Online conformal inference for multi-step time series forecasting

Xiaoqian Wang*
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
and
Rob J Hyndman
Department of Econometrics & Business Statistics, Monash University

2 February 2026

Abstract

We consider the problem of constructing distribution-free prediction intervals for multi-step time series forecasting, with a focus on the temporal dependencies inherent in multi-step forecast errors. We establish that the optimal h-step-ahead forecast errors exhibit serial correlation up to lag $(h-1)$ under a general non-stationary autoregressive data generating process. To leverage these properties, we propose the Autocorrelated Multi-step Conformal Prediction (AcMCP) method, which effectively incorporates autocorrelations in multi-step forecast errors, resulting in more statistically efficient prediction intervals. This method guarantees asymptotic marginal coverage for multi-step prediction intervals, though we note that, for finite samples, the coverage error admits an upper bound that increases with the forecasting horizon. Additionally, we extend several easy-to-implement conformal prediction methods, originally designed for single-step forecasting, to accommodate multi-step scenarios. Through empirical evaluations, including simulations and applications to data, we demonstrate that AcMCP achieves coverage that closely aligns with the target within local windows, while providing adaptive prediction intervals that effectively respond to varying conditions.

Keywords: Conformal prediction, Coverage guarantee, Distribution-free inference, Exchangeability, Weighted quantile estimate

*Corresponding author. Xiaoqian Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China. E-mail address: xiaoqian.wang@amss.ac.cn (X. Wang).

1 Introduction

Conformal prediction (Papadopoulos et al. 2002, Vovk et al. 2005) is a simple yet powerful framework for uncertainty quantification. It constructs valid prediction intervals that achieve nominal coverage without imposing stringent assumptions on the data generating distribution, other than requiring the data to be i.i.d. or, more generally, exchangeable. Its credibility and potential make it widely used to quantify uncertainty for predictions produced by black-box machine learning models (Papadopoulos et al. 2007, Papadopoulos 2008, Shafer & Vovk 2008, Barber et al. 2021) or non-parametric models (Lei & Wasserman 2014).

Three widely used classes of conformal prediction methods for constructing distribution-free prediction intervals are split conformal prediction (Vovk et al. 2005), full conformal prediction (Vovk et al. 2005), and jackknife+ (Barber et al. 2021). Split conformal, which relies on a holdout set, offers computational efficiency but sacrifices some statistical efficiency due to data splitting. Full conformal prediction avoids data splitting, providing higher accuracy at the cost of increased computational complexity. Both split and full conformal prediction methods guarantee coverage at the target level under the assumption of data exchangeability. Jackknife+ strikes a balance between these methods, offering a compromise between statistical precision and computational demands. Under assumptions of algorithmic stability, Jackknife+ attains near-nominal coverage, while under the weaker assumption of exchangeability alone, it guarantees coverage of at least $1-2\alpha$ in the worst case. Gupta et al. (2022) further generalize jackknife+ and related out-of-bag conformal methods through a unified framework based on nested families of prediction sets.

Nevertheless, the data exchangeability assumption is often violated in many applied domains, where challenges such as non-stationarity, distributional drift, temporal and spatial dependencies are prevalent. In response, several extensions to conformal prediction have been proposed to handle non-exchangeable data. Notable examples include methods for handling covariate shift (Tibshirani et al. 2019, Yang, Kuchibhotla & Tchetgen Tchetgen 2024), online distribution shift (Gibbs &

Candès 2021, 2024, Zaffran et al. 2022, Bastani et al. 2022), time series data (Chernozhukov et al. 2018, Gibbs & Candès 2021, Xu & Xie 2021, 2023, Zaffran et al. 2022), and methods based on certain distributional assumptions of the data rather than exchangeability (Chernozhukov et al. 2018, Oliveira et al. 2024, Xu & Xie 2021, 2023). Additionally, some methods propose weighting nonconformity scores (e.g., prediction errors) differently, either using non-data-dependent weights (Barber et al. 2023) or weights based on observed feature values (Tibshirani et al. 2019, Guan 2023).

Recent work has sought to extend conformal prediction to time series settings, where exchangeability obviously fails due to inherent temporal dependencies. One line of research has focused on developing conformal-type methods that offer coverage guarantees under certain relaxations of exchangeability. For example, within the full conformal prediction framework, Chernozhukov et al. (2018) and Yu et al. (2022) construct prediction sets for time series by using a group of permutations that are specifically designed to preserve the dependence structure in the data, ensuring validity under weak assumptions on the nonconformity score. In the split conformal prediction framework, Xu & Xie (2021) and Xu & Xie (2023) extend conformal prediction methods to time series settings and establish asymptotically valid conditional coverage under the assumption that model errors are stationary and strongly mixing. Barber et al. (2023) use weighted residual distributions to provide robustness against distribution drift. Additionally, Oliveira et al. (2024) introduce a general framework based on concentration inequalities and data decoupling properties of the data to retain asymptotic coverage guarantees across several dependent data settings.

In a separate strand of research, Gibbs & Candès (2021) develop adaptive conformal inference (ACI, denoted as ACP hereafter) in an online manner to manage temporal distribution shifts and ensure long-run coverage guarantees. The basic idea is to adapt the miscoverage rate, α, based on historical miscoverage frequencies. However, ACP may yield infinite or empty prediction intervals when the α drifts below 0 or exceeds 1, respectively. Follow-up work has refined this idea by introducing time-dependent step sizes to respond to arbitrary distribution shifts, as seen in

studies by Bastani et al. (2022), Zaffran et al. (2022), and Gibbs & Candès (2024). Recent research has proposed a generalized updating process that tracks the quantile of the nonconformity score sequence, rather than the miscoverage rate, as discussed by Bhatnagar et al. (2023), Angelopoulos et al. (2023), and Angelopoulos et al. (2024).

Existing conformal prediction methods for time series primarily focus on single-step forecasting, even though many applications require reliable uncertainty quantification over multiple future horizons. The literature on multi-step conformal prediction is comparatively limited and, when it does consider multi-step forecasts, it often reduces the problem to H horizon-specific tasks by constructing separate prediction sets for each y_{t+h}, $h \in [H]$, without considering how forecasts at different horizons are related. For example, Stankeviciute et al. (2021) integrate conformal prediction with recurrent neural networks for multi-step forecasting and then apply Bonferroni correction to control coverage. This approach, however, assumes data independence, which is often unrealistic for time series. Yang, Candès & Lei (2024) propose Bellman conformal inference to jointly control multi-step miscoverage by minimizing a loss that balances average interval length across horizons and miscoverage, which however ignores temporal dependencies across horizons and can be computationally intensive due to related optimization at each time step. Related extensions to multivariate targets have also been studied; see, for example, Schlembach et al. (2025).

We employ a unified notation to formalize the mathematical representation of conformal prediction for time series data. We consider a general sequential setting in which we observe a time series $\{y_t\}_{t\geq 1}$ generated by an unknown data generating process (DGP), which may depend on its own past, along with other exogenous predictors, $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{p,t})'$, and their histories. The joint distribution of $\{(\boldsymbol{x}_t, y_t)\}_{t\geq 1}$, where $\boldsymbol{x}_t \in \mathbb{R}^p$ and $y_t \in \mathbb{R}$, is allowed to vary over time, thereby accommodating non-stationary processes. At each time point t, we aim to forecast H steps into the future, providing a *prediction set* (which is a prediction interval in this setting), $\hat{\mathscr{C}}_{t+h|t}$, for the realization y_{t+h} for each $h \in [H]$. The h-step-ahead forecast uses the previously observed data

$\{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq t}$ along with the new information of the exogenous predictors $\{\boldsymbol{x}_{t+j}\}_{1 \leq j \leq h}$. Note that we can generate ex-ante forecasts by using forecasts of the predictors based on information available up to and including time t. Alternatively, ex-post forecasts are generated assuming that actual values of the predictors from the forecast period are available. Given a nominal *miscoverage rate* $\alpha \in (0,1)$ specified by the user, we expect to construct prediction intervals $\hat{\mathscr{C}}_{t+h|t}$ that achieve long-run coverage guarantees, in the sense that $\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left\{y_{t+h} \in \hat{\mathscr{C}}_{t+h|t}\right\} \geq 1 - \alpha$.

Our goal is to achieve long-run coverage for multi-step univariate time series forecasting. All the proposed methods are grounded in the split conformal prediction framework and an online learning scheme, which are well-suited to the sequential nature of time series data. First, we extend several widely used conformal prediction methods that are originally designed for single-step forecasting to the multi-step setting by constructing horizon-specific prediction intervals. These extensions follow the common practice of treating each horizon independently and, in general, do not provide theoretical long-run coverage guarantees for multi-step prediction intervals, except for the proposed multi-step conformal PID control (MPID) method, which admits such a guarantee. Second, we provide theoretical results showing that, under a general non-stationary autoregressive DGP, the forecast errors of optimal h-step-ahead forecasts can be well approximated by a linear combination of at most its lag $(h-1)$ with respect to forecast horizon. Third, building on the theoretical results, we introduce the autocorrelated multi-step conformal prediction (AcMCP) method, which explicitly accounts for the autocorrelations of multi-step forecast errors at the calibration stage by preserving the dependence structure among nonconformity scores. AcMCP is proven to guarantee asymptotic marginal coverage for multi-step prediction intervals, with the coverage gap converging to zero as the sample size increases, without making any assumptions on data distribution shifts. Our method targets pointwise prediction intervals for each specific forecast horizon, $h \in [H]$. Finally, we illustrate the practical utility of these proposed methods through two simulations and two applications to electricity demand and eating-out expenditure forecasting. The results show that the proposed AcMCP method adapts effectively to changes in

the observed data distribution and, among the proposed extensions that achieve coverage close to the target level, generally yields more informative and narrower prediction intervals, particularly for larger forecast horizons.

We developed the `conformalForecast` package for R to implement the proposed multi-step conformal prediction methods, the package is publicly available on CRAN. All the data and code to reproduce the experiments are made available at https://github.com/xqnwang/cpts.

