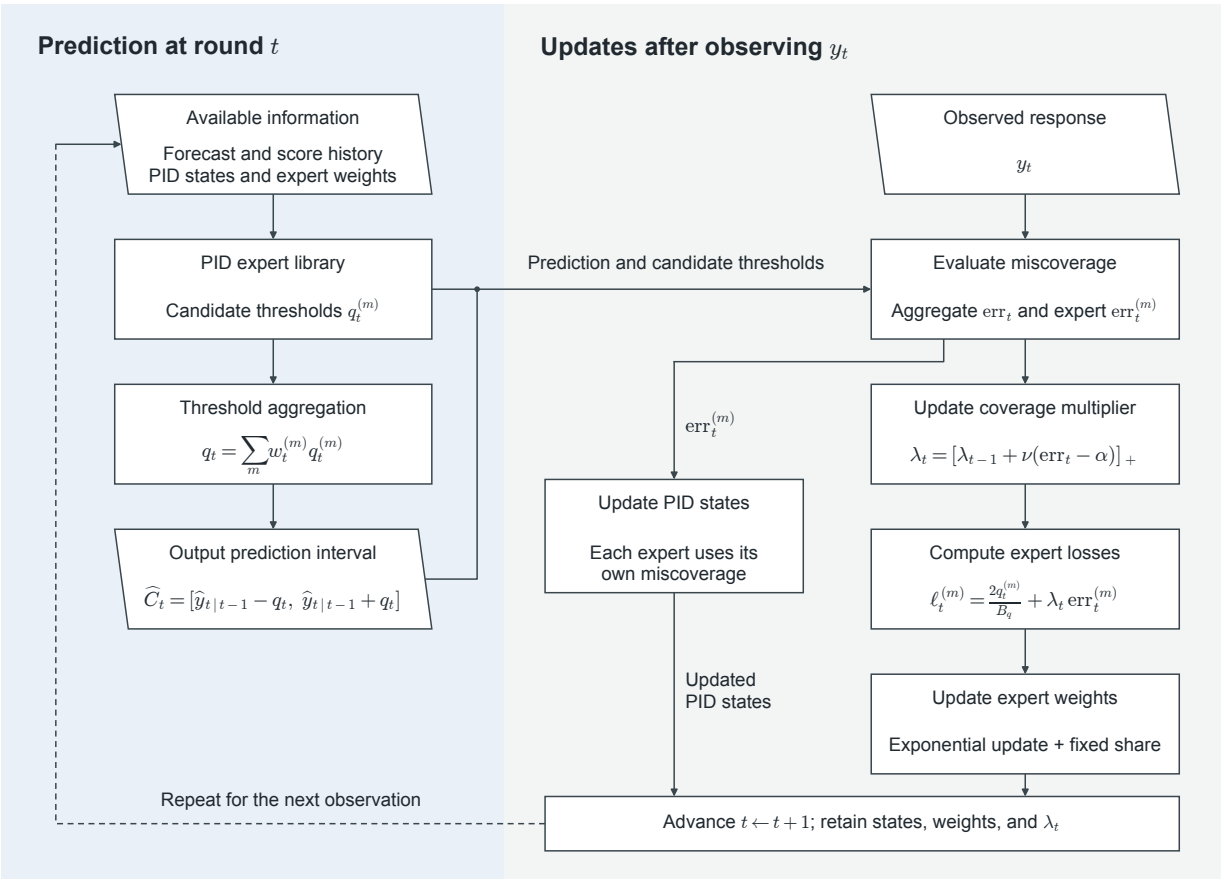


Figure 2: Framework of PID-Expert aggregation for sequential forecasting.

As illustrated in Figure 2, the method proceeds recursively through prediction and updating based on feedback. At each round, the candidate PID thresholds are combined using the current expert weights to form the reported prediction interval. Once the response is observed, aggregate coverage feedback adjusts the emphasis placed on miscoverage, while each expert's realized coverage and interval width inform the subsequent weight update. In parallel, each expert updates its own PID state using its individual coverage feedback. The resulting states and weights are then carried forward to the next round, allowing the ensemble to adapt its coverage–efficiency balance over time while preserving the distinct calibration dynamics of the individual experts.

3.2. Expert library

We construct a fixed library of PID controllers with heterogeneous configurations to provide a range of threshold responses to changing distributions of nonconformity scores. Expert m is specified by the parameter vector

$$\theta_m = (\eta_m, K_{I,m}, C_{\text{sat},m}), \qquad m = 1, \ldots, M.$$

All experts share the same point forecast, sequence of nonconformity scores, predictable score range R_t, and scorecaster contribution. Each expert maintains its own proportional state $q_t^{P,(m)}$, cumulative coverage error $I_t^{(m)}$, and candidate threshold $q_t^{(m)}$.

After observing the response, we compute each expert's miscoverage indicator

$$err_t^{(m)} = \mathbf{1}\{s_t > q_t^{(m)}\}.$$

Each controller follows the PID formulation in Equation (1), using its own parameters and miscoverage feedback. Specifically, its threshold is updated as

$$q_{t+1}^{(m)} = \left[q_{t+1}^{P,(m)} + K_{I,m} \tan\left\{ \frac{I_t^{(m)} \log t}{t C_{\text{sat},m}} \right\} + \widehat{s}_{t+1|t} \right]_+ .$$

Thus, the experts retain distinct threshold trajectories driven by their own coverage histories. We aggregate these trajectories rather than average the PID parameters and run a single controller. Averaging parameters would collapse the separate coverage histories and would generally produce a different threshold path, since each expert's miscoverage indicator depends on its own threshold. The aggregation weights leave the individual PID updates unchanged.

3.3. Coverage–efficiency loss and online aggregation

Coverage alone cannot distinguish an efficient interval from an unnecessarily wide interval that also covers the response. We therefore compare experts using both width and miscoverage. After observing the response at round t, expert m receives the loss

$$\ell_t^{(m)} = \frac{2q_t^{(m)}}{B_q} + \lambda_t err_t^{(m)}. \tag{2}$$

The width term favors narrower intervals, while the miscoverage term penalizes miscoverage according to the current multiplier λ_t. We use an initial window of L_0 nonconformity scores during an initial period before online evaluation and set the nominal normalization scale to

$$B_q = 2Q_{1-\alpha}\left(\frac{1}{L_0} \sum_{t=1}^{L_0} \delta_{s_t} \right), \tag{3}$$

with a positive lower bound applied in the implementation. The resulting $B_q > 0$ is held fixed throughout online evaluation. Thus, expert comparisons account for both interval

width and realized miscoverage, so a covered expert may still have a larger total loss if its interval is sufficiently wide.

An expert's coverage feedback controls its own threshold path, but the reported interval uses an average of thresholds. Its miscoverage indicator generally differs from the weighted average of expert miscoverage indicators. To feed this final outcome into expert evaluation, the coverage multiplier λ_t is updated using aggregate miscoverage err_t:

$$\lambda_t = [\lambda_{t-1} + \nu(err_t - \alpha)]_+ , \qquad \lambda_0 = 0, \tag{4}$$

where the multiplier step size $\nu > 0$ controls the response to aggregate coverage feedback. A miss increases the multiplier, whereas a covered response decreases it toward zero. This adjustment strengthens the miscoverage penalty when coverage falls short of the target and relaxes it when coverage exceeds the target. In the following experiments, we set $\nu = 0.02$. The updated λ_t enters the current expert losses, which in turn determine the weights for the next round.

Given the losses in Equation (2), we first apply exponential reweighting:

$$\widetilde{w}_{t+1}^{(m)} = \frac{w_t^{(m)} \exp\{-\xi \ell_t^{(m)}\}}{\sum_{j=1}^{M} w_t^{(j)} \exp\{-\xi \ell_t^{(j)}\}}. \tag{5}$$

Here, $\widetilde{w}_{t+1}^{(m)}$ is the intermediate normalized weight, and the learning rate $\xi > 0$ controls sensitivity to loss differences. The exponential update increases the relative weight of experts with lower loss while preserving the information contained in the current weight distribution.

A pure exponential update may give an expert so little weight that it recovers slowly when its subsequent losses improve. To keep early losses from effectively removing a configuration from future comparisons, we apply a fixed-share step (Herbster and Warmuth, 1998):

$$w_{t+1}^{(m)} = (1-\sigma)\widetilde{w}_{t+1}^{(m)} + \frac{\sigma}{M}, \tag{6}$$

where $0 \le \sigma \le 1$ is the share parameter. This update retains a fraction $1-\sigma$ of the allocation determined by the losses while redistributing the remaining mass uniformly across the expert library. For $\sigma > 0$, every expert keeps weight at least σ/M and remains available for later comparisons.

The resulting weights are then used to aggregate the experts' candidate thresholds at the next time step,

$$q_{t+1} = \sum_{m=1}^{M} w_{t+1}^{(m)} q_{t+1}^{(m)},$$

yielding the prediction interval $[\widehat{y}_{t+1|t} - q_{t+1}, \widehat{y}_{t+1|t} + q_{t+1}]$.

Algorithm 1 summarizes the procedure. Conditional on the point forecast and shared score information, the computational and storage costs at each round are both $O(M)$,

assuming that the controller states have fixed dimension. The loss and PID updates for individual experts can be performed in parallel.

Algorithm 1 PID-Expert aggregation

Require: Target miscoverage level α; PID library $\{\theta_m\}_{m=1}^M$; normalization scale $B_q > 0$; learning rate $\xi > 0$; share parameter $0 \le \sigma \le 1$; multiplier step size $\nu > 0$.

Initialize all PID states, $w_1^{(m)} \leftarrow 1/M$, for $m = 1, \dots, M$, and $\lambda_0 \leftarrow 0$.
for $t = 1, \dots, T$ **do**
  Obtain the shared forecast $\widehat{y}_{t|t-1}$ and candidate thresholds $\{q_t^{(m)}\}_{m=1}^M$.
  Set $q_t \leftarrow \sum_m w_t^{(m)} q_t^{(m)}$ and output $[\widehat{y}_{t|t-1} - q_t, \widehat{y}_{t|t-1} + q_t]$.
  Observe y_t and compute err_t and $\{err_t^{(m)}\}_{m=1}^M$.
  Update $\lambda_t = [\lambda_{t-1} + \nu(err_t - \alpha)]_+$.
  Compute the expert loss $\ell_t^{(m)} = 2q_t^{(m)}/B_q + \lambda_t err_t^{(m)}, \quad m = 1, \dots, M$.
  Update the intermediate weights $\widetilde{w}_{t+1}^{(m)} = \frac{w_t^{(m)} \exp\{-\xi \ell_t^{(m)}\}}{\sum_{j=1}^M w_t^{(j)} \exp\{-\xi \ell_t^{(j)}\}}, \quad m = 1, \dots, M$.
  Apply the fixed-share update $w_{t+1}^{(m)} = (1-\sigma)\widetilde{w}_{t+1}^{(m)} + \frac{\sigma}{M}, \quad m = 1, \dots, M$.
  Update each expert's PID states using its own coverage feedback $err_t^{(m)}$.
end for

4. Aggregation properties and coverage analysis

The aggregation rule raises two questions: how the expert weights compare with a fixed configuration, and how the reported interval controls miscoverage. These questions require separate arguments because the weighted average of expert miscoverage indicators generally differs from the miscoverage indicator of the aggregate threshold. We first study the comparison made by the expert weighting layer, then analyze the aggregate coverage feedback and state what the two results imply for interval efficiency.