2 Online learning with sequential splits

Let $z_t = (\boldsymbol{x}_t, y_t)$ denote the data point (including the response y_t and possibly predictors $\boldsymbol{x}_t$) at time t. Suppose that, at each time t, a forecasting model $\hat{f}_t$ is trained on historical data $z_{1:t}$ and updated (re-trained) as new observations arrive. We assume that the predictors are known into the future, corresponding to an ex-post forecasting setting commonly used when predictors are deterministic, policy-driven, or externally specified. By conditioning on future predictors, we focus exclusively on the uncertainty in the response variable y_t and avoid introducing additional uncertainty from forecasting the exogenous inputs. Using the forecasting model $\hat{f}_t$, we are able to produce H-step point forecasts, $\{\hat{y}_{t+h|t}\}_{h\in[H]}$, using the future values for the predictors. We define the *nonconformity score* as the (signed) forecast error

$$s_{t+h|t} = \mathscr{S}(z_{1:t}, y_{t+h}) := y_{t+h} - \hat{f}_t(z_{1:t}, \boldsymbol{x}_{(t+1):(t+h)}) = y_{t+h} - \hat{y}_{t+h|t}.$$

The task is to employ conformal inference to build H-step prediction intervals, $\{\hat{\mathscr{C}}^{\alpha}_{t+h|t}(z_{1:t}, \boldsymbol{x}_{(t+1):(t+h)})\}_{h\in[H]}$, at the target coverage level $1-\alpha$. For brevity, we will use $\hat{\mathscr{C}}^{\alpha}_{t+h|t}$ to denote the h-step-ahead $100(1-\alpha)\%$ prediction interval.

Sequential split. In a time series context, it is inappropriate to perform *random splitting*, a standard strategy in much of the conformal prediction literature, due to the temporal dependency present in the data. Following related work such as Wisniewski et al. (2020) and Zaffran et al. (2022), throughout the conformal prediction methods proposed in this paper, we use a *sequential*

split to preserve the temporal structure. For example, the t available data points, $z_{1:t}$, are sequentially split into two consecutive sets, a *proper training set* $\mathscr{D}_{\text{tr}} \subset \{1,\ldots,t_r\}$ and a *calibration set* $\mathscr{D}_{\text{cal}} \subset \{t_r+1,\ldots,t\}$, where $t_c = t - t_r \gg H$, $\mathscr{D}_{\text{tr}} \cup \mathscr{D}_{\text{cal}} = \{1,\ldots,t\}$, and $\mathscr{D}_{\text{tr}} \cap \mathscr{D}_{\text{cal}} = \varnothing$. Here, "proper" means that the training set is used exclusively for fitting the model, with no overlap into the calibration set, which is essential for ensuring the validity of coverage in conformal prediction (Papadopoulos et al. 2002, Vovk et al. 2005). With sequential splitting, multiple H-step forecasts and their respective nonconformity scores can be computed on the calibration set.

Online learning. We will adapt the following generic online learning framework, which is closely related to Zaffran et al. (2022), for all conformal prediction methods to be discussed in later sections. The entire procedure for the online learning framework with sequential splits is also illustrated in Figure 1. This framework updates prediction intervals as new data points arrive, allowing us to assess their long-run coverage behavior. It adopts a standard rolling window evaluation strategy, although expanding windows could easily be used instead.

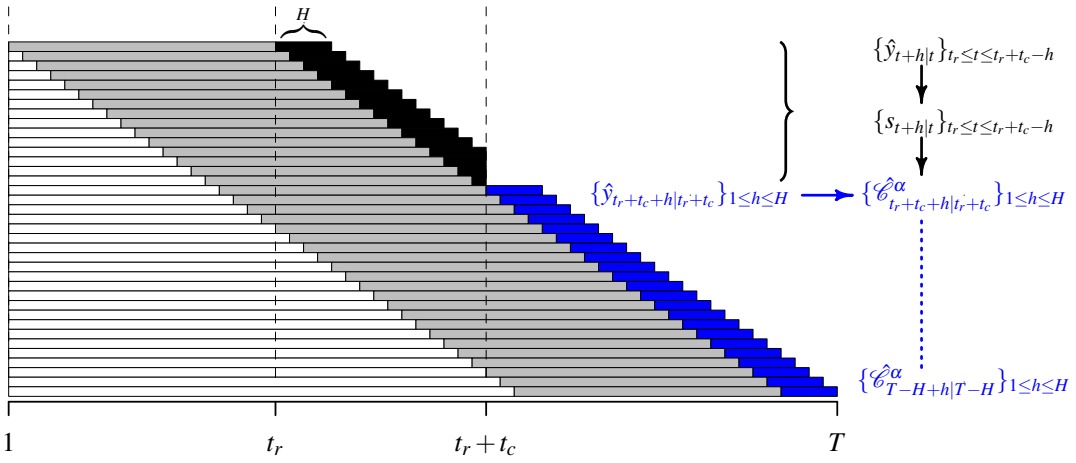

Figure 1: *Online learning framework with sequential splits. White: unused data; Gray: training data; Black: forecasts in calibration set; Blue: forecasts in test set.*

1. *Initialization*. Train a forecasting model on the initial proper training set $z_{(1+t-t_r):t}$, setting $t = t_r$. Then generate H-step-ahead point forecasts $\{\hat{y}_{t+h|t}\}_{h\in[H]}$ and compute the corresponding nonconformity scores $\{s_{t+h|t} = \mathscr{S}(z_{(1+t-t_r):t}, y_{t+h})\}_{h\in[H]}$ based on the true values H steps ahead, i.e. $\{y_{t+h}\}_{h\in[H]}$.

2. *Recurring procedure*. Roll the training set forward by one data point by setting $t \to t+1$. Then repeat step 1 until the nonconformity scores on the entire initial calibration set, $\{s_{t+h|t}\}_{t_r \le t \le t_r+t_c-h}$ for $h \in [H]$, are computed.
3. *Quantile estimation and prediction interval calculation*. Use nonconformity scores obtained from the calibration set to perform quantile estimation and compute H-step-ahead prediction intervals on the test set.
4. *Online updating*. Continuously roll the training set and calibration set forward, one data point at a time, to update the nonconformity scores for calibration, and then repeat step 3 until prediction intervals for the entire test set are obtained. That is, $\{\hat{\mathscr{C}}^{\alpha}_{t+h|t}\}_{t_r+t_c \le t \le T-H}$ for $h \in [H]$, where $T-t_r-t_c$ is the length of the test set used for testing coverage. Our goal is to achieve long-run coverage in time.

3 Extending conformal prediction for multi-step forecasting

In this section, we apply the online learning framework outlined in Section 2 to extend several popular conformal prediction methods to multi-step forecasting. In time series forecasting, when a model is properly specified and well-trained, the forecasts that minimize the mean squared error can be considered optimal in the sense of achieving the lowest possible expected squared forecast error. A key property of optimal forecast errors, which holds generally for linear models, is that the variance of the forecast error $e_{t+h|t}$ is non-decreasing in h (Diebold & Lopez 1996, Patton & Timmermann 2007). Therefore, a horizon-specific conformal prediction procedure is required for each $h \in [H]$. We note that the extensions of existing approaches introduced in this section do not provide theoretical long-run coverage guarantees, except for the multi-step conformal PID control (MPID) method, which admits such a guarantee.

3.1 Online multi-step split conformal prediction

Split conformal prediction (SCP, also called inductive conformal prediction, Papadopoulos et al. 2002, Vovk et al. 2005), is a holdout method for building prediction intervals using a pre-trained model on a training set. A key advantage of SCP is its ability to guarantee coverage by assuming data exchangeability. In regression setting, SCP randomly separates the available n data points, $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$, into a proper training set of size n_t and a calibration set of size n_c. Given a regression model $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ fitted on the training set and a score function $\mathscr{S}$, nonconformity scores $s_i = \mathscr{S}(X_i, Y_i)$ are computed on the calibration set to measure the nonconformity between the calibration's response values and the predicted values obtained from the fitted model $\hat{\mu}$. Then SCP computes the prediction interval for the test data Y_{n+1} using

$$\hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1}) = \left\{ y \in \mathbb{R} : \mathscr{S}(X_{n+1}, y) \leq Q_{1-\alpha}\left(\sum_{i \in \mathscr{D}_{\text{cal}}} \frac{1}{n_c + 1} \cdot \delta_{s_i} + \frac{1}{n_c + 1} \cdot \delta_{+\infty} \right) \right\},$$

where $\mathrm{Q}_\tau(\cdot)$ denotes the τ-quantile of its argument, and δ_a denotes the point mass at a. Time series data are inherently nonexchangeable due to their temporal dependence and autocorrelation. Therefore, directly applying SCP to time series data would violate the method's exchangeability assumption and compromise its coverage guarantee.

Here we introduce online **multi-step split conformal prediction** (MSCP) as a generalization of SCP to recursively update all h-step-ahead prediction intervals over time. MSCP applies conformal inference in an online fashion, updating prediction intervals as new data points are received. Specifically, for each $h \in [H]$, we consider the following simple online update to construct prediction intervals on the test set:

$$\hat{\mathscr{C}}^{\alpha}_{t+h|t} = \left\{ y \in \mathbb{R} : s^{y}_{t+h|t} \leq Q_{1-\alpha}\left(\sum_{i=t-t_c+1}^{t} \frac{1}{t_c + 1} \cdot \delta_{s_{i|i-h}} + \frac{1}{t_c + 1} \cdot \delta_{+\infty} \right) \right\}, \tag{1}$$

where $s^{y}_{t+h|t} := \mathscr{S}(z_{1:t}, y)$ denotes the h-step-ahead nonconformity score calculated at time t using a hypothesized test observation y.

3.2 Online multi-step weighted conformal prediction

Barber et al. (2023) propose nonexchangeable conformal prediction procedure (NexCP) that generalizes SCP to certain nonexchangeable settings by assigning higher weights to observations believed to share the same distribution as the test data. NexCP assumes these weights are fixed and data-independent. Under exchangeability, NexCP retains the same coverage guarantees as SCP; when exchangeability is violated, the coverage gap is governed by the total variation distance between swapped nonconformity score vectors, which can be substantial in time series settings.

The online **multi-step weighted conformal prediction** (MWCP) method we propose here adapts the NexCP method to the online setting for time series forecasting. MWCP uses weighted quantile estimate for constructing prediction intervals, contrasting with the MSCP definitions where all nonconformity scores for calibration are implicitly assigned equal weight.

For the subsequent empirical study, we choose fixed weights $w_i = b^{t+1-i}$, $b \in (0,1)$ and $i = t - t_c + 1, \ldots, t$, for nonconformity scores on the corresponding calibration set. In this setting, weights decay exponentially as the nonconformity scores get older, akin to the rationale behind the simple exponential smoothing method in time series forecasting (Hyndman & Athanasopoulos 2021). Then for each $h \in [H]$, MWCP updates the h-step-ahead prediction interval:

$$\hat{\mathscr{C}}^{\alpha}_{t+h|t} = \left\{ y \in \mathbb{R} : s^{y}_{t+h|t} \leq Q_{1-\alpha}\left(\sum_{i=t-t_c+1}^{t} \tilde{w}_i \cdot \delta_{s_{i|i-h}} + \tilde{w}_{t+1} \cdot \delta_{+\infty} \right) \right\},$$

where $\tilde{w}_i$ and $\tilde{w}_{t+1}$ are normalized weights given by

$$\tilde{w}_i = \frac{w_i}{\sum_{i=t-t_c+1}^{t} w_i + 1}, \text{ for } i \in \{t - t_c + 1, \ldots, t\} \quad \text{and} \quad \tilde{w}_{t+1} = \frac{1}{\sum_{i=t-t_c+1}^{t} w_i + 1}.$$

3.3 Multi-step adaptive conformal prediction

Next we extend the adaptive conformal inference (ACI, denoted as ACP hereafter) method proposed by Gibbs & Candès (2021) to address multi-step time series forecasting, introducing the **multi-step adaptive conformal prediction** (MACP) method. Assuming that $\beta \mapsto \mathbb{P}(y_{t+h} \in \hat{\mathscr{C}}^{\beta}_{t+h|t})$

is continuous and non-increasing, with $\mathbb{P}(y_{t+h} \in \hat{\mathscr{C}}^{0}_{t+h|t}) = 1$ and $\mathbb{P}(y_{t+h} \in \hat{\mathscr{C}}^{1}_{t+h|t}) = 0$, an optimal value $\alpha^{*}_{t+h|t} \in [0,1]$ exists such that the realised miscoverage rate of the corresponding prediction interval closely approximates the nominal miscoverage rate α. Specifically, for each $h \in [H]$, we iteratively estimate $\alpha^{*}_{t+h|t}$ by updating a parameter $\alpha_{t+h|t}$ through a sequential adjustment process

$$\alpha_{t+h|t} := \alpha_{t+h-1|t-1} + \gamma(\alpha - \text{err}_{t|t-h}). \tag{2}$$

Then the h-step-ahead prediction interval is computed using Equation (1) by setting $\alpha = \alpha_{t+h|t}$. Here, $\gamma > 0$ denotes a fixed step size parameter, $\alpha_{2h|h}$ denotes the initial estimate typically set to α, and $\text{err}_{t|t-h}$ denotes the miscoverage event $\text{err}_{t|t-h} = \mathbb{1}\left\{y_t \notin \hat{\mathscr{C}}^{\alpha_{t|t-h}}_{t|t-h}\right\}$.

Equation (2) indicates that the correction to the estimation of $\alpha^{*}_{t+h|t}$ at time $t+h$ is determined by the historical miscoverage frequency up to time t. At each iteration, the estimate used for quantile calibration is increased if $\hat{\mathscr{C}}^{\alpha_{t|t-h}}_{t|t-h}$ covers y_t and decreased otherwise. Therefore, the miscoverage event has a delayed impact on the estimation of $\alpha^{*}_{t+h|t}$ over h periods, indicating that the correction of the $\alpha^{*}_{t+h|t}$ estimate becomes less prompt with increasing values of h. In particular, Equation (2) reduces to the update for ACP for $h = 1$.

We do not consider the update equation $\alpha_{t+1|t-h+1} := \alpha_{t|t-h} + \gamma(\alpha - \text{err}_{t|t-h})$ in this context, as the available information at time t is insufficient to estimate $\alpha^{*}_{t+h|t}$ required for h-step forecasts.

Selecting the parameter γ is pivotal yet challenging. Gibbs & Candès (2021) suggest setting γ in proportion to the degree of variation of the unknown α^{*}_{t} over time. Several strategies have been proposed to avoid the necessity of selecting γ. For example, Zaffran et al. (2022) use an adaptive aggregation of multiple ACPs with a set of candidate values for γ, determining weights based on their historical performance. Bastani et al. (2022) propose a multivalid prediction algorithm in which the prediction set is established by selecting a threshold from a sequence of candidate thresholds. Both methods rely on update schemes that place substantial weight on older historical data, which may impede rapid adaptation to abrupt changes (Gibbs & Candès 2024). Therefore, Gibbs & Candès (2024) propose an alternative expert selection scheme that adaptively tunes

the step-size parameter over time and places greater emphasis on more recent observations by construction, enabling faster responses to sudden environmental shifts.

The theoretical coverage properties of ACP suggest that a larger value for γ generally results in less deviation from the target coverage. As there is no restriction on $\alpha_{t+h|t}$, and it can drift below 0 or above 1, a larger γ may lead to frequent output of null or infinite prediction sets in order to quickly adapt to the current miscoverage status.

3.4 Multi-step conformal PID control

We introduce **multi-step conformal PID control** method (hereafter MPID), which extends the PID method (Angelopoulos et al. 2023), originally developed for one-step-ahead forecasting. For each individual forecast horizon $h \in [H]$, the estimated $1-\alpha$ quantile of the h-step-ahead score at time t is updated iteratively as

$$q_{t+h|t} = \underbrace{q_{t+h-1|t-1} + \eta(\text{err}_{t|t-h} - \alpha)}_{\text{P}} + \underbrace{r_t\left(\sum_{i=h+1}^{t}(\text{err}_{i|i-h} - \alpha)\right)}_{\text{I}} + \underbrace{\hat{s}_{t+h|t}}_{\text{D}}, \tag{3}$$

where $\eta > 0$ is a constant learning rate, and r_t is a saturation function that adheres to

$$x \geq c \cdot g(t-h) \Longrightarrow r_t(x) \geq b, \quad \text{and} \quad x \leq -c \cdot g(t-h) \Longrightarrow r_t(x) \leq -b, \tag{4}$$

for constant $b, c > 0$, and an admissible function g that is sublinear, nonnegative, and nondecreasing. With this updating equation, we can obtain all required h-step-ahead prediction intervals using information available at time t. When $h = 1$, Equation (3) simplifies to the PID update, which guarantees long-run coverage. More importantly, Equation (3) represents a specific instance of Equation (10) that we will introduce later, thereby ensuring long-run coverage for each individual forecast horizon h according to Corollary 2.

The "P" control shows an h-period delay in updating the quantile estimate. The underlying intuition is similar to that of MACP: it increases (or decreases) the h-step-ahead quantile estimate if the prediction set at time t miscovered (or covered) the corresponding realization. MACP can

be considered as a special case of the P control, while the P control has the ability to prevent the generation of null or infinite prediction sets after a sequence of miscoverage events.

The "I" control captures cumulative historical coverage errors associated with h-step-ahead prediction intervals during updates, thereby enhancing the stability of the interval coverage.

The "D" control involves $\hat{s}_{t+h|t}$ as the h-step-ahead forecast of the nonconformity score (i.e., the forecast error), produced by any suitable forecasting model (or "scorecaster", see Section 4 for a detailed discussion) trained using the h-step-ahead nonconformity scores available up to and including time t. The effectiveness of the D control depends on the ability of the scorecaster to capture systematic and predictable patterns in the nonconformity scores, such as temporal dependence or conditional heteroskedasticity. Only when such structure is present and adequately modeled, this module provides additional benefits; otherwise, it may increase variability in the coverage and prediction intervals.

The MPID method will be improved in Section 4 by proposing the AcMCP method, which replaces the D-control component of MPID with a structured mechanism that explicitly captures autocorrelation in multi-step forecast errors. As a result, the long-run coverage guarantees established for AcMCP in Section 4.3 also apply to MPID.

4 Autocorrelated multi-step conformal prediction

In the PID method proposed by Angelopoulos et al. (2023), a notable feature is the inclusion of a scorecaster, a model trained on the score sequence to forecast the future score. The rationale behind it is to identify any leftover signal in the score distribution not captured by the base forecasting model. While this is appropriate in the context of huge data sets and potentially weak learners, it is unlikely to be a useful strategy for time series models. We expect to use a forecasting model that leaves only white noise in the residuals (equivalent to the one-step nonconformity scores). Moreover, the inclusion of a scorecaster often only introduces variance to the quantile estimate, resulting in inefficient (wider) prediction intervals.

On the other hand, in any non-trivial context, multi-step forecasts are correlated with each other. Specifically, the h-step-ahead forecast errors $e_{t+h|t}$ are correlated with the forecast errors from the previous $h-1$ steps, as errors propagate through the recursive forecasting procedure over the forecast horizon. However, no conformal prediction methods have taken this potential dependence into account in their methodological construction.

In this section, we will explore the properties of multi-step forecast errors and propose a novel conformal prediction method that accounts for their autocorrelations while providing theoretical long-run coverage guarantees.

4.1 Properties of multi-step forecast errors

We assume that a time series $\{y_t\}_{t\geq 1}$ follows a general non-stationary autoregressive process:

$$y_t = f_t(y_{(t-d):(t-1)}, \boldsymbol{x}_{(t-k):t}) + \varepsilon_t, \tag{5}$$

where f_t is a nonlinear function in d lagged values of y_t, the current value of the exogenous predictors, along with the preceding k values, and ε_t is white noise innovation with mean zero and constant variance. The sequence of model coefficients that parameterizes the function f is restricted to ensure that the stochastic process is locally stationary (Dahlhaus 2012).

Proposition 1 (Autocorrelations of multi-step optimal forecast errors)**.** *Let $\{y_t\}_{t\geq 1}$ be a time series generated by a general non-stationary autoregressive process as given in Equation (5), and assume that any exogenous predictors are known into the future. The forecast errors for optimal h-step-ahead forecasts can be approximately expressed as*

$$e_{t+h|t} = \omega_{t+h} + \phi_1 e_{t+h-1|t} + \cdots + \phi_p e_{t+h-p|t}, \tag{6}$$

where $p = \min\{d, h-1\}$, and ω_t is white noise. The approximation is obtained in the proof via a first-order Taylor series expansion of the forecasting function, where higher-order remainder terms are neglected. Therefore, the optimal h-step-ahead forecast errors are at most serially

correlated to lag $(h-1)$.

Proof. See Appendix A1. □

Proposition 1 suggests that the optimal h-step ahead forecast error, $e_{t+h|t}$, is serially correlated with the forecast errors from at most the past $h-1$ steps, i.e., $e_{t+1|t}, \ldots, e_{t+h-1|t}$. However, the autocorrelation among errors associated with optimal forecasts can not be used to improve forecasting performance, as it does not incorporate any new information available when the forecast was made. If we could forecast the forecast error, then we could improve the forecast, indicating that the initial forecast couldn't have been optimal.