4.1. Comparison properties of the expert weights

Fix an interval $\mathcal{I} = [r, s]$ of length $L = s - r + 1$. For the expert losses in Equation (2), define the weighted first and second moments

$$\bar{\ell}_t := \sum_{m=1}^M w_t^{(m)} \ell_t^{(m)}, \qquad \bar{v}_t := \sum_{m=1}^M w_t^{(m)} \big(\ell_t^{(m)}\big)^2.$$

The local regret relative to a fixed expert m is

$$\mathrm{Reg}_{\mathcal{I}}(m) := \sum_{t=r}^s \bar{\ell}_t - \sum_{t=r}^s \ell_t^{(m)}.$$

The comparison uses the same realized sequence of coverage multipliers $\{\lambda_t\}$ for every expert. Thus, it measures how the weighting layer performs relative to a fixed expert under the losses generated by the actual aggregation procedure.

Assumption 1 (Well-defined expert feedback)**.** *The expert library is finite, $M < \infty$, and the losses satisfy $0 \le \ell_t^{(m)} < \infty$ for all t and m. The aggregation parameters are fixed over time with $\xi > 0$ and $0 < \sigma \le 1/2$. The initial weights are uniform, $w_1^{(m)} = 1/M$, and subsequent weights are updated according to Equations (5) and (6).*

The restriction on σ is used for the regret analysis, while the weight update itself is defined for $0 \le \sigma \le 1$. Positive sharing ensures the uniform lower bound $w_t^{(m)} \ge \frac{\sigma}{M}$, while $\sigma \le 1/2$ allows the simplification $-\log(1-\sigma) \le 2\sigma$.

Theorem 1 (Local fixed-expert regret)**.** *Under Assumption 1, for any interval $\mathcal{I} = [r, s]$ of length $L = s - r + 1$ and any expert m,*

$$\mathrm{Reg}_{\mathcal{I}}(m) \le \xi \sum_{t=r}^{s} \bar{v}_t + \frac{\log(M/\sigma) - L\log(1-\sigma)}{\xi} \tag{7}$$

$$\le \xi \sum_{t=r}^{s} \bar{v}_t + \frac{\log(M/\sigma) + 2\sigma L}{\xi}. \tag{8}$$

The result follows from a fixed-share comparison argument (Herbster and Warmuth, 1998) and holds pathwise, even when the expert losses depend on aggregate coverage feedback through λ_t. It therefore requires no assumptions of independence, stationarity, or exchangeability. The bound depends on the accumulated second moments of the expert losses, so finiteness of the individual losses alone is not sufficient to guarantee sublinear regret.

The proof is given in Appendix A.2. An extension to an external benchmark sequence, conditional on a library approximation assumption, is stated and proved in Appendix A.3.

Application to aggregation parameters. We use the simplified fixed-expert bound in Equation (8) to derive practical reference scales for the aggregation parameters. We begin by fixing a local time scale $L \ge 1$ and optimizing the σ-dependent complexity term.

Corollary 1 (Fixed-share scale)**.** *For a prescribed local interval length L, the value minimizing the σ-dependent complexity term in Equation (8) is*

$$\sigma^\star = \frac{1}{2L}.$$

Next, define

$$S_{\mathcal{I}} := \sum_{t=r}^{s} \bar{v}_t, \qquad V_{\mathcal{I}} := \frac{S_{\mathcal{I}}}{L}.$$

Substituting $\sigma = 1/(2L)$ into Equation (8) leaves the ξ-dependent terms $\xi S_{\mathcal{I}} + \frac{\log(2ML)+1}{\xi}$. For a reference calculation that holds $S_{\mathcal{I}} > 0$ fixed, minimizing this expression over $\xi > 0$ gives

$$\xi_{\mathcal{I}}^\star = \sqrt{\frac{\log(2ML)+1}{S_{\mathcal{I}}}} = \sqrt{\frac{\log(2ML)+1}{LV_{\mathcal{I}}}}. \tag{9}$$

Since $V_{\mathcal{I}}$ is not available when the parameters are chosen, we use a reference scale $V > 0$ and set

$$\boxed{\sigma = \frac{1}{2L}, \qquad \xi = \sqrt{\frac{\log(2ML) + 1}{LV}}} \tag{10}$$

throughout each run.

The choice of L sets the local time scale for these reference parameters; the bound itself holds over arbitrary intervals without reoptimizing them. The scale V is fixed before evaluation using the reference calculation in Appendix A.4. The proof of Corollary 1 and the learning-rate calculation are included in Appendix A.2.

4.2. Coverage feedback for the reported interval

The multiplier recursion in Equation (4) yields the following pathwise relation between average aggregate miscoverage and multiplier growth.

Proposition 1 (Pathwise upper bound on aggregate miscoverage)**.** *For every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=1}^{T} err_t \leq \alpha + \frac{\lambda_T - \lambda_0}{\nu T}. \tag{11}$$

Consequently, along any realized path for which $\lambda_T = o(T)$,

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} err_t \leq \alpha. \tag{12}$$

The relation gives asymptotic control when the multiplier grows sublinearly. To obtain a sufficient condition for this behavior, we next impose negative drift once the multiplier is sufficiently large.

Let $\mathcal{F}_{t-1}$ denote the information available before the response at time t is observed, so that q_t and λ_{t-1} are $\mathcal{F}_{t-1}$-measurable.

Assumption 2 (High-multiplier negative drift)**.** *There exist constants $\nu\alpha \leq \Lambda < \infty$ and $0 < \delta \leq \alpha$ such that, almost surely for every $t \geq 1$,*

$$\mathbb{E}[err_t - \alpha \mid \mathcal{F}_{t-1}] \leq -\delta \qquad \textit{whenever } \lambda_{t-1} \geq \Lambda. \tag{13}$$

Moreover, the multiplier evolves according to the recursion in Equation (4) from a fixed finite initial value $\lambda_0 \geq 0$. The algorithm uses $\lambda_0 = 0$.

This assumption requires the conditional aggregate miscoverage probability to be at most $\alpha - \delta$ whenever $\lambda_{t-1} \geq \Lambda$. Since $\Lambda \geq \nu\alpha$, the projection is inactive in this region, which implies $\mathbb{E}[\lambda_t - \lambda_{t-1} \mid \mathcal{F}_{t-1}] \leq -\nu\delta$. The restriction on Λ entails no loss of generality, since any finite threshold above which the condition holds can be increased to at least $\nu\alpha$. A larger multiplier places greater emphasis on miscoverage in the expert losses and thereby

influences subsequent weights. This negative-drift property is an additional assumption on the combined expert library, controller dynamics, and data-generating process.

Proposition 2 (Mean-square sublinearity of the multiplier)**.** *Under Assumption 2, define*

$$C_\lambda := \nu^2 + 2\nu\Lambda(1-\alpha).$$

Then, for every $T \geq 1$,

$$\mathbb{E}[\lambda_T^2] \leq \lambda_0^2 + C_\lambda T. \tag{14}$$

Consequently,

$$\mathbb{E}\left[\left(\frac{\lambda_T}{T}\right)^2\right] \leq \frac{\lambda_0^2}{T^2} + \frac{C_\lambda}{T} \longrightarrow 0. \tag{15}$$

Thus, $\lambda_T/T \to 0$ in L^2 and hence in probability. Together with the bounded increments of the multiplier recursion, the second-moment bound also yields $\lambda_T/T \to 0$ almost surely; see Appendix A.6.

Combining Propositions 1 and 2 yields a nonasymptotic bound for expected average miscoverage.

Corollary 2 (Expected aggregate miscoverage)**.** *Under Assumption 2,*

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} err_t\right] &\leq \alpha + \frac{\sqrt{\mathbb{E}[\lambda_T^2]}}{\nu T} \\ &\leq \alpha + \frac{\sqrt{\lambda_0^2 + C_\lambda T}}{\nu T} = \alpha + O(T^{-1/2}).\end{aligned} \tag{16}$$

Under the same assumption, the almost-sure sublinearity of the multiplier further implies that Equation (12) holds almost surely. Proofs of the coverage results appear in Appendix A.5–Appendix A.7.

4.3. Role of the aggregation components

The pathwise coverage result is obtained from the realized aggregate error after the reported threshold has been formed. The corresponding feedback pathway can be summarized as

$$(\{\theta_m\}, B_q, \xi, \sigma, \nu) \longrightarrow \{q_t^{(m)}, w_t^{(m)}\}_{m=1}^M \longrightarrow q_t \longrightarrow err_t \longrightarrow \lambda_t \longrightarrow \{\ell_t^{(m)}, w_{t+1}^{(m)}\}_{m=1}^M.$$

The quantities $\{\theta_m\}$, B_q, ξ, and σ do not appear explicitly in Equation (11) because the pathwise bound follows directly from the scalar multiplier update. They nevertheless remain behaviorally important because, through the expert thresholds and weights, they affect q_t, the realized miscoverage sequence, and the evolution of λ_t. These components therefore influence whether Assumption 2 holds and, when it does, the admissible values of the stability constants Λ and δ.

More generally, Proposition 1 remains algebraically valid for any predictable rule that generates q_t and subsequently applies the same multiplier update. The expected and almost-sure asymptotic guarantees additionally rely on the high-multiplier negative-drift condition in Assumption 2.

5. Experiments

In this section, we evaluate the proposed PID-Expert aggregation method through simulations and two real-data applications. The following experiment choices are used throughout:

(a) All experiments use one-step-ahead forecasting with symmetric prediction intervals at the nominal coverage level $1 - \alpha = 0.9$.

(b) The benchmark methods are configured following Angelopoulos et al. (2023) for Conformal PID and Gibbs and Candès (2024) for DtACI and AgACI. AgACI is implemented here by aggregating calibration levels; the original method aggregates interval endpoints separately (Zaffran et al., 2022). Further implementation details are provided in Appendix B and Appendix C. We also include two natural variants of PID-Expert as benchmarks, using the same expert threshold trajectories: PID-Expert selection adaptively uses the threshold from whichever expert receives the highest weight in PID-Expert aggregation at each time step; PID-Expert average takes an arithmetic mean of all expert thresholds. These comparisons change how the final threshold is formed; selection reuses the aggregation weights without a separate weight update.

(c) For PID-Expert aggregation, we set the multiplier step size to $\nu = 0.02$ and fix the width normalization scale using the initial score window, as defined in Equation (3). We use the prespecified reference value $V = 1.3533$ in the parameter rule $\sigma = 1/(2L)$ and $\xi = \sqrt{(\log(2ML) + 1)/(LV)}$ from Equation (10). The expert library is constructed from a parameter grid around the recommended Conformal PID configuration; the grid is specified in Appendix B.

(d) We assess coverage using empirical and rolling coverage and interval efficiency using mean and median width. We also report the mean scaled interval score (MSIS, Gneiting and Raftery, 2007), which accounts jointly for interval width and the magnitude of coverage failures. For a symmetric interval with half-width q_t and absolute forecast error s_t, the interval score (Winkler, 1972) is

$$\mathrm{IS}_t = 2q_t + \frac{2}{\alpha}(s_t - q_t)_+.$$

MSIS is the average interval score over the evaluation period divided by the mean absolute error of a seasonal naive forecast for seasonal series, or a one-step naive

forecast for nonseasonal series. Lower MSIS values indicate better interval forecasts under this scoring criterion.

5.1. Simulations

In this subsection, we evaluate the proposed aggregation method using two simulation designs: a nonlinear autoregressive process with exogenous predictors and Gaussian sequences with stationary, gradually changing, or abruptly changing means.