The proof of Proposition 1, provided in Appendix A1, suggests that, if f_t is a linear autoregressive model, then the coefficients in Equation (6) are the linear coefficients of the optimal forecasting model. However, when f_t takes on a more complex nonlinear structure, the coefficients become complicated functions of observed data and unobserved model coefficients.

It is well-established in the forecasting literature that, for a zero-mean covariance-stationary time series, the optimal linear least-squares forecasts have h-step-ahead errors that are at most MA$(h-1)$ process (Harvey et al. 1997), a property that can be derived using Wold's representation theorem. The statement can be extended to time series generated by a general non-stationary autoregressive process (Sommer 2023), as described in Proposition 2. We provide the proof of this proposition in Appendix A2 based on Proposition 1.

Proposition 2 (MA$(h-1)$ process for h-step-ahead optimal forecast errors)**.** *Let* $\{y_t\}_{t\geq 1}$ *be a time series generated by a general non-stationary autoregressive process as given in Equation* (5)*, and assume that any exogenous predictors are known into the future. Under the approximation in Proposition 1, the forecast errors of optimal h-step-ahead forecasts follow an approximate MA*$(h-1)$ *process*

$$e_{t+h|t} = \omega_{t+h} + \theta_1 \omega_{t+h-1} + \cdots + \theta_{h-1} \omega_{t+1}. \tag{7}$$

where ω_t *is white noise.*

Proof. See Appendix A2. □

4.2 The AcMCP method

We can exploit these properties of multi-step forecast errors, leading to a new method that we call the **autocorrelated multi-step conformal prediction** (AcMCP) method. Unlike extensions of existing conformal prediction methods, which treat multi-step forecasting as independent events (see Section 3), the AcMCP method integrates the autocorrelations inherent in multi-step forecast errors, thereby making the output multi-step prediction intervals more statistically efficient.

The AcMCP method updates the quantile estimate q_t in an online setting to achieve the goal of long-run coverage. Specifically, the iteration of the h-step-ahead quantile estimate for the score is

$$q_{t+h|t} = q_{t+h-1|t-1} + \eta(\text{err}_{t|t-h} - \alpha) + r_t\left(\sum_{i=h+1}^{t}(\text{err}_{i|i-h} - \alpha)\right) + \tilde{e}_{t+h|t}, \tag{8}$$

for $h \in [H]$. Obviously, the AcMCP method can be viewed as a further extension of the PID method by Angelopoulos et al. (2023). Nevertheless, AcMCP diverges from PID with several innovations and differences.

First, we are no longer confined to predicting just one step ahead. Instead, we can make multi-step forecasts with accompanying theoretical coverage guarantees, constructing distribution-free prediction intervals for steps $t+1,\ldots,t+H$ based on available information up to time t.

Additionally, in AcMCP, $\tilde{e}_{t+h|t}$ is a forecast combination of two simple models: one being an MA$(h-1)$ model trained on the h-step-ahead forecast errors available up to and including time t (i.e. $e_{1+h|1},\ldots,e_{t|t-h}$), and the other a linear regression model trained by regressing $e_{t+h|t}$ on forecast errors from past steps (i.e. $e_{t+h-1|t},\ldots,e_{t+1|t}$). Unlike MPID, which treats multi-step prediction problems independently across horizons, AcMCP performs multi-step conformal prediction recursively, explicitly accounting for the serial dependence among multi-step forecast errors. Importantly, the role of $\tilde{e}_{t+h|t}$ is not to forecast the nonconformity scores themselves, but to incorporate the autocorrelation structure of multi-step forecast errors into the construction of the

resulting prediction intervals.

4.3 Coverage guarantees

Proposition 3. *Let* $\{s_{t+h|t}\}_{t\in\mathbb{N}}$ *be any sequence of numbers in* $[-b,b]$ *for any* $h \in [H]$*, where* $b > 0$*, and may be infinite. Assume that* r_t *is a saturation function obeying Equation* (4)*, for an admissible function* g*. Then the iteration* $q_{t+h|t} = r_t\left(\sum_{i=h+1}^{t}(\mathrm{err}_{i|i-h} - \alpha)\right)$ *satisfies*

$$\left|\frac{1}{T-h}\sum_{t=h+1}^{T}(\mathrm{err}_{t|t-h} - \alpha)\right| \leq \frac{c\cdot g(T-h)+h}{T-h}, \tag{9}$$

for any $T \geq h+1$*, where* $c > 0$ *is the constant in Equation* (4)*.*

Therefore the prediction intervals obtained by the iteration yield the correct long-run coverage; i.e., $\lim_{T\to\infty}\frac{1}{T-h}\sum_{t=h+1}^{T}\mathrm{err}_{t|t-h} = \alpha$*.*

Proof. See Appendix A3. □

When $h = 1$, Proposition 3 reduces to Proposition 2 of Angelopoulos et al. (2023). While Angelopoulos et al. (2023) only consider one-step-ahead forecasting, Proposition 3 extends their result to the multi-step setting and provides an explicit upper bound on the coverage gap. For finite T, Proposition 3 indicates that increasing the forecast horizon h amplifies in-sample deviations from the target coverage because $g(T-h)/(T-h)$ is non-increasing, given that the admissible function g is sublinear, nonnegative, and nondecreasing. Importantly, this reflects finite-sample behavior rather than a breakdown of the asymptotic coverage guarantee. As expected, forecast uncertainty increases with the horizon, since predictions further into the future are affected by more sources of variability. In finite samples, conformal prediction intervals may not fully scale with this growing uncertainty, leading to a larger gap between nominal and empirical coverage.

The quantile iteration $q_{t+h|t} = q_{t+h-1|t-1} + \eta(\mathrm{err}_{t|t-h} - \alpha)$ can be seen as a particular instance of the iteration outlined in Proposition 3 if we set $q_{2h|h} = 0$ without losing generality. Thus, its coverage bounds can be easily derived as a result of Proposition 3.

Corollary 1. *Let* $\{s_{t+h|t}\}_{t\in\mathbb{N}}$ *be any sequence of numbers in* $[-b,b]$ *for any* $h \in [H]$*, where* $b > 0$*,*

and may be infinite. Then the iteration $q_{t+h|t} = q_{t+h-1|t-1} + \eta(\mathrm{err}_{t|t-h} - \alpha)$ *satisfies*

$$\left| \frac{1}{T-h} \sum_{t=h+1}^{T} (\mathrm{err}_{t|t-h} - \alpha) \right| \leq \frac{b + \eta h}{\eta (T-h)},$$

for any learning rate $\eta > 0$ *and* $T \geq h+1$.

Therefore the prediction intervals obtained by the iteration yield the correct long-run coverage; i.e., $\lim_{T\to\infty} \frac{1}{T-h} \sum_{t=h+1}^{T} \mathrm{err}_{t|t-h} = \alpha$.

Proof. See Appendix A4. □

When $h = 1$, Corollary 1 reduces to Proposition 1 of Angelopoulos et al. (2023). More generally, Corollary 1 extends this result to the multi-step setting and establishes an explicit upper bound on the finite-sample coverage gap. The boundedness assumption on the nonconformity scores is imposed solely for theoretical analysis, ensuring control over the accumulation of miscoverage errors in the quantile tracking recursion. When the update is unraveled, the quantile tracker can be interpreted as an error integrator that accumulates past deviations of the empirical coverage from the target level. Bounded scores guarantee the stability of this accumulation and enable explicit bounds on the finite-sample coverage gap. Importantly, the bound itself is not required in practice, and the update proceeds agnostically and remains valid as long as the scores are finite.

More importantly, Proposition 3 is also adequate for establishing the coverage guarantee of AcMCP given by Equation (8). We first reformulate Equation (8) as

$$q_{t+h|t} = \hat{q}_{t+h|t} + r_t \left(\sum_{i=h+1}^{t} (\mathrm{err}_{i|i-h} - \alpha) \right), \tag{10}$$

where $\hat{q}_{t+h|t}$ can be any function of the past observations $\{(x_i, y_i)\}_{1 \leq i \leq t}$ and quantile estimates $q_{i+h|i}$ for $i \leq t-1$. Taking $\hat{q}_{t+h|t} = q_{t+h-1|t-1} + \eta(\mathrm{err}_{t|t-h} - \alpha) + \tilde{e}_{t+h|t}$ will recover Equation (8). We can consider $\hat{q}_{t+h|t}$ as the forecast of the quantile $q_{t+h|t}$ based on available historical data. We then present the long-run coverage guarantee for AcMCP given by Equation (10).

Corollary 2. *Let* $\{\hat{q}_{t+h|t}\}_{t\in\mathbb{N}}$ *be any sequence of numbers in* $[-\frac{b}{2}, \frac{b}{2}]$*, and* $\{s_{t+h|t}\}_{t\in\mathbb{N}}$ *be any sequence of numbers in* $[-\frac{b}{2}, \frac{b}{2}]$*, for any* $h \in [H]$*,* $b > 0$ *and may be infinite. Assume that* r_t *is*

a saturation function obeying Equation (4), for an admissible function g. Then the prediction intervals obtained by the AcMCP iteration given by Equation (10) yield the correct long-run coverage; i.e., $\lim_{T\to\infty} \frac{1}{T-h} \sum_{t=h+1}^{T} \mathrm{err}_{t|t-h} = \alpha$.

Proof. See Appendix A5. □

5 Experiments

We evaluate the proposed multi-step conformal prediction methods using two simulated settings and two real-world data sets. The following parameter choices are used throughout:

(a) we focus on the target coverage level $1-\alpha = 0.9$;

(b) for the MWCP method, we use $b = 0.99$ as per Barber et al. (2023);

(c) following Angelopoulos et al. (2023), MACP uses a step size $\gamma = 0.005$; MPID uses a Theta model as the scorecaster; and both MPID and AcMCP use a learning rate of $\eta = 0.01\hat{B}_t$ for quantile tracking, where $\hat{B}_t = \max\{s_{t-\Delta+1|t-\Delta-h+1}, \ldots, s_{t|t-h}\}$ is the highest score over a tailing window of length Δ, which is set equal to the calibration set size;

(d) we adopt a nonlinear saturation function $r_t(x) = K_{\mathrm{I}} \tan\left(x \log(t)/(tC_{\mathrm{sat}})\right)$, where $\tan(x) = \mathrm{sign}(x) \cdot \infty$ for $x \notin [-\pi/2, \pi/2]$, and constants $C_{\mathrm{sat}}, K_{\mathrm{I}} > 0$ are chosen heuristically as in Angelopoulos et al. (2023);

(e) we use a clipped version of MACP, replacing infinite intervals with the largest score observed to date. Other methods do not produce infinite intervals by construction and therefore require no clipping.