5.1.1. Nonlinear simulation

The first experiment follows the nonlinear data-generating process of Wang and Hyndman (2024),

$$y_t = \sin(y_{t-1}) + 0.5\log(y_{t-2}+1) + 0.1y_{t-1}x_{1,t} + 0.3x_{2,t} + \varepsilon_t,$$

where $x_{1,t}, x_{2,t} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ and ε_t is Gaussian white noise with standard deviation 0.1. After an appropriate burn-in period, we generate 2,000 data points.

One-step-ahead point forecasts are produced by a feed-forward neural network with a single hidden layer, using lagged responses and the two exogenous predictors as inputs. Following the neural autoregression design of Wang and Hyndman (2024), the model is refitted at each forecast origin using a rolling training window of 500 observations, with a calibration window of the same length. Conformal PID and PID-Expert use the same Theta model scorecaster (Assimakopoulos and Nikolopoulos, 2000). The PID-Expert library contains $M = 343$ controllers. We set the calibration window size $L_0 = 500$ used to give B_q and $L = 100$ as a local time scale, yielding $\sigma = 0.005$ and $\xi = 0.299$.

Table 1: Performance in the nonlinear simulation in terms of empirical coverage, mean and median interval width, minimum rolling coverage, and MSIS.

Method	Coverage	Mean width	Median width	Min roll. cov.	MSIS
DtACI	0.892	**0.337**	**0.333**	0.820	**2.453**
AgACI	0.894	**0.337**	0.335	0.810	**2.453**
Conformal PID	0.904	0.360	0.357	**0.860**	2.513
PID-Expert Selection	0.895	0.350	0.349	0.840	2.503
PID-Expert Average	0.909	0.363	0.358	**0.860**	2.504
PID-Expert Aggregation	**0.903**	0.352	0.350	0.850	2.489

Minimum rolling coverage is computed over a window of 100 observations.

Table 1 reports the results for the simulated series. Bold entries indicate empirical coverage closest to the nominal level, the highest minimum rolling coverage, and the smallest width and MSIS values. DtACI and AgACI yield the narrowest intervals and lowest MSIS, but their empirical coverage falls below the nominal level and their minimum rolling coverage is also lower. PID-Expert aggregation achieves empirical coverage of 0.903, the closest to the nominal level among the six methods. Compared with Conformal PID, it reduces the mean interval width from 0.360 to 0.352 and MSIS from 2.513 to 2.489, while

maintaining a comparable minimum rolling coverage of 0.850, versus 0.860 for Conformal PID. PID-Expert selection further reduces the mean width slightly to 0.350, but its coverage falls to 0.895. By contrast, PID-Expert average attains higher coverage of 0.909 but produces a wider interval with a mean width of 0.363. Both alternatives have higher MSIS than PID-Expert aggregation. Overall, PID-Expert aggregation provides a favorable balance between coverage and interval efficiency.

Figure 3 shows how coverage and interval width evolve over rolling windows of 100 observations. Conformal PID and PID-Expert aggregation exhibit shallower declines in rolling coverage than DtACI and AgACI, with minimum rolling coverage of 0.860 and 0.850, compared with 0.820 and 0.810, respectively. PID-Expert selection reaches a lower minimum rolling coverage of 0.840, whereas PID-Expert average matches Conformal PID at 0.860. In terms of interval width, PID-Expert aggregation generally produces narrower rolling mean and median widths than Conformal PID and PID-Expert average while maintaining rolling coverage close to the target level.

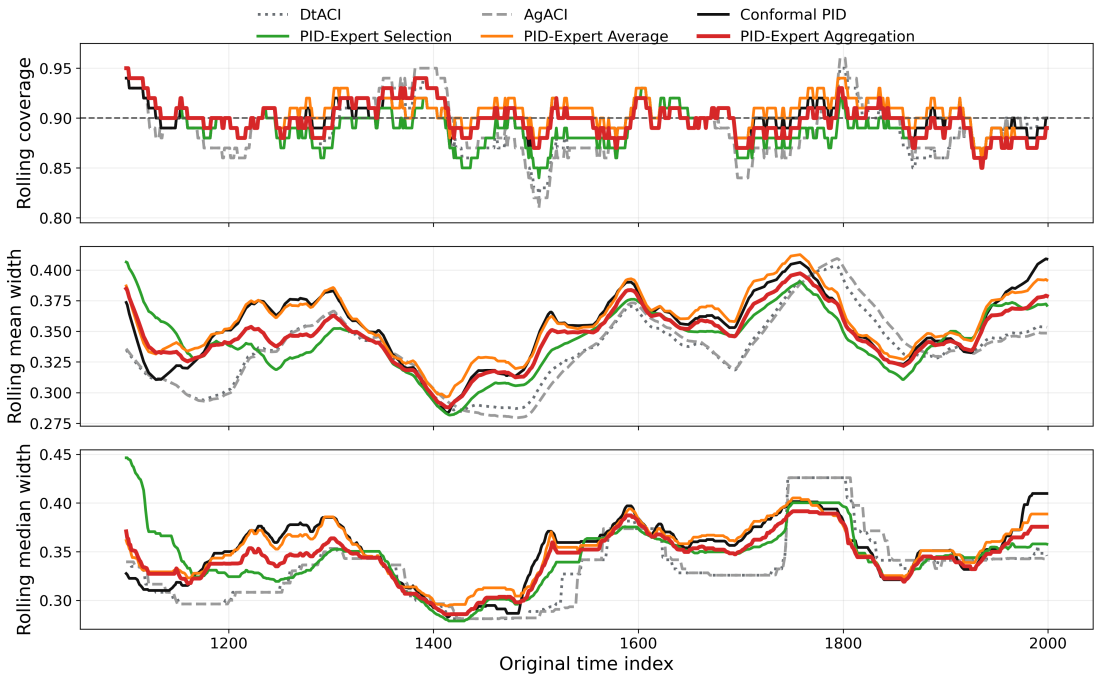


Figure 3: Rolling coverage, mean interval width, and median interval width in the nonlinear simulation, computed over a window of 100 observations. The horizontal dashed line marks the target coverage level of 0.9.

Figure 4 compares PID-Expert aggregation with all 343 candidate PID controllers in the library. Across these experts, empirical coverage ranges from 0.901 to 0.909, while mean interval width varies considerably. For reference, we define the hindsight best feasible expert as the controller with the smallest mean width among those achieving empirical coverage of at least 0.9 on this realization. This expert has a mean width of 0.348, compared with 0.352 for PID-Expert aggregation and 0.360 for Conformal PID. Based on these reported values, PID-Expert aggregation closes approximately 67% of the gap in mean interval width between Conformal PID and the hindsight benchmark. PID-Expert selection

produces a slightly narrower interval, with a mean width of 0.350, but its empirical coverage falls to 0.895, below the target level. In contrast, PID-Expert average raises coverage to 0.909 but increases the mean width to 0.363, exceeding that of Conformal PID. This comparison illustrates how online aggregation can approach the performance of a favorable library member without selecting that controller ex post using the evaluation outcomes.

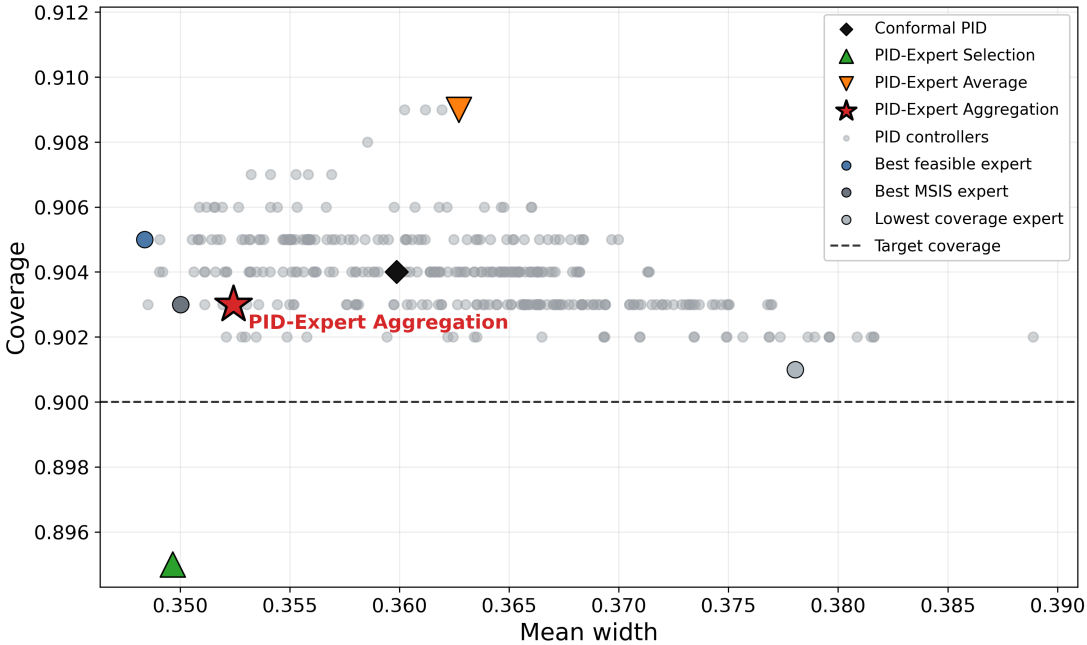


Figure 4: Empirical coverage and mean interval width for the 343 PID controllers, Conformal PID, PID-Expert selection, PID-Expert average, and PID-Expert aggregation in the nonlinear simulation. The best feasible expert is identified ex post from the evaluation sample and therefore represents a hindsight oracle.

Figure 5 illustrates the evolution of the aggregation weights for the five experts attaining the highest peak weights. E23 receives greater weight early in the evaluation period, followed by E27 over much of the middle portion. The allocation subsequently shifts toward E13, with E136 and E128 gaining greater influence later in the sequence. The experts favored in the earlier periods use $\eta = 0.025$, whereas E136 and E128 use $\eta = 0.075$ together with larger integral gains K_I. These evolving weight patterns illustrate how PID-Expert reallocates influence across heterogeneous controller configurations as their relative performance changes over time, rather than relying on a single prespecified configuration throughout.

5.1.2. Distribution-shift simulations

Following Gibbs and Candès (2024), we consider three Gaussian simulation settings with stationary, gradually changing, and abruptly changing means. Conditional on the mean trajectory, observations are independent and satisfy

$$Y_t \mid \mu_t \sim N(\mu_t, 1).$$

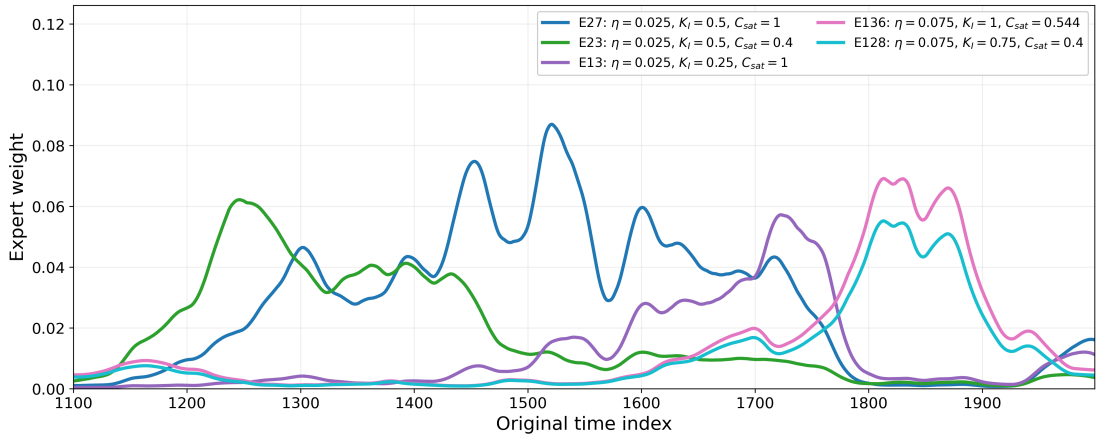


Figure 5: Evolution of the PID-Expert weights for the five experts attaining the highest peak weights in the nonlinear simulation. The displayed trajectories are smoothed using a moving average over 15 time steps for visualization, and labels such as E23 denote expert indices.

For each setting, we generate a sequence of 6,000 observations. The three settings differ only in the evolution of μ_t.

Stationary. The mean remains fixed at $\mu_t = 0$, so the observations are independent standard normal random variables.

Smooth shift. The mean evolves gradually through serially correlated increments. At the update times, it follows

$$m_{j+1} = m_j + \frac{1}{2}(m_j - m_{j-1}) + \frac{1}{2}\varepsilon_j, \qquad \varepsilon_j \overset{\text{iid}}{\sim} N(0, 0.006),$$

starting from $m_0 = m_1 = 0$. Each mean value is held fixed for five consecutive observations, so that $\mu_t = m_{\lceil t/5 \rceil}$.

Jump shift. The sequence is divided into three equal segments. Within each segment, the mean alternates every 500 observations between -0.075 and 0.075 in the first and last segments, and between -1.5 and 1.5 in the middle segment. The resulting design introduces two abrupt transitions between regimes with markedly different magnitudes of distributional change.

We adapt these settings to symmetric prediction intervals, using the nonconformity score $s_t = |Y_t|$ and reporting $[-q_t, q_t]$ for all six methods considered. In the jump setting, the distribution of $|Y_t|$ changes with the magnitude of the mean but not with its sign, so sign reversals within a segment leave the score distribution unchanged.

PID-Expert aggregation uses the same library of $M = 343$ controllers in all three scenarios, with $L_0 = L = 500$, giving $\sigma = 0.001$ and $\xi = 0.143$. Each scenario is evaluated over 100 trials with independent observation noise and a common mean trajectory, as shown in Figure 6. Performance is evaluated after an initial period of 500 observations, and rolling metrics use windows of 500 observations.

Table 2 summarizes performance averaged over 100 trials. PID-Expert aggregation consistently improves interval efficiency relative to Conformal PID, reducing mean width from 3.414 to 3.338 under stationarity, from 5.084 to 4.976 under smooth shifts, and from

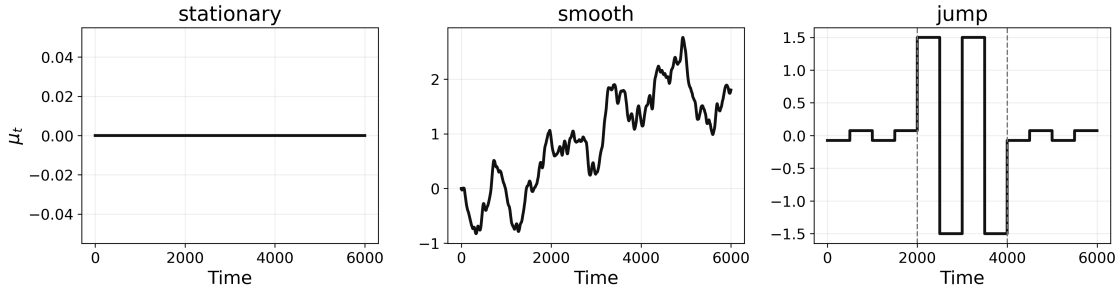


Figure 6: Mean trajectories used in the stationary, smooth-shift, and jump-shift simulation designs.

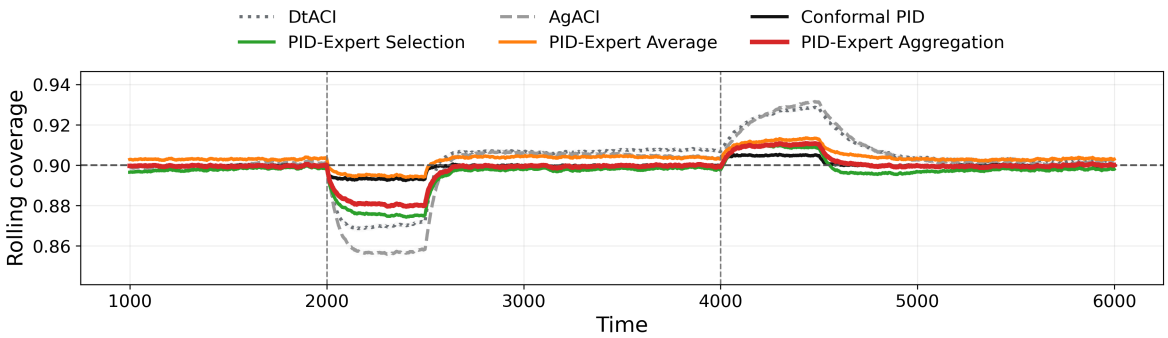
Table 2: Performance in the stationary, smooth-shift, and jump-shift simulations, averaged over 100 trials.

Scenario	Method	Coverage	Mean width	Min roll. cov.	MSIS
Stationary	DtACI	0.8996	3.298	0.878	3.661
	AgACI	0.8997	**3.297**	0.875	**3.658**
	Conformal PID	**0.8998**	3.414	**0.892**	3.788
	PID-Expert Selection	0.8983	3.313	0.884	3.694
	PID-Expert Average	0.9034	3.420	0.891	3.757
	PID-Expert Aggregation	0.8997	3.338	0.886	3.705
Smooth	DtACI	0.9057	5.802	0.870	5.940
	AgACI	0.9032	5.687	0.870	5.862
	Conformal PID	**0.8990**	5.084	**0.888**	5.367
	PID-Expert Selection	0.8951	**4.945**	0.867	**5.262**
	PID-Expert Average	0.9025	5.091	0.887	5.323
	PID-Expert Aggregation	0.8970	4.976	0.873	5.267
Jump	DtACI	0.9027	4.526	0.861	4.796
	AgACI	0.9009	4.436	0.848	4.732
	Conformal PID	**0.8997**	4.259	**0.888**	4.596
	PID-Expert Selection	0.8967	**4.147**	0.867	**4.504**
	PID-Expert Average	0.9034	4.270	**0.888**	4.558
	PID-Expert Aggregation	0.8988	4.172	0.873	4.507

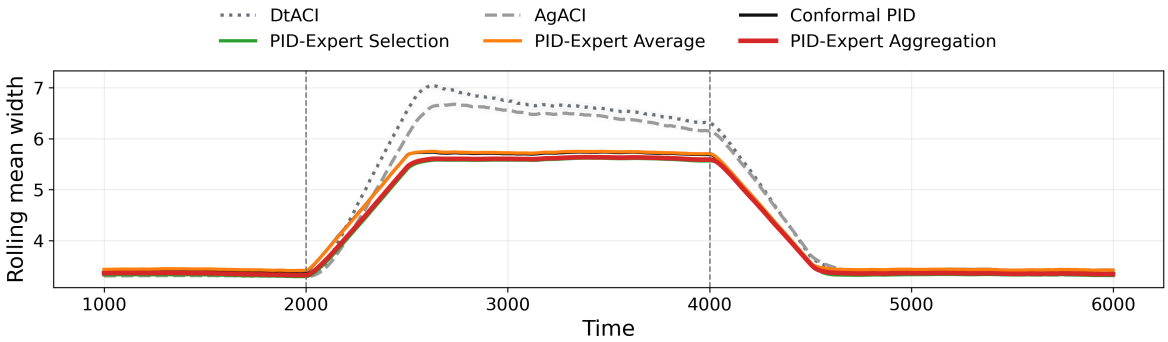
Minimum rolling coverage over 500-observation windows is calculated within each trial and averaged across 100 trials.

4.259 to 4.172 under jump shifts. MSIS is also lower in all three scenarios, while empirical coverage remains close to the nominal level at 0.8997, 0.8970, and 0.8988, respectively.

Among the PID-Expert variants, selection produces the narrowest intervals and the lowest MSIS under smooth and jump shifts, but also has the lowest empirical and minimum rolling coverage. In contrast, PID-Expert average maintains higher coverage and minimum rolling coverage, but produces slightly wider intervals than Conformal PID in all three scenarios. PID-Expert aggregation lies between these two alternatives, retaining much of the width reduction from selection while providing better coverage. Under stationarity, DtACI and AgACI achieve the smallest widths and MSIS with coverage close to 0.9. Under smooth and jump shifts, however, both methods produce substantially wider intervals and higher MSIS than the PID-based methods, despite coverage at or above the nominal level.



(a) Rolling coverage



(b) Rolling mean width

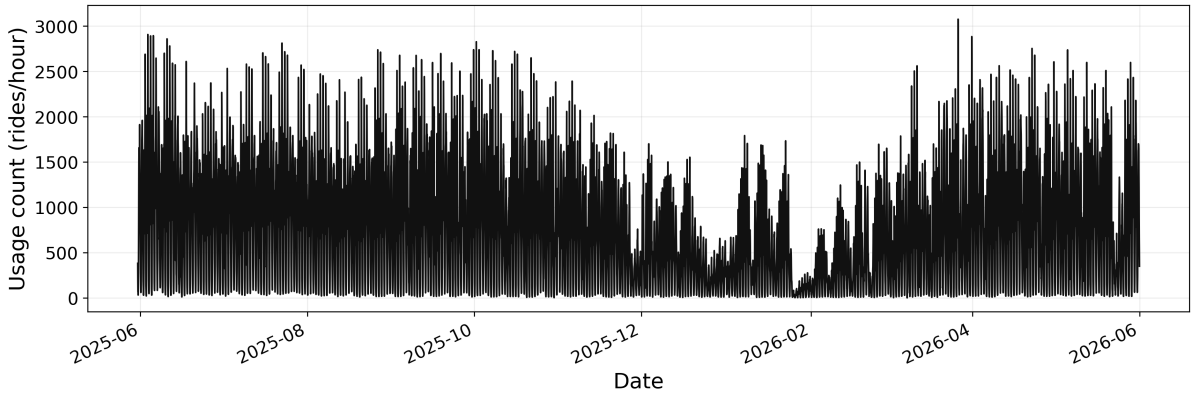
Figure 7: Jump-shift simulation: coverage and mean interval width over rolling windows of 500 observations, averaged over 100 trials. Shaded regions show approximate pointwise 95% Monte Carlo confidence intervals for the mean trajectories. The horizontal dashed line marks nominal coverage of 0.9, and vertical dashed lines mark changes in mean magnitude.

Figure 7 shows the rolling coverage and mean width in the jump scenario, averaged over trials. The increase in mean magnitude leads to a temporary decline in coverage for all six methods. This decline is smaller for the PID-based methods than for DtACI and AgACI, although PID-Expert aggregation experiences a somewhat larger dip than Conformal PID. PID-Expert selection exhibits a deeper coverage decline than PID-Expert aggregation and produces slightly narrower intervals, whereas PID-Expert average shows a shallower decline but wider intervals. When the mean magnitude decreases, the ACI variants exhibit a larger and more persistent coverage overshoot. Throughout most of the high-magnitude regime, PID-Expert aggregation closely tracks PID-Expert selection in mean width and generally remains slightly below PID-Expert average and Conformal PID.