5.1 Simulated linear autoregressive process

We first consider a simulated stationary time series generated from a simple $\mathrm{AR}(2)$ process

$$y_t = 0.8y_{t-1} - 0.5y_{t-2} + \varepsilon_t,$$

where ε_t is white noise with error variance $\sigma^2 = 1$. After an appropriate burn-in period, we generate $N = 5000$ data points. Under the sequential split and online learning settings, we create training sets $\mathscr{D}_{\text{tr}}$ and calibration sets $\mathscr{D}_{\text{cal}}$, each with a length of 500. We use AR(2) models to generate 1- to 3-step-ahead point forecasts (i.e. $H = 3$), using the `Arima()` function from the `forecast` R package (Hyndman & Khandakar 2008, Hyndman et al. 2026). The goal is to generate prediction intervals using various proposed conformal prediction methods and evaluate whether they can achieve the nominal long-run coverage for each separate forecast horizon.

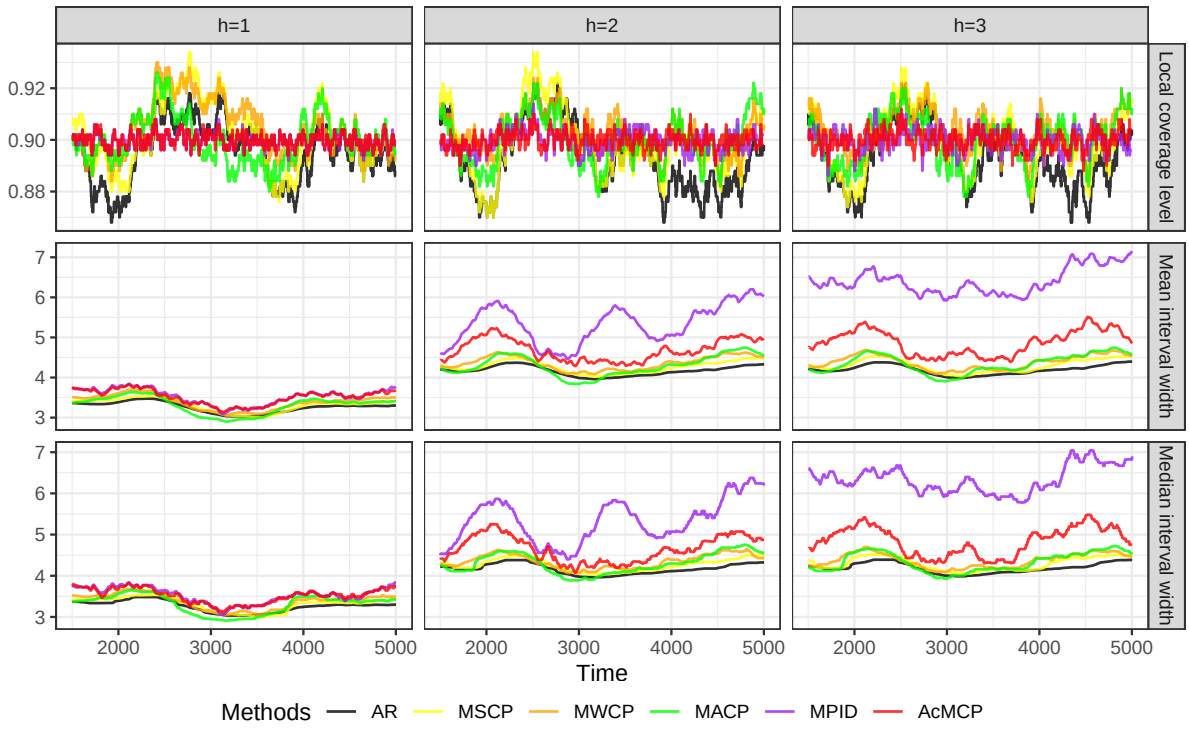

Figure 2: *AR(2) simulation results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 500*. The target coverage level is* $1 - \alpha = 0.9$.

Figure 2 presents the rolling coverage and interval width of each method for each forecast horizon, with metrics computed over a rolling window of size 500. The analytic (and optimal) intervals obtained from the AR model are also shown. We observe that MPID and AcMCP achieve approximately the desired 90% coverage level over the rolling windows, while other methods, including the AR model, undergo much wider swings away from the desired level,

showing high coverage volatility over time. For each forecast horizon and method, the rolling mean and median interval widths follow very similar trajectories across rolling windows. AcMCP constructs narrower prediction intervals than MPID, despite both methods achieving similar coverage. Moreover, we see that AcMCP tends to offer coverage-adaptive prediction intervals, automatically adjusting their width based on past coverage performance, and results in wider intervals especially when competing methods undercover, which is to be expected. In short, AcMCP intervals offer greater adaptivity and more precise coverage compared to AR, MSCP, MWCP and MACP. Both MPID and AcMCP achieve tight coverage, but AcMCP outputs more informative and smaller intervals. The benefits of AcMCP are more noticeable when h grows, and it does not deteriorate for smaller h. This improvement can be attributed to the inclusion of a second model (scorecaster) which introduces additional variance into the generated prediction intervals. The results can be further elucidated with Figure 3, which presents boxplots of rolling coverage and interval width for each method and each forecast horizon.

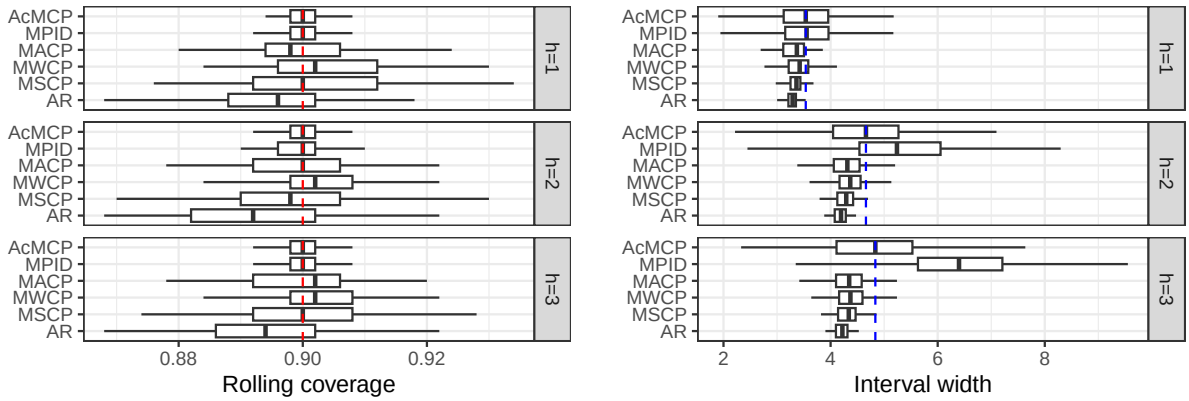

Figure 3: *AR(2) simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

The inclusion of the last term $\tilde{e}_{t+h|t}$ in AcMCP should only result in a slight difference compared to the version without this term, which we henceforth refer to as MPI. This is because, the inclusion of $\tilde{e}_{t+h|t}$ aims to capture autocorrelations inherent in multi-step forecast errors and focuses on the

mean of forecast errors, whereas the whole update of AcMCP operates on quantiles of scores. To illustrate the subtle difference in their results and explore their origins, we visualize their prediction intervals over a truncated period of length 500, as shown in Figure 4. We observe that AcMCP and MPI indeed construct similar prediction intervals so their lower and upper bounds mostly overlap with each other. The main differences occur around the time 1320 and during the period 1470-1500, where AcMCP tends to have a fanning-out effect, increasing the interval width as the forecast horizon increases, compared to MPI. Figure 4 also presents the prediction interval bounds given by MACP. Since both MACP and AcMCP rely on past miscoverage information to update prediction intervals, we include MACP to support a direct comparison of their resulting prediction intervals while preserving a clear and concise presentation. The prediction intervals of both AcMCP and MACP can capture certain patterns in the actual observations, and there is no consistent pattern indicating dominance of one method over the other in terms of interval width.

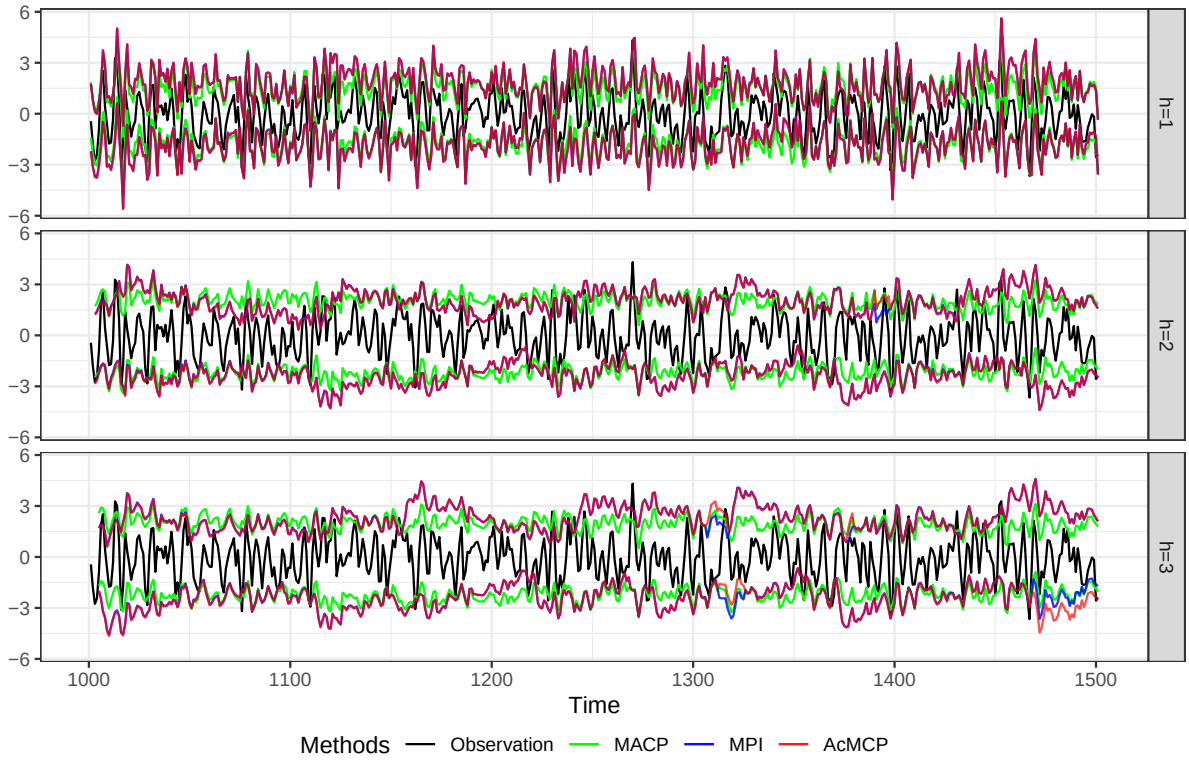

Figure 4: *AR(2) simulation results showing the prediction interval bounds for the MACP, MPI, and AcMCP methods over a truncated period of length 500.*