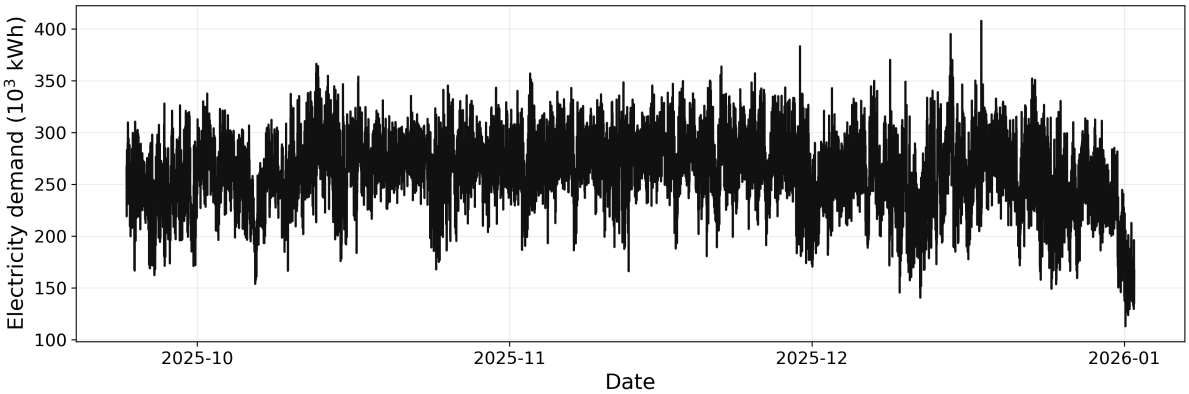
5.2. Real-data applications

In this subsection, we use two real-world datasets to evaluate the proposed method. Following Montero-Manso et al. (2020), we consider six forecasting methods: random walk with drift, automatic exponential smoothing (Auto ETS), feed-forward neural networks with a single hidden layer (NNAR), Theta, TBATS, and STLM-AR. These methods are implemented using the *forecast* package in R (Hyndman et al., 2026). To examine performance under different forecasting specifications, we report results for the first three individual models and for an equally weighted average forecast (Average model) constructed

from all six models.



(a) Capital Bikeshare hourly usage from 31 May 2025 to 31 May 2026.



(b) Ordos electricity demand from 24 September 2025 to 1 January 2026.

Figure 8: Observed Capital Bikeshare usage and Ordos electricity demand.

5.2.1. Capital Bikeshare data

The Capital Bikeshare dataset contains hourly trip counts for the Washington, DC metropolitan area. We use 8,784 hourly observations spanning from 00:00 on 31 May 2025 to 23:00 on 31 May 2026. As shown in the upper panel of Figure 8, the series exhibits pronounced high-frequency periodicity together with substantial seasonal variation in both its level and volatility. Usage is generally higher during the warmer months, declines markedly over the winter, and rises again in spring. The pronounced changes in the level and amplitude of the series provide a challenging setting for online prediction interval calibration. We use the first 28 days to train the forecasting models and set $L = L_0 = 168$, corresponding to one week of hourly observations. This yields $\sigma = 0.00298$ and $\xi = 0.243$. Interval widths are measured in rides per hour. The data are publicly available at `https://capitalbikeshare.com/system-data`.

As shown in Table 3, PID-Expert aggregation consistently produces narrower mean intervals than DtACI, AgACI, and Conformal PID across all four forecasting specifications, while maintaining empirical coverage close to the nominal level of 0.9. Relative to Conformal PID, it reduces mean interval width by 4.8%, 5.5%, 6.4%, and 6.6% for the random walk with drift, Auto ETS, NNAR, and Average model, respectively, with only

Table 3: Capital Bikeshare results across the four forecasting specifications, reporting empirical coverage, mean and median interval widths, and MSIS. Widths are measured in rides per hour.

Forecasting model	Method	Coverage	Mean width	Median width	MSIS
Random walk with drift	DtACI	0.896	881.2	906.5	6.00
	AgACI	0.906	899.1	**886.7**	**5.91**
	Conformal PID	**0.899**	910.8	920.0	6.09
	PID-Expert Selection	0.888	**857.9**	898.4	6.21
	PID-Expert Average	0.911	907.1	916.1	5.98
	PID-Expert Aggregation	0.898	866.8	898.3	6.03
Auto ETS	DtACI	0.896	577.0	619.2	4.65
	AgACI	**0.900**	590.4	620.9	**4.59**
	Conformal PID	0.899	603.7	614.0	4.76
	PID-Expert Selection	0.886	**561.8**	**597.8**	4.72
	PID-Expert Average	0.907	609.8	624.7	4.68
	PID-Expert Aggregation	0.897	570.6	602.5	4.65
NNAR	DtACI	0.894	334.0	**313.9**	2.84
	AgACI	0.903	349.1	320.4	**2.81**
	Conformal PID	0.899	355.0	327.4	2.89
	PID-Expert Selection	0.890	**326.8**	318.7	2.89
	PID-Expert Average	0.912	358.3	331.9	2.83
	PID-Expert Aggregation	**0.900**	332.2	322.2	2.84
Average model	DtACI	0.896	446.2	466.9	**3.56**
	AgACI	**0.900**	461.9	**463.4**	3.58
	Conformal PID	0.899	477.7	484.1	3.65
	PID-Expert Selection	0.890	**440.5**	471.0	3.59
	PID-Expert Average	0.906	482.9	491.2	3.61
	PID-Expert Aggregation	0.896	446.1	471.4	3.57

small changes in empirical coverage. It also achieves lower MSIS than Conformal PID in all four cases. PID-Expert selection yields the smallest mean widths, but at the cost of lower coverage, ranging from 0.886 to 0.890. In contrast, PID-Expert average attains higher coverage, ranging from 0.906 to 0.912, but produces wider intervals. PID-Expert aggregation lies between these two alternatives, keeping coverage closer to the target while retaining much of the width reduction achieved by selection. Although it does not uniformly attain the smallest median width or MSIS, both measures remain competitive in each case.

Figure 9 further examines the distribution of rolling performance. Across all four forecasting specifications, the rolling mean width distribution of PID-Expert aggregation is consistently shifted below that of Conformal PID, while its rolling coverage remains centered close to the nominal level of 0.9. The three PID-Expert variants also reveal a clear coverage–width trade-off. PID-Expert selection generally produces the narrowest intervals, but its rolling coverage distributions are systematically shifted below the target. PID-Expert average shifts coverage upward, but at the cost of noticeably wider intervals. PID-Expert aggregation lies between these two alternatives, retaining much of the width reduction from selection while bringing coverage closer to the target than selection and producing narrower intervals than averaging. The consistency of this pattern across the four forecasting models suggests that the gain from aggregation is not specific to a particular forecasting specification.

5.2.2. Electricity demand in Ordos

Accurate electricity demand forecasting is an important component of power system operation and planning, with implications for power dispatch, generation scheduling, and

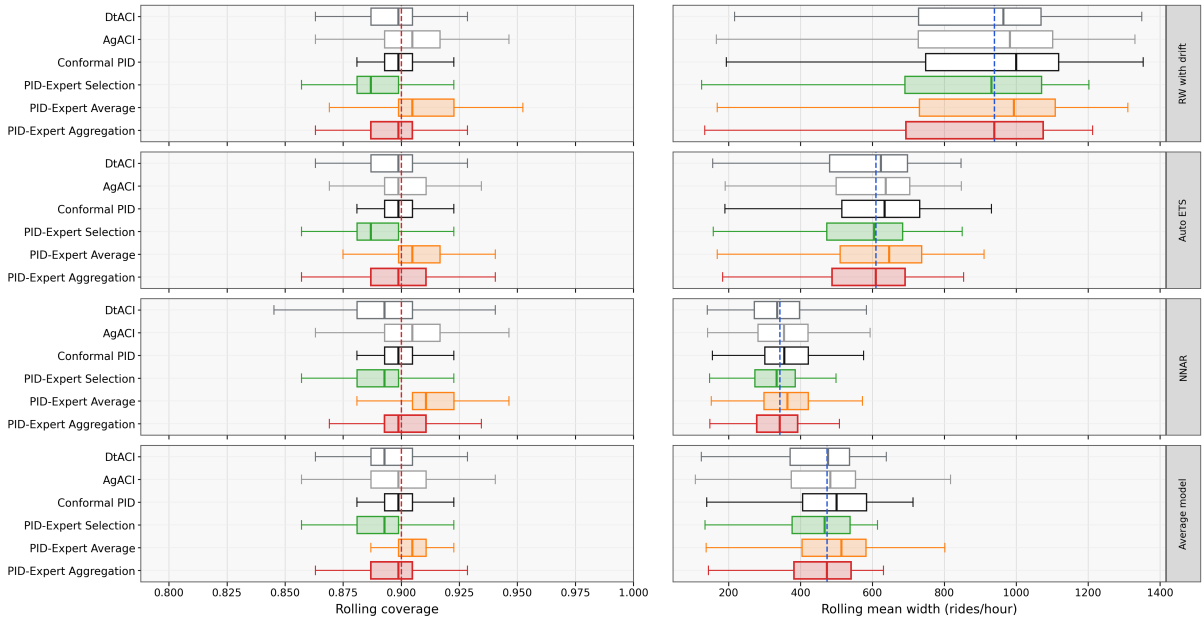


Figure 9: Rolling coverage and mean interval width for the Capital Bikeshare application. Dashed lines indicate the nominal coverage level of 0.9 and the median rolling mean width of PID-Expert aggregation.

grid reliability (Weron, 2014; Kath and Ziel, 2021; Devine and Bertsch, 2026; Lipiecki et al., 2026). The electricity demand dataset consists of operational metering records from 153 industrial enterprises above designated size in Ordos, China, sampled at 15-minute intervals. We use 9,600 observations spanning from 00:00 on 24 September 2025 to 23:45 on 1 January 2026. Electricity demand, measured in kWh, is used as the response variable. The Ordos electricity demand data used in this study are available from the corresponding author upon request.

As shown in the lower panel of Figure 8, the electricity series exhibits substantial high-frequency variation around a relatively stable level over most of the sample period. The variability becomes more pronounced in parts of December, followed by a marked decline in demand toward the end of the observation period. These time-varying features provide a useful setting for evaluating sequential prediction intervals under changing demand conditions. We use the first 14 days to train the forecasting models described above and set the local adaptation and initial score windows to $L = L_0 = 672$, corresponding to one week of 15-minute observations. This yields $\sigma = 0.000744$ and $\xi = 0.128$. Interval widths are reported in units of 10^3 kWh.

Table 4 reports the results across the four forecasting specifications. PID-Expert aggregation consistently produces narrower mean intervals than Conformal PID while maintaining empirical coverage close to the nominal level of 0.9. Relative to Conformal PID, it reduces mean interval width by approximately 1.7%–3.3% and MSIS by 1.2%–2.4%. PID-Expert selection yields still narrower mean intervals, but with lower coverage under the random walk with drift, Auto ETS, and NNAR specifications. In contrast, PID-Expert average is more conservative, with coverage of 0.907–0.908 and wider intervals than both Conformal PID and PID-Expert aggregation.

Comparisons with DtACI and AgACI vary across forecasting models. For the random walk with drift and Auto ETS specifications, AgACI attains the lowest MSIS, although PID-Expert aggregation remains competitive in interval width. Under NNAR, PID-Expert aggregation shares the lowest reported MSIS with DtACI and AgACI, while under the Average model it achieves both a narrower mean interval and a lower MSIS than the two ACI methods. Overall, the results show that PID-Expert aggregation improves interval efficiency relative to Conformal PID consistently across forecasting specifications, while maintaining coverage close to the target.