5.2 Simulated nonlinear autoregressive process

Next consider the case of a nonlinear data generation process, defined as

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

where $x_{1,t}$ and $x_{2,t}$ are uniformly distributed on $[0,1]$, and ε_t is white noise with variance $\sigma^2 = 0.1$. Thus, y_t nonlinearly depends on its lagged values y_{t-1}, y_{t-2}, and exogenous variables $x_{1,t}$ and $x_{2,t}$.

After an appropriate burn-in period, we generate $N = 2000$ data points. Under the sequential split and online learning settings, both training set $\mathscr{D}_{\text{tr}}$ and calibration set $\mathscr{D}_{\text{cal}}$ have size 500. Given the nonlinear DGP, we use feed-forward neural networks with a single hidden layer and lagged inputs to generate 1- to 3-step-ahead point forecasts ($H = 3$), using the `nnetar()` function from the `forecast` R package (Hyndman et al. 2026). Note that analytic prediction intervals are unavailable for neural networks, thus we do not include them when presenting the results.

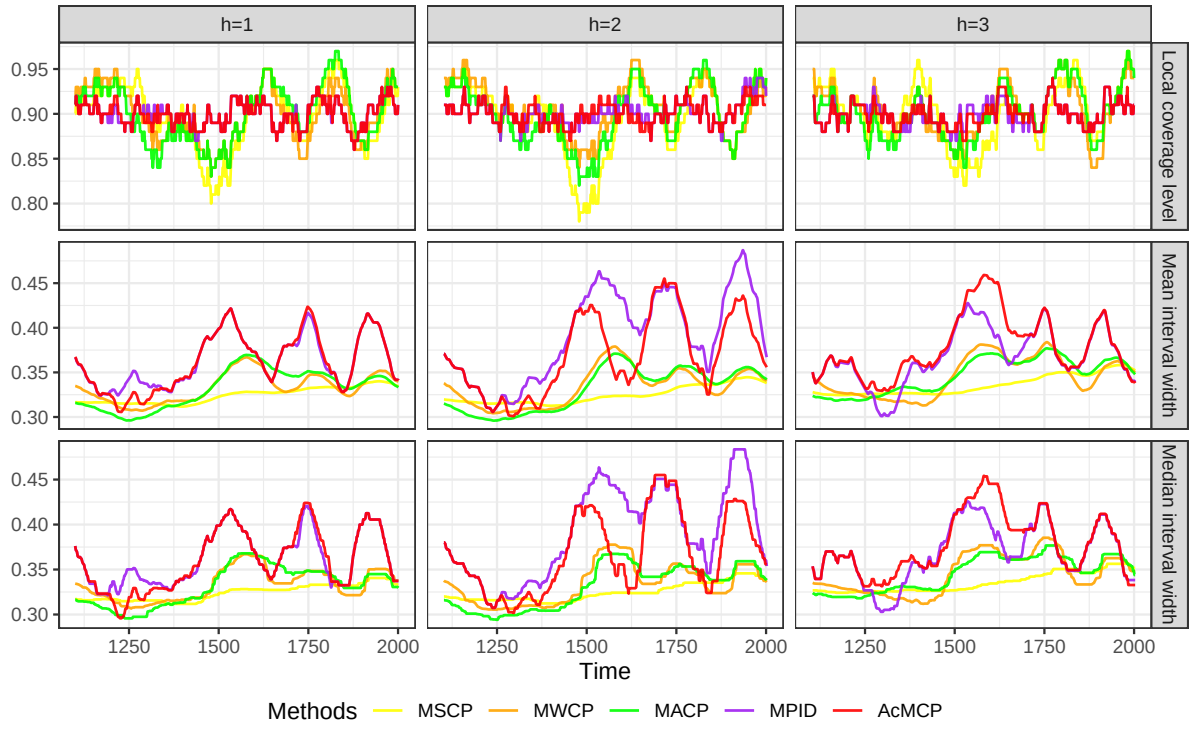

Figure 5: *Nonlinear simulation results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 100*. The target coverage level is* $1-\alpha = 0.9$.

Figure 5 shows the rolling coverage and interval width of each method, computed on a rolling window of size 100. MPID and AcMCP are able to maintain minor fluctuations around the target coverage of 90% over time, whereas MSCP, MWCP, and MACP struggle to sustain the target level and display pronounced fluctuations over time. Moreover, all methods, except for MSCP, adapt interval widths to distributional changes, widening intervals in response to undercoverage and narrowing them under overcoverage. Notably, MPID and AcMCP demonstrate greater adaptability, with higher variability in interval widths compared to competing methods in order to uphold the desired coverage. The outcome of the rolling mean and median interval widths indicates that AcMCP produces generally narrower intervals than MPID for 2-step-ahead forecasts but wider intervals for 3-step-ahead forecasts. Furthermore, under AcMCP, the interval width at $h = 3$ exceeds that at $h = 2$ in 53.41% of the test cases, compared with only 38.08% under MPID. This pattern suggests more plausible behavior for AcMCP, as forecast uncertainty typically increases with the forecast horizon, leading to a greater tendency for wider prediction intervals.

We provide further insights into the performance of these conformal prediction methods by presenting boxplots of the rolling coverage and interval width for each method, as depicted in Figure 6. We observe that coverage variability is higher for MSCP, MWCP and MACP than for MPID and AcMCP, while MPID and AcMCP lead to a lower effective interval size.

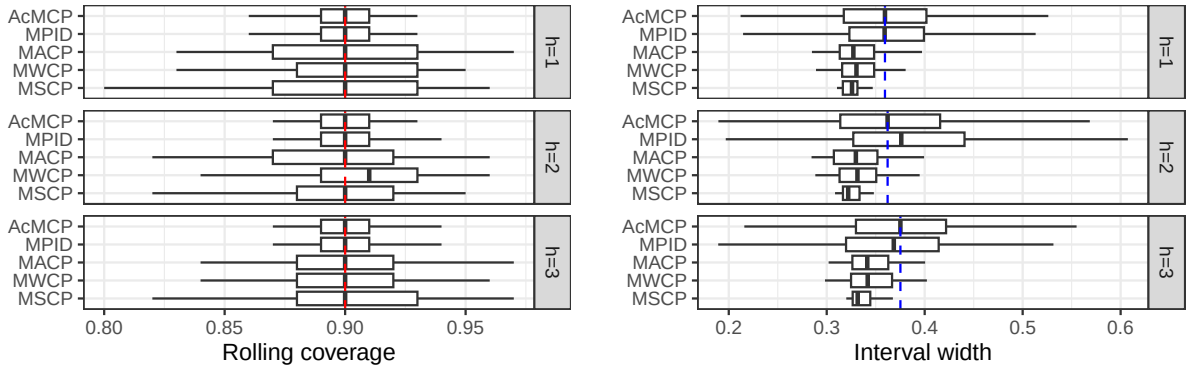

Figure 6: *Nonlinear simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

5.3 Electricity demand data

We apply the conformal prediction methods using data comprising daily electricity demand (GW), daily maximum temperature (degrees Celsius), and holiday information for Victoria, Australia, from 2012 to 2014. Temperatures are taken from the Melbourne Regional Office weather station. The left panel of Figure 7 displays the daily electricity demand and temperatures over the period. The right panel shows a nonlinear relationship between demand and temperature, with demand increasing for low temperatures (due to heating) and increasing for high temperatures (due to cooling); the two clouds of points correspond to working days and non-working days.

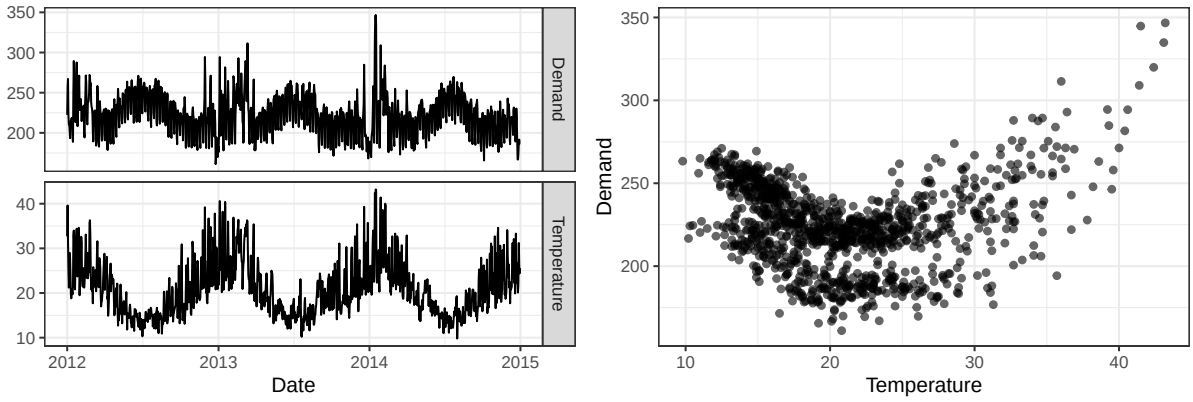

Figure 7: *Daily electricity demand and corresponding daily maximum temperatures in 2012–2014, Victoria, Australia.*

The response variable is `Demand`, with two covariates: `Temperature`, and `Workday` (an indicator variable for working days). Following Hyndman & Athanasopoulos (2021), we fit a dynamic regression model with a piecewise linear function of temperature (with a knot at 18 degrees) to generate 1- to 7-step-ahead point forecasts ($H = 7$), modeling the errors with an ARIMA process to capture autocorrelation. We use two years of data for training and 100 observations for calibration.

Figure 8 and Figure 9 compare the rolling coverage and interval width across method. These computations are based on a rolling window of size 100. The DR method corresponds to the analytic intervals obtained from the dynamic regression model. First, DR consistently attains coverage well above the 90% target, resulting in much wider intervals than other methods for

$h = 1,2,3,4$. While this overcoverage improves reliability, the resulting intervals are overly conservative and may lead to inefficient operational decisions in electricity markets, such as excessive reserve procurement. Second, MSCP, MWCP, and MACP fail to sustain the target coverage and noticeably undercover after September 2014 for all horizons, thus leading to narrower intervals than others. In electricity market applications, such undercoverage underestimates demand uncertainty and increases the risk of insufficient reserve allocation. Third, MPI, MPID, and AcMCP offer wider intervals that largely mitigate the post–September 2014 undercoverage. Among these, MPID shows slightly poorer coverage than MPI and AcMCP at $h = 3$, despite producing to wider intervals. Finally, while MPI and AcMCP display similar coverage pattern, AcMCP is capable of constructing narrower intervals than MPI, particularly at larger horizons. Relative to MPID, AcMCP also achieves greater interval width reduction, with advantages that become more pronounced as h increases and without sacrificing performance at shorter horizons.

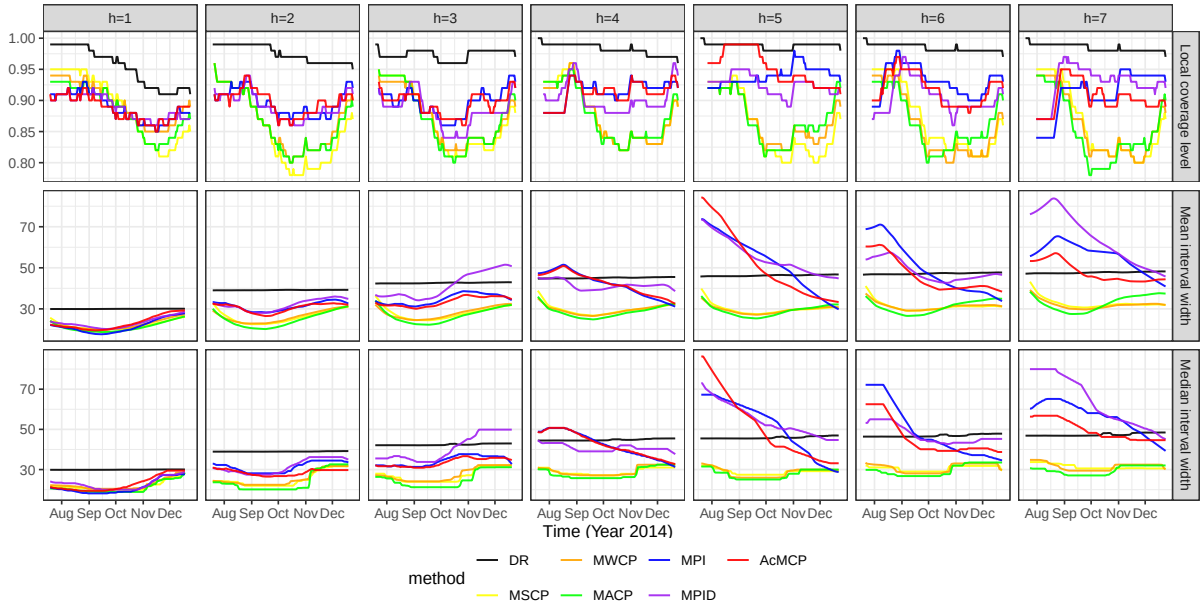

Figure 8: *Electricity demand data results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 100*. The target coverage level is* $1 - \alpha = 0.9$.

5.4 Eating out expenditure data

In our final example, we apply the conformal prediction methods to forecast the eating out expenditure ($ million) in Victoria, Australia. The data includes monthly expenditure on cafes, restaurants and takeaway food services from April 1982 to December 2019, as shown in Figure 10. The data shows an overall upward trend, obvious annual seasonal, and evel-dependent variability.

We consider three models: ARIMA with logarithmic transformation, ETS, and STL-ETS (Hyndman & Athanasopoulos 2021), and then output their simple average as final point forecasts. STL-ETS applies ETS to the seasonally adjusted series obtained via STL decomposition. All models are automatically trained using the `forecast` R package (Hyndman et al. 2026). We aim to forecast 12 months ahead ($H = 12$) using 20 years of data for training and 5 year for calibration.

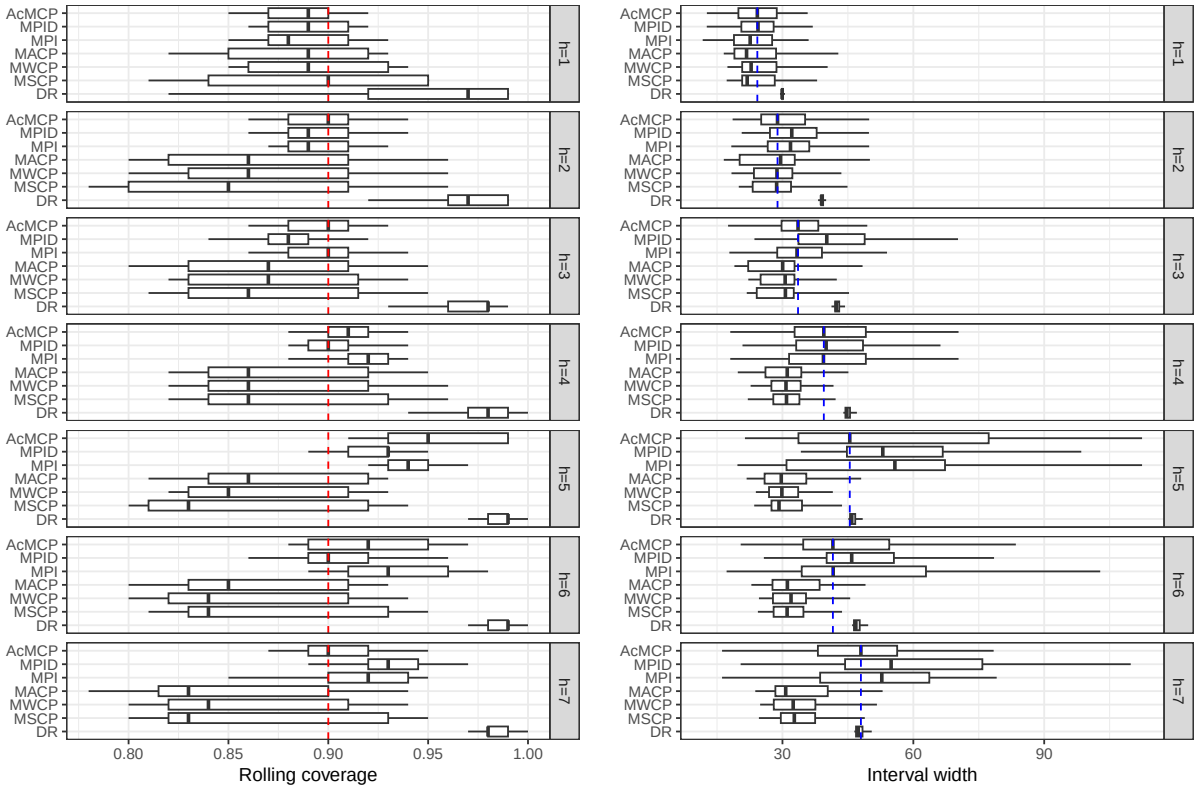

Figure 9: *Electricity demand data results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

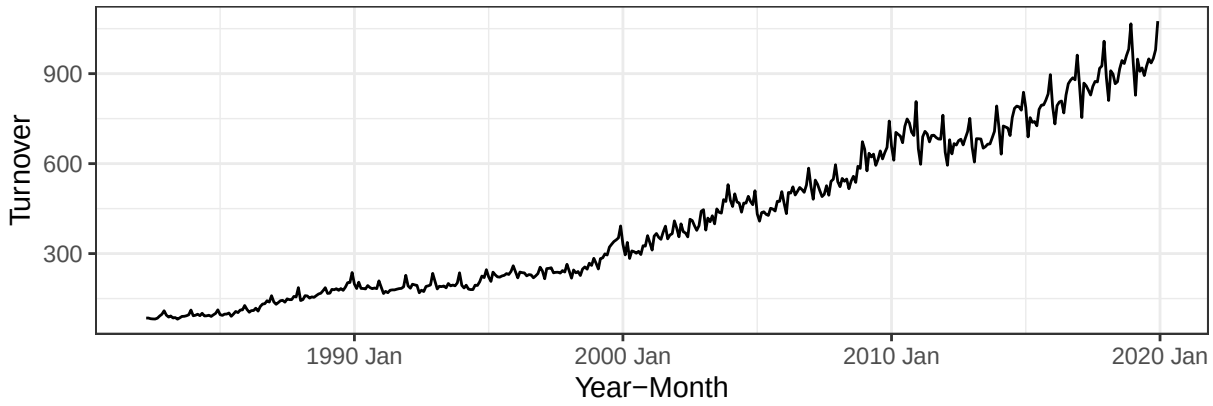

Figure 10: *Monthly expenditure on cafes, restaurants and takeaway food services in Victoria, Australia, from April 1982 to December 2019.*

Because the test set contains only 152 months, we report coverage and interval width averaged over the full test period rather than rolling windows.

As the forecast horizons considered in this application are relatively long, we summarize performance in Figure 11 by reporting the average coverage gap and average interval width across the whole test set for each method and each forecast horizon. The results first show that MSCP, MWCP and MACP provide valid prediction intervals for smaller forecast horizon but fail to achieve the desired coverage for larger forecast horizons ($h > 5$). Second, for $h \leq 5$, MPI and AcMCP can approximately achieve the desired coverage and provide comparable mean interval widths with other methods, except for MPID. Third, the coverage of MSCP, MWCP and MACP declines gradually as the forecast horizon increases, while MPI and AcMCP maintain coverage within a tighter range, albeit at the cost of interval efficiency. Lastly, compared to MPI, AcMCP exhibits slightly less deviation from the desired coverage across most forecast horizons.

5.5 Post-hoc comparison

For MACP, the step size γ controls both the calibration tightness and the responsiveness of the interval widths: smaller values lead to smoother interval evolution but slower correction of local coverage errors, while larger values improve responsiveness at the cost of increased

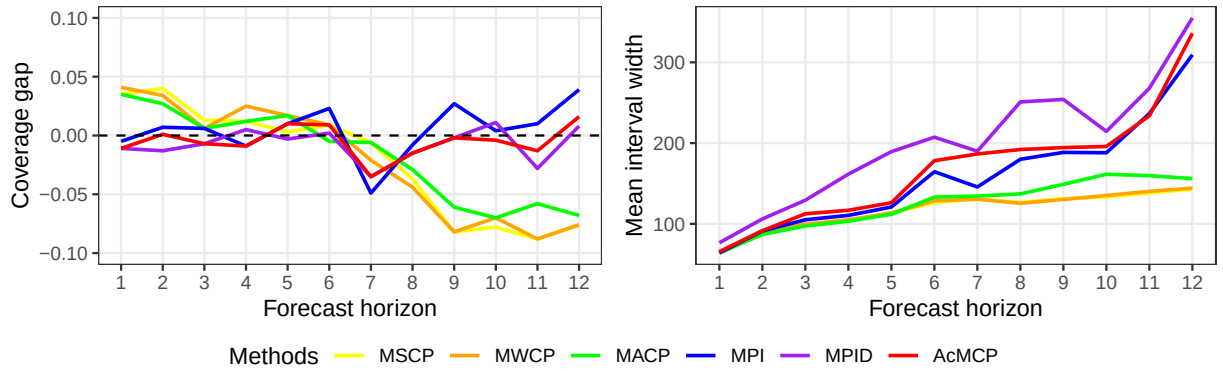

Figure 11: *Eating out expenditure data results showing coverage gap and interval width averaged over the entire test set for each forecast horizon. The black dashed line in the top panel indicates no difference from the* 90% *target level.*

variability. An analogous trade-off arises in AcMCP, where the learning rate η governs the balance between adaptivity and stability. This naturally raises the question of whether AcMCP's empirical advantages persist when the benchmark MACP is optimally tuned for high adaptivity.

To examine this, we conduct a post-hoc comparison in which η for AcMCP is held fixed, while γ for MACP is selected via grid search over $[0.005, 0.5]$. Specifically, γ is chosen to match the sample variance of local coverage achieved by AcMCP over rolling windows of the test set.

Table 1 summarizes, for each forecast horizon, the mean coverage over the full test set, the minimum and maximum local coverage across rolling windows, and the mean and median interval widths for MACP and AcMCP across the data sets considered above. In simple linear and nonlinear simulations, both methods attain mean coverage close to the nominal 90%, with AcMCP consistently producing comparable or narrower intervals when coverage is comparable. More pronounced differences arise in real data sets and at longer horizons. For electricity demand, MACP displays systematic undercoverage as the horizon increases, particularly in terms of minimum local coverage, whereas AcMCP maintains more stable coverage around the target by adaptively adjusting interval widths. A similar pattern appears for eating out expenditure data, where MACP suffers substantial coverage degradation at larger horizons despite tuning, while AcMCP preserves near-nominal coverage, albeit with wider but informative intervals. Overall,

these results indicate that AcMCP provides superior coverage control and, when coverage is comparable, yields prediction intervals that are shorter or at least comparable in length.

Table 1: *Performance of MACP and AcMCP across the four datasets, summarized by mean test-set coverage, minimum and maximum local coverage over rolling windows, and mean and median interval widths for each forecast horizon.*

	Mean coverage (%)		Min. local coverage (%)		Max. local coverage (%)		Mean interval width		Median interval width	
	MACP	AcMCP	MACP	AcMCP	MACP	AcMCP	MACP	AcMCP	MACP	AcMCP
					AR(2) simulation data					
h=1	89.95	90.00	89.20	89.40	90.80	90.80	4.04	3.55	3.62	3.54
h=2	89.89	89.94	89.00	89.00	91.00	91.00	5.08	4.68	4.85	4.66
h=3	89.94	90.02	88.80	88.80	91.00	91.20	5.16	4.68	5.00	4.84
					Nonlinear simulation data					
h=1	90.00	90.00	87.00	86.00	94.00	93.00	0.35	0.36	0.33	0.36
h=2	89.98	90.18	86.00	87.00	94.00	93.00	0.37	0.37	0.33	0.36
h=3	89.86	90.06	87.00	87.00	94.00	94.00	0.35	0.38	0.33	0.38
					Electricity demand data					
h=1	89.92	89.15	86.00	85.00	91.00	92.00	25.02	24.44	22.52	24.24
h=2	88.67	90.23	83.00	86.00	91.00	94.00	30.56	30.72	29.03	28.87
h=3	88.58	90.94	81.00	86.00	92.00	93.00	34.81	34.35	31.64	33.57
h=4	89.29	90.87	84.00	88.00	93.00	94.00	41.93	40.74	34.90	39.48
h=5	89.60	94.80	86.00	91.00	93.00	99.00	41.17	55.62	34.72	45.43
h=6	90.32	90.73	85.00	88.00	92.00	97.00	36.34	46.65	35.08	41.56
h=7	87.80	89.84	84.00	87.00	90.00	95.00	36.55	47.45	35.02	47.98
					Eating out expenditure data					
h=1	85.62	88.89	80.00	83.33	90.00	93.33	76.70	65.51	64.21	63.08
h=2	84.77	90.01	78.33	83.33	93.33	96.67	97.93	91.50	84.26	89.10
h=3	80.54	89.26	68.33	83.33	95.00	95.00	104.10	112.42	101.46	107.83
h=4	59.18	89.12	15.00	81.67	91.67	98.33	89.74	116.80	77.68	111.16
h=5	46.90	91.03	13.33	81.67	65.00	96.67	59.72	126.12	68.74	120.97
h=6	36.36	90.91	8.33	85.00	41.67	96.67	29.26	178.17	62.05	178.60
h=7	53.90	86.52	31.67	73.33	61.67	98.33	63.10	186.58	81.44	198.27
h=8	47.48	88.29	23.33	83.33	58.33	98.33	55.30	192.07	78.62	192.59
h=9	32.12	89.78	1.67	86.67	48.33	100.00	20.29	194.40	47.47	168.96
h=10	71.85	89.63	53.33	85.00	90.00	93.34	116.88	195.83	121.56	193.68
h=11	69.92	88.72	41.67	85.00	98.33	96.67	102.20	233.80	99.26	220.92
h=12	69.47	91.60	65.00	83.33	83.33	98.33	127.19	336.02	128.28	297.48

6 Conclusion and discussion

We introduced a unified notation for conformal prediction in time series, with a focus on multi-step forecasting within a general online learning framework. We extended several accessible conformal prediction methods to address the challenges of multi-step forecasting scenarios.

Under a general nonstationary autoregressive DGP, we showed that the optimal h-step-ahead forecast errors can be approximated by a linear combination of at most its lag $(h-1)$ with respect to forecast horizon. Building on this foundation, we introduce AcMCP, a novel method that explicitly accounts for autocorrelations in multi-step forecast errors and achieves long-run coverage guarantees without assumptions on data distributional shifts. Across simulations and

real data applications, AcMCP achieves coverage closer to the target within local windows while offering adaptive prediction intervals that respond effectively to varying conditions.

Our analysis is restricted to an ex-post forecasting setting, in which future values of exogenous predictors are assumed to be observed. In this setting, interval width reflects both *aleatoric uncertainty*, arising from intrinsic randomness in DGP, and *epistemic uncertainty*, stemming from a lack of knowledge about the underlying DGP (Sale et al. 2025). In many practical forecasting scenarios, however, predictors must themselves be forecast. In such settings, both types of uncertainty are amplified and interdependent: aleatoric uncertainty expands to include the variability induced by stochastic predictor forecasts, while epistemic uncertainty compounds through the propagation of uncertainty from the predictor models to the primary predictive model. Consequently, conformal inference procedures require adaptation to explicitly incorporate and propagate this joint uncertainty in order to preserve valid coverage and statistical efficiency. Additionally, our methodology does not incorporate an automated procedure for tuning the learning-rate parameter.

These limitations suggest several directions for future research. One avenue is to further improve interval efficiency while preserving coverage guarantees, with the goal of producing the narrowest possible prediction intervals. Our focus here is on methods that admit efficient online implementation through simple weighting or updating schemes, making them well suited for practical multi-step forecasting. More computationally intensive methods such as Bellman conformal inference (Yang, Candès & Lei 2024), which solve an optimization problem at each time step to trade off interval length and miscoverage, are not considered here. Integrating such optimization-based ideas with the proposed framework to further tighten intervals while retaining coverage guarantees is an important topic for future work. Another promising direction is the development of refined conformal methodologies for ex-ante forecasting, where uncertainty in future predictors should be explicitly incorporated into the inference procedure.

Acknowledgments

This work was supported by the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA) under Grant No. IC200100009; the Presidential Foundation of the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China, under Grant No. E555930101.

Disclosure statement

The authors report there are no competing interests to declare.

Supplementary materials

All proofs of the propositions and corollaries presented in the paper are provided in the Appendix.

References

Angelopoulos, A. N., Barber, R. F. & Bates, S. (2024), Online conformal prediction with decaying step sizes, *in* 'Proceedings of the 41st International Conference on Machine Learning', Vol. 235, PMLR, pp. 1616–1630.

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

Barber, R. F., Candès, E. J., Ramdas, A. & Tibshirani, R. J. (2021), 'Predictive inference with the jackknife+', *The Annals of Statistics* **49**(1), 486–507.

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

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R. & Roth, A. (2022), 'Practical adversarial multivalid conformal prediction', *Advances in Neural Information Processing Systems* **35**, 29362–29373.

Bhatnagar, A., Wang, H., Xiong, C. & Bai, Y. (2023), Improved online conformal prediction via strongly adaptive online learning, *in* 'Proceedings of the 40th International Conference on Machine Learning', Vol. 202, PMLR, pp. 2337–2363.

Chernozhukov, V., Wüthrich, K. & Yinchu, Z. (2018), Exact and robust conformal inference methods for predictive machine learning with dependent data, *in* 'Conference On learning theory', Vol. 75, PMLR, pp. 732–749.

Dahlhaus, R. (2012), Locally stationary