Table 4: Ordos electricity demand results across the four forecasting specifications. The table reports empirical coverage, mean and median interval widths, and the mean scaled interval score (MSIS). Interval widths are measured in 10^3 kWh.

Forecasting model	Method	Coverage	Mean width	Median width	MSIS
Random walk with drift	DtACI	0.899	**127.5**	**126.0**	4.83
	AgACI	0.903	129.0	126.1	**4.79**
	Conformal PID	**0.900**	130.4	128.5	4.88
	PID-Expert Selection	0.898	**127.5**	128.3	4.89
	PID-Expert Average	0.907	131.4	129.7	4.80
	PID-Expert Aggregation	0.901	128.2	128.5	4.82
Auto ETS	DtACI	0.901	89.0	**86.2**	3.45
	AgACI	**0.900**	89.1	86.5	**3.44**
	Conformal PID	**0.900**	92.0	89.8	3.53
	PID-Expert Selection	0.898	**88.6**	87.2	3.45
	PID-Expert Average	0.908	92.5	90.0	3.48
	PID-Expert Aggregation	0.899	89.0	87.7	3.45
NNAR	DtACI	**0.899**	83.4	**78.6**	**3.23**
	AgACI	**0.899**	83.8	79.1	**3.23**
	Conformal PID	**0.899**	85.5	81.9	3.29
	PID-Expert Selection	0.896	**83.2**	78.9	3.25
	PID-Expert Average	0.907	86.2	82.2	3.25
	PID-Expert Aggregation	**0.899**	83.6	80.0	**3.23**
Average model	DtACI	**0.900**	84.4	**80.7**	3.26
	AgACI	0.901	84.5	81.0	3.26
	Conformal PID	**0.900**	86.9	83.7	3.33
	PID-Expert Selection	**0.900**	**83.6**	81.3	3.26
	PID-Expert Average	0.908	87.5	84.8	3.29
	PID-Expert Aggregation	**0.900**	84.0	82.0	**3.25**

Figure 10 further reveals the coverage–width trade-off at the rolling-window level. Across all four forecasting specifications, PID-Expert aggregation keeps the rolling coverage distribution centered close to the nominal level of 0.9, while its rolling mean widths are consistently lower than those of Conformal PID. The lower median rolling width supports the efficiency gains reported in Table 4. The three PID-Expert variants display a clear ordering: selection generally attains the narrowest intervals but shifts the rolling coverage distribution downward, whereas averaging shifts coverage upward at the cost of wider intervals. PID-Expert aggregation lies between these two extremes, retaining much of the width reduction from selection and generally raising rolling coverage, while producing narrower intervals than averaging. The same pattern across the four forecasting models highlights the stability of this coverage–efficiency trade-off across different underlying forecasts.

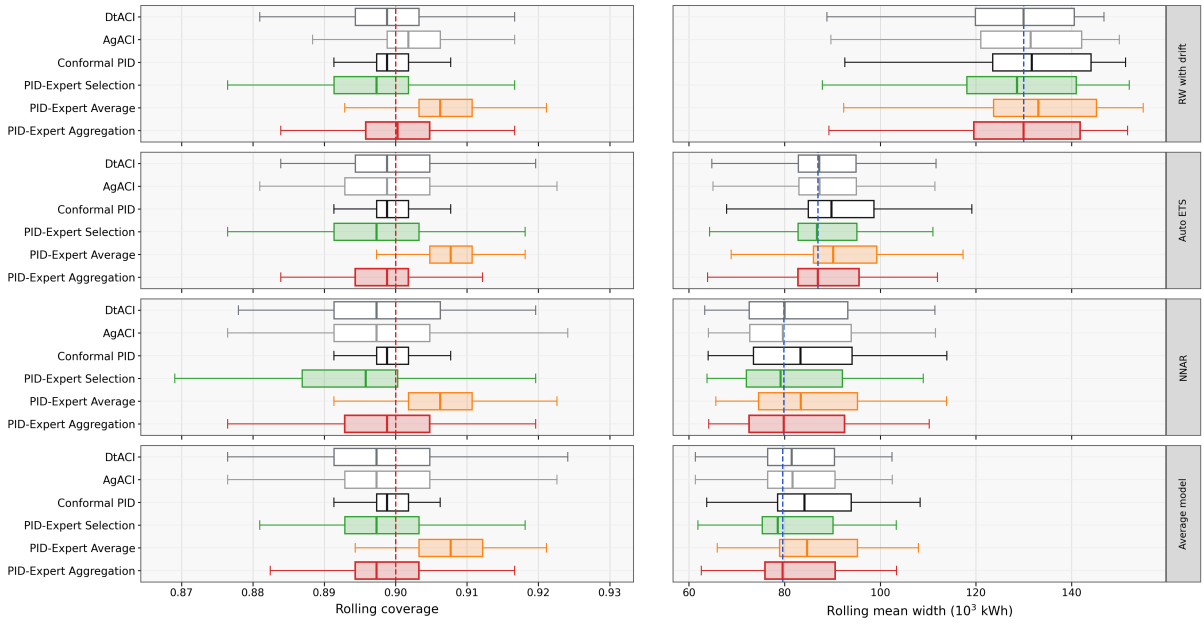


Figure 10: Rolling coverage and mean interval width for the Ordos electricity demand application, computed over windows of 672 observations. The red dashed lines indicate the nominal coverage level of 0.9, and the blue dashed lines indicate the median rolling mean width of PID-Expert aggregation. Interval widths are measured in 10^3 kWh.

6. Conclusions

In this paper, we propose PID-Expert aggregation, an online conformal framework that adaptively combines the score thresholds generated by a heterogeneous library of PID controllers. Rather than relying on a single prespecified controller, the proposed method updates expert weights through a loss that jointly accounts for normalized interval width and expert-level miscoverage. An aggregate coverage multiplier further adjusts the relative penalty on miscoverage in response to realized coverage errors. This threshold-level aggregation provides a direct mechanism for adapting the coverage–efficiency tradeoff while preserving the distinct calibration dynamics of the individual controllers.

The theoretical analysis characterizes both the aggregation and coverage properties of the method. Local regret bounds compare the weighted average of expert losses with any fixed expert over arbitrary time intervals and motivate practical reference choices for the learning rate and share parameter. Separately, the multiplier recursion yields a deterministic, pathwise upper bound on time-averaged aggregate miscoverage. Under a high-multiplier negative-drift condition, the multiplier grows sublinearly, leading to a nonasymptotic expected miscoverage bound of order $O(T^{-1/2})$ above the target and almost-sure asymptotic upper control.

Across four simulation settings and two real-world forecasting applications, PID-Expert consistently produces narrower mean prediction intervals than the prespecified Conformal PID benchmark while maintaining empirical coverage close to the nominal level. Its relative performance against DtACI and AgACI varies across settings, indicating that aggregating score thresholds and aggregating adaptive miscoverage levels offer complementary advan-

tages across different data regimes. The two additional PID-Expert benchmarks clarify the coverage–efficiency tradeoff. Selection generally produces narrower intervals with lower coverage, whereas equal averaging produces wider intervals with higher coverage. Aggregation retains much of the width reduction from selection while generally bringing empirical coverage closer to the nominal level, but it does not uniformly achieve the lowest MSIS or eliminate local coverage shortfalls. Overall, the results support online aggregation as a way to reduce reliance on a single PID configuration while balancing interval width and coverage.

The present study is limited to one-step-ahead symmetric intervals based on absolute-error scores and a fixed, finite controller library. Its performance therefore depends on whether the library contains controllers with suitable score scales and adaptation dynamics. Future work could extend the framework to asymmetric tail-specific intervals, horizon-dependent aggregation for multistep forecasting, and data-driven procedures for constructing or expanding the controller library online. Such extensions would broaden the applicability of PID-Expert aggregation while retaining its central principle of jointly adapting calibration and interval efficiency. Proofs and implementation details are given in the appendices. The implementation and available data are provided at https://github.com/Guodong-Liu/PID_Expert.

Acknowledgments

This research was supported by an IIF-SAS Grant, which promotes research in forecasting. Xiaoqian Wang is supported by the National Natural Science Foundation of China (No. 72601399) and the Presidential Foundation of the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (No. E555930101). Yanfei Kang is supported by the National Natural Science Foundation of China (No. 72571014).

Declaration of generative AI use

During the preparation of this work the authors used ChatGPT to improve English language and assist in coding to create graphical illustrations. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Angelopoulos, A. N. and Bates, S. (2023), 'Conformal prediction: A gentle introduction', *Foundations and Trends in Machine Learning* **16**(4), 494–591.

Angelopoulos, A. N., Candès, E. J. and Tibshirani, R. J. (2023), 'Conformal PID control for time series prediction', *Advances in Neural Information Processing Systems* **36**, 23047–23074.

Areces, F., Mohri, C., Hashimoto, T. and Duchi, J. (2025), Online conformal prediction via online optimization, *in* 'Proceedings of the 42nd International Conference on Machine Learning', Vol. 267 of *Proceedings of Machine Learning Research*, PMLR, pp. 1604–1649.

Assimakopoulos, V. and Nikolopoulos, K. (2000), 'The theta model: a decomposition approach to forecasting', *International Journal of Forecasting* **16**(4), 521–530.

Barber, R. F., Candès, E. J., Ramdas, A. and Tibshirani, R. J. (2023), 'Conformal prediction beyond exchangeability', *The Annals of Statistics* **51**(2), 816–845.

Chatfield, C. (1993), 'Calculating interval forecasts', *Journal of Business & Economic Statistics* **11**(2), 121–135.

Chernozhukov, V., Wüthrich, K. and Yinchu, Z. (2018), Exact and robust conformal inference methods for predictive machine learning with dependent data, *in* 'Conference on Learning Theory', PMLR, pp. 732–749.

Devine, M. T. and Bertsch, V. (2026), 'Analysing the interactions between demand side and supply side investment decisions in an oligopolistic electricity market using a stochastic equilibrium model', *European Journal of Operational Research* . In press.

Gibbs, I. and Candès, E. J. (2021), 'Adaptive conformal inference under distribution shift', *Advances in Neural Information Processing Systems* **34**, 1660–1672.

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

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007), 'Probabilistic forecasts, calibration and sharpness', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**(2), 243–268.

Gneiting, T. and Katzfuss, M. (2014), 'Probabilistic forecasting', *Annual Review of Statistics and Its Application* **1**(1), 125–151.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Hajihashemi, E. and Shen, Y. (2024), Multi-model ensemble conformal prediction in dynamic environments, *in* 'Advances in Neural Information Processing Systems', Vol. 37.

Herbster, M. and Warmuth, M. K. (1998), 'Tracking the best expert', *Machine Learning* **32**, 151–178.

Hu, D., Wu, J., Xia, S.-T. and Zou, C. (2026), Distribution-informed online conformal prediction, *in* 'International Conference on Learning Representations', Vol. 2026, pp. 58230–58255.

Hyndman, R. J. and Athanasopoulos, G. (2021), *Forecasting: Principles and Practice*, 3rd edn, OTexts.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. and Yasmeen, F. (2026), *forecast: Forecasting functions for time series and linear models*. R package version 9.0.0. https://pkg.robjhyndman.com/forecast/.

Kath, C. and Ziel, F. (2021), 'Conformal prediction interval estimation and applications to day-ahead and intraday power markets', *International Journal of Forecasting* **37**(2), 777–799.

Koo, B. and Seo, M. H. (2015), 'Structural-break models under mis-specification: Implications for forecasting', *Journal of Econometrics* **188**(1), 166–181.

Koo, B., Wong, B. and Zhong, Z.-Y. (2026), 'Disentangling structural breaks in factor models for macroeconomic data', *Journal of Business & Economic Statistics* . In press.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018), 'Distribution-free predictive inference for regression', *Journal of the American Statistical Association* **113**(523), 1094–1111.

Lipiecki, A., Bilińska, K., Kourentzes, N. and Weron, R. (2026), 'Stealing accuracy: Predicting day-ahead electricity prices with temporal hierarchy forecasting (THieF)', *International Journal of Forecasting* **42**(4), 1215–1225.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J. and Talagala, T. S. (2020), 'FFORMA: Feature-based forecast model averaging', *International Journal of Forecasting* **36**(1), 86–92.

Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002), Inductive confidence machines for regression, *in* 'European Conference on Machine Learning', Springer, pp. 345–356.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E. et al. (2022), 'Forecasting: theory and practice', *International Journal of Forecasting* **38**(3), 705–871.

Petropoulos, F., Hyndman, R. J. and Bergmeir, C. (2018), 'Exploring the sources of uncertainty: Why does bagging for time series forecasting work?', *European Journal of Operational Research* **268**(2), 545–554.

Schlembach, F., Smirnov, E., Koprinska, I. and Winands, M. H. M. (2025), 'Conformal multistep-ahead multivariate time-series forecasting', *Machine Learning* **114**(7), 165.

Shafer, G. and Vovk, V. (2008), 'A tutorial on conformal prediction', *Journal of Machine Learning Research* **9**(12), 371–421.

Shang, H. L. (2026), 'Conformal prediction for functional time series: application to age-specific mortality rates', *Journal of Population Research* **43**(2), 14.

Stankeviciute, K., Alaa, A. M. and van der Schaar, M. (2021), 'Conformal time-series forecasting', *Advances in Neural Information Processing Systems* **34**, 6216–6228.

Sun, S. and Yu, R. (2025), 'Conformal prediction for time-series forecasting with change points', *Advances in Neural Information Processing Systems* **38**, 12243–12272.

Tibshirani, R. J., Barber, R. F., Candès, E. J. and Ramdas, A. (2019), Conformal prediction under covariate shift, *in* 'Advances in Neural Information Processing Systems', Vol. 32, pp. 2530–2540.

Vovk, V., Gammerman, A. and Shafer, G. (2005), *Algorithmic Learning in a Random World*, Springer-Verlag.

Wang, X. and Hyndman, R. J. (2024), 'Online conformal inference for multi-step time series forecasting', *arXiv preprint arXiv:2410.13115* .

Wang, X., Hyndman, R. J., Li, F. and Kang, Y. (2023), 'Forecast combinations: An over 50-year review', *International Journal of Forecasting* **39**(4), 1518–1547.

Weron, R. (2014), 'Electricity price forecasting: A review of the state-of-the-art with a look into the future', *International Journal of Forecasting* **30**(4), 1030–1081.

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

Wintenberger, O. (2017), 'Optimal learning with bernstein online aggregation', *Machine Learning* **106**(1), 119–141.

Xu, C. and Xie, Y. (2021), Conformal prediction interval for dynamic time-series, *in* 'International Conference on Machine Learning', Vol. 139, PMLR, pp. 11559–11569.

Xu, C. and Xie, Y. (2023), Sequential predictive conformal inference for time series, *in* 'International Conference on Machine Learning', PMLR, pp. 38707–38727.

Yang, Z., Candès, E. and Lei, L. (2024), 'Bellman conformal inference: Calibrating prediction intervals for time series', *arXiv preprint arXiv:2402.05203* .

Zaffran, M., Féron, O., Goude, Y., Josse, J. and Dieuleveut, A. (2022), Adaptive conformal predictions for time series, *in* 'International Conference on Machine Learning', PMLR, pp. 25834–25866.

Appendix A. Proofs for the theoretical results

Appendix A.1. Fixed-share auxiliary results

Define the exponential normalizer

$$Z_t := \sum_{m=1}^{M} w_t^{(m)} \exp\{-\xi \ell_t^{(m)}\}.$$

Then

$$\widetilde{w}_{t+1}^{(m)} = \frac{w_t^{(m)} \exp\{-\xi \ell_t^{(m)}\}}{Z_t}.$$

Lemma 1 (One-step exponential-weights inequality)**.** *Under Assumption 1,*

$$\log Z_t \le -\xi \bar{\ell}_t + \xi^2 \bar{v}_t. \tag{A.1}$$

Proof. For $x \ge 0$, $\exp(-x) \le 1 - x + x^2$. Applying this inequality with $x = \xi \ell_t^{(m)}$, multiplying by $w_t^{(m)}$, and summing over m gives

$$Z_t \le 1 - \xi \bar{\ell}_t + \xi^2 \bar{v}_t.$$

The right-hand side is positive because it is at least $Z_t > 0$. Monotonicity of the logarithm and $\log(1+u) \le u$ yield Equation (A.1). □

Lemma 2 (Expert recoverability)**.** *Under Assumption 1, for every $t \ge 1$ and every expert m,*

$$w_t^{(m)} \ge \frac{\sigma}{M}.$$

Proof. At initialization, $w_1^{(m)} = 1/M \ge \sigma/M$. For $t \ge 1$, Equation (6) and $\widetilde{w}_{t+1}^{(m)} \ge 0$ imply $w_{t+1}^{(m)} \ge \sigma/M$. □

Lemma 3 (Local comparison with a fixed expert)**.** *Under Assumption 1, for every interval $\mathcal{I} = [r, s]$ and every expert m,*

$$\sum_{t=r}^{s} \log Z_t \ge \log \frac{\sigma}{M} + L \log(1-\sigma) - \xi \sum_{t=r}^{s} \ell_t^{(m)}. \tag{A.2}$$

Proof. The fixed-share update satisfies

$$\begin{aligned} w_{t+1}^{(m)} &\ge (1-\sigma) \widetilde{w}_{t+1}^{(m)} \\ &= (1-\sigma) \frac{w_t^{(m)} \exp\{-\xi \ell_t^{(m)}\}}{Z_t}. \end{aligned}$$

Consequently,

$$Z_t \ge (1-\sigma) \frac{w_t^{(m)}}{w_{t+1}^{(m)}} \exp\{-\xi \ell_t^{(m)}\}.$$

Multiplying from $t = r$ to s telescopes the weight ratios and gives

$$\prod_{t=r}^{s} Z_t \geq (1-\sigma)^L \frac{w_r^{(m)}}{w_{s+1}^{(m)}} \exp\left\{-\xi \sum_{t=r}^{s} \ell_t^{(m)}\right\}.$$

Because $w_{s+1}^{(m)} \leq 1$ and, by Lemma 2, $w_r^{(m)} \geq \sigma/M$, taking logarithms proves Equation (A.2). □

Appendix A.2. Proofs of the regret and parameter results

Proof of Theorem 1. Summing Equation (A.1) over $\mathcal{I}$ and combining it with Equation (A.2) yields

$$\log \frac{\sigma}{M} + L \log(1-\sigma) - \xi \sum_{t=r}^{s} \ell_t^{(m)} \leq -\xi \sum_{t=r}^{s} \bar{\ell}_t + \xi^2 \sum_{t=r}^{s} \bar{v}_t.$$

Rearranging and dividing by ξ proves Equation (7). Since $-\log(1-\sigma) \leq 2\sigma$ for $0 < \sigma \leq 1/2$, Equation (8) follows. □

Proof of Corollary 1. For fixed ξ, the relevant term is $g(\sigma) = \log(M/\sigma) + 2\sigma L$. It is strictly convex because $g''(\sigma) = 1/\sigma^2 > 0$, and $g'(\sigma) = -1/\sigma + 2L$ vanishes at $\sigma = 1/(2L)$. For $L \geq 1$, this value satisfies $0 < \sigma \leq 1/2$. □

Reference calculation of the learning rate. After substituting $\sigma = 1/(2L)$ into Equation (8), hold $S_{\mathcal{I}} > 0$ fixed and consider the sum

$$\xi S_{\mathcal{I}} + \frac{\log(2ML) + 1}{\xi}.$$

As a function of $\xi > 0$ with $S_{\mathcal{I}}$ fixed, its first and second derivatives are

$$S_{\mathcal{I}} - \frac{\log(2ML) + 1}{\xi^2}, \qquad \frac{2\{\log(2ML) + 1\}}{\xi^3} > 0,$$

respectively. Its unique minimum occurs at Equation (9). In an actual run, the learning rate affects the realized weights and aggregate coverage feedback, and hence $S_{\mathcal{I}}$. The minimization therefore treats $S_{\mathcal{I}}$ as fixed. Replacing $V_{\mathcal{I}}$ in Equation (9) with the prespecified scale V gives the learning-rate rule in Equation (10).

Appendix A.3. Comparison with an external benchmark sequence

The fixed-expert comparison extends naturally to an external benchmark sequence when the library contains a fixed expert that approximates that benchmark well over the interval of interest. We formalize this requirement as follows.

Assumption 3 (Expert-library approximation)**.** *For a benchmark sequence with finite losses* $\{\ell_t^\star\}_{t\in\mathcal{I}}$*, suppose that the library contains an expert* $m_\mathcal{I}$ *satisfying*

$$\sum_{t=r}^{s} \ell_t^{(m_\mathcal{I})} - \sum_{t=r}^{s} \ell_t^\star \le \Delta_{\text{grid}}(\mathcal{I}), \qquad 0 \le \Delta_{\text{grid}}(\mathcal{I}) < \infty. \tag{A.3}$$

Corollary 3 (Local comparison with a target sequence)**.** *Under Assumptions 1 and 3,*

$$\sum_{t=r}^{s} \bar{\ell}_t - \sum_{t=r}^{s} \ell_t^\star \le \xi \sum_{t=r}^{s} \bar{v}_t + \frac{\log(M/\sigma) + 2\sigma L}{\xi} + \Delta_{\text{grid}}(\mathcal{I}). \tag{A.4}$$

The corollary separates aggregation regret from library approximation error. Since $\Delta_{\text{grid}}(\mathcal{I})$ is not controlled here as a function of library resolution, the result is an error decomposition rather than a quantitative approximation guarantee.

Proof of Corollary 3. Apply Theorem 1 with $m = m_\mathcal{I}$ and add Equation (A.3). The intermediate loss of $m_\mathcal{I}$ cancels, yielding Equation (A.4). □

Appendix A.4. An idealized reference-distribution derivation of the second-moment scale

To motivate a prespecified second-moment scale for the coverage–efficiency loss, we consider an idealized population reference problem. Let $S \ge 0$ have a continuous distribution function F and density f, with $(1-\alpha)$-quantile $0 < Q = F^{-1}(1-\alpha) < \infty$. Assume that F is differentiable at Q, with derivative $f(Q) > 0$. Consider the population width minimization problem

$$\min_{q\ge 0} \frac{2q}{B_q} \quad \text{subject to} \quad \Pr(S > q) \le \alpha.$$

At the calibrated solution $q = Q$, with the ideal normalization $B_q = 2Q$, the normalized width equals one. The Lagrangian is

$$\mathcal{L}(q, \lambda) = \frac{2q}{B_q} + \lambda\{\Pr(S > q) - \alpha\}.$$

Since $Q > 0$, stationarity with respect to q at Q gives

$$\lambda^\star = \frac{2}{B_q f(Q)} = \frac{1}{Q f(Q)}.$$

We use the standardized reference score $U = S/Q$ and take $U \sim \text{Unif}(0, 1/(1-\alpha))$. This reference law satisfies $\Pr(U \le 1) = 1-\alpha$ and has $f_U(1) = 1-\alpha$, so

$$\lambda^\star = \frac{1}{f_U(1)} = \frac{1}{1-\alpha}.$$

For a reference calculation over a single update, set every expert threshold, and hence the aggregate threshold, equal to Q, and set the multiplier before the update equal to

$\lambda^\star$. All expert miscoverage indicators then coincide with the aggregate indicator, with miscoverage probability α.

As in Equations (2) and (4), the current aggregate error updates the multiplier before the expert loss is evaluated. With $B_q = 2Q$, a covered observation therefore has loss 1, while a miscovered observation has loss

$$1 + \lambda^\star + \nu(1 - \alpha).$$

Since these events have probabilities $1 - \alpha$ and α, respectively, the resulting reference second moment is

$$V_{\text{ref}} = (1 - \alpha) + \alpha \left[1 + \frac{1}{1 - \alpha} + \nu(1 - \alpha)\right]^2.$$

For the experiments, $\alpha = 0.1$ and $\nu = 0.02$, yielding $V_{\text{ref}} \approx 1.3533$; before evaluation, we set V in Equation (10) to this rounded value. The calculation describes a population reference state and a single update. It makes no claim that the online multiplier converges to $\lambda^\star$ or that V_{ref} is a stationary second moment of the algorithm. The resulting V serves as a reference tuning scale, with no claim that it estimates or bounds the realized $V_{\mathcal{I}}$ or yields optimal online tuning.

Appendix A.5. Proof of the pathwise aggregate miscoverage bound

Proof of Proposition 1. Projection onto the nonnegative half-line implies

$$\lambda_t \geq \lambda_{t-1} + \nu(err_t - \alpha).$$

Hence $err_t - \alpha \leq (\lambda_t - \lambda_{t-1})/\nu$. Summing from $t = 1$ to T telescopes and gives Equation (11). Taking the upper limit along any realized path for which $\lambda_T/T \to 0$ proves Equation (12). □

Appendix A.6. Proof of stochastic multiplier sublinearity

Proof of Proposition 2. Because $[x]_+^2 \leq x^2$, the multiplier update satisfies

$$\begin{aligned}\lambda_t^2 &\leq \{\lambda_{t-1} + \nu(err_t - \alpha)\}^2 \\ &= \lambda_{t-1}^2 + 2\nu\lambda_{t-1}(err_t - \alpha) + \nu^2(err_t - \alpha)^2.\end{aligned}$$

Since $err_t \in \{0, 1\}$ and $0 < \alpha < 1$, we have $(err_t - \alpha)^2 \leq 1$. Using the $\mathcal{F}_{t-1}$-measurability of λ_{t-1} gives

$$\mathbb{E}[\lambda_t^2 \mid \mathcal{F}_{t-1}] \leq \lambda_{t-1}^2 + 2\nu\lambda_{t-1}\mathbb{E}[err_t - \alpha \mid \mathcal{F}_{t-1}] + \nu^2.$$

We bound the conditional drift term in two cases. If $\lambda_{t-1} < \Lambda$, then

$$2\nu\lambda_{t-1}\mathbb{E}[err_t - \alpha \mid \mathcal{F}_{t-1}] \leq 2\nu\Lambda(1 - \alpha).$$

If $\lambda_{t-1} \geq \Lambda$, then Equation (13) gives

$$2\nu\lambda_{t-1}\mathbb{E}[err_t - \alpha \mid \mathcal{F}_{t-1}] \leq -2\nu\delta\lambda_{t-1} \leq 0.$$

Therefore, in both cases,

$$\mathbb{E}[\lambda_t^2 \mid \mathcal{F}_{t-1}] \leq \lambda_{t-1}^2 + C_\lambda.$$

Taking expectations and iterating from $t = 1$ to T yields

$$\mathbb{E}[\lambda_T^2] \leq \lambda_0^2 + C_\lambda T.$$

Dividing by T^2 proves Equation (15). Mean-square convergence implies convergence in probability.

To establish almost-sure sublinearity, apply Markov's inequality along the subsequence $T = k^2$. For every $\varepsilon > 0$,

$$\sum_{k=1}^{\infty} \Pr\left(\frac{\lambda_{k^2}}{k^2} > \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{k=1}^{\infty} \left(\frac{\lambda_0^2}{k^4} + \frac{C_\lambda}{k^2}\right) < \infty.$$

The first Borel–Cantelli lemma, applied to $\varepsilon = 1/j$ for positive integers j, yields $\lambda_{k^2}/k^2 \to 0$ almost surely. The multiplier update also gives $|\lambda_t - \lambda_{t-1}| \leq \nu$. Hence, for $k^2 \leq T < (k+1)^2$,

$$0 \leq \frac{\lambda_T}{T} \leq \frac{\lambda_{k^2} + \nu(2k+1)}{k^2} \longrightarrow 0 \qquad \text{almost surely.}$$

This proves the almost-sure conclusion stated after the proposition. □

Appendix A.7. Proof of expected aggregate miscoverage

Proof of Corollary 2. Taking expectations in Equation (11), using $\lambda_0 \geq 0$, and then applying the Cauchy–Schwarz inequality gives

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} err_t\right] &\leq \alpha + \frac{\mathbb{E}[\lambda_T] - \lambda_0}{\nu T} \\
&\leq \alpha + \frac{\mathbb{E}[\lambda_T]}{\nu T} \\
&\leq \alpha + \frac{\sqrt{\mathbb{E}[\lambda_T^2]}}{\nu T}.
\end{aligned}$$

Substituting Equation (14) proves Equation (16). □

Appendix B. PID expert libraries

Appendix B.1. Reference saturation constant

We construct the PID expert libraries around the reference Conformal PID configuration. The reference saturation constant is determined by the logarithmic rule

$$C^{\star}_{\text{sat}}(T_{\text{ref}}) = \frac{2}{\pi}\left\{\lceil \delta_{\text{sat}} \log T_{\text{ref}} \rceil - \frac{1}{\log T_{\text{ref}}}\right\}, \qquad \delta_{\text{sat}} = 0.01. \tag{B.1}$$

The logarithm in Equation (B.1) is natural, $\lceil \cdot \rceil$ denotes the ceiling function, and $T_{\text{ref}} > 1$ is the horizon used to construct the reference controller. The subscript in δ_{sat} distinguishes this numerical saturation constant.

Appendix B.2. Nonlinear simulation library

For the nonlinear simulation, we set $T_{\text{ref}} = 1000$, giving

$$C^{\star}_{\text{sat}}(1000) = 0.5444596\ldots.$$

The library comprises all $M = 7^3 = 343$ combinations of the following parameter values:

$$\begin{aligned}
\eta &\in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.2\},\\
K_I &\in \{0.1, 0.25, 0.4, 0.5, 0.75, 1.0, 1.5\},\\
C_{\text{sat}} &\in \{0.2, 0.3, 0.4, C^{\star}_{\text{sat}}(1000), 0.7, 0.85, 1.0\}.
\end{aligned}$$

The reference saturation constant is reported as 0.544 when rounded.

Appendix B.3. Distribution-shift simulation library

All three distribution-shift scenarios use the same PID expert library, constructed with reference sequence length $T_{\text{ref}} = 6000$. The resulting reference saturation constant is

$$C^{\star}_{\text{sat}}(6000) = 0.5634410\ldots.$$

The library consists of all $M = 7^3 = 343$ combinations of the following parameter values:

$$\begin{aligned}
\eta &\in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.2\},\\
K_I &\in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0\},\\
C_{\text{sat}} &\in \{0.2, 0.3, 0.4, 0.56, 0.7, 0.85, 1.0\}.
\end{aligned}$$

The grid uses the rounded reference value 0.56, whereas the Conformal PID benchmark uses $C^{\star}_{\text{sat}}(6000)$ without rounding, together with $\eta = 0.1$ and $K_I = 2$.

Appendix B.4. Real-data expert libraries

For each forecasting model, let T_s be the number of nonconformity scores available for calibration and L_0 the length of the initial score window. We define the reference scales

$$B' = \max_{1 \le t \le \min(L_0, T_s)} |s_t|, \qquad C^\star_{\text{sat}} = C^\star_{\text{sat}}(T_s).$$

In both applications, $T_s \ge L_0$ and the scores are absolute forecast errors, so $B' = \max_{1 \le t \le L_0} s_t$.

The Ordos energy consumption and Capital Bikeshare applications each use all $M = 9^3 = 729$ parameter combinations below:

$$\begin{aligned} \eta &\in \{0.02, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.2, 0.25\}, \\ K_I &\in B' \times \{0.125, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 4\}, \\ C_{\text{sat}} &\in C^\star_{\text{sat}} \times \{0.25, 0.35, 0.5, 0.75, 1, 1.25, 1.5, 2, 2.5\}. \end{aligned}$$

The Conformal PID benchmark uses the same reference scales:

$$\eta = 0.1, \qquad K_I = B', \qquad C_{\text{sat}} = C^\star_{\text{sat}}.$$

Appendix C. DtACI and AgACI configurations

The ACI benchmarks aggregate adaptive miscoverage levels before converting them to thresholds. AgACI here denotes an adaptation to calibration levels; the original method aggregates interval endpoints separately (Zaffran et al., 2022).

DtACI and AgACI share a library of nine ACI experts, distinguished by their step sizes:

$$\gamma \in \{0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128, 0.256\}, \qquad M_{\text{ACI}} = 9.$$

For DtACI, the share parameter and learning rate are set to

$$\sigma_{\text{DtACI}} = \frac{1}{2L}, \qquad \eta_{\text{DtACI}} = \sqrt{\frac{\log(2M_{\text{ACI}}L) + 1}{LV_{\text{pin}}}}, \qquad V_{\text{pin}} = \frac{\alpha^2(1-\alpha)^2}{3},$$

where L is the local adaptation window length. AgACI uses adaptive learning rates for individual experts, with stabilization constant $\varepsilon_{\text{AgACI}} = 0.001$.

For the nonlinear simulation and real-data applications, both methods use the rolling calibration window

$$\mathcal{W}_t = \{\max(1, t-L), \ldots, t-1\}.$$

Using only preceding scores, the empirical upper-tail probability of the observed s_t is

$$\beta_t = \frac{1 + \sum_{r \in \mathcal{W}_t} \mathbf{1}\{s_r \ge s_t\}}{|\mathcal{W}_t| + 1}.$$

The aggregate level $\overline{\alpha}_t$, chosen before observing s_t, determines the reported threshold:

$$q_t = \widehat{Q}_{1-\overline{\alpha}_t}\big(\{s_r : r \in \mathcal{W}_t\}\big),$$

where $\widehat{Q}$ is the higher empirical quantile. We set $q_t = 0$ when $\mathcal{W}_t$ is empty.

For the Gaussian distribution-shift simulations, DtACI and AgACI use a fixed standard normal reference to relate the observed nonconformity score to a tail probability and the adaptive miscoverage level to an interval threshold:

$$\beta_t = 2\{1 - \Phi(|Y_t|)\}, \qquad q_t = \Phi^{-1}\left(1 - \frac{\overline{\alpha}_t}{2}\right),$$

where Φ is the standard normal distribution function and $\overline{\alpha}_t$ is determined before observing Y_t. The reference distribution remains fixed across scenarios; adaptation operates through $\overline{\alpha}_t$, without using the time-varying mean μ_t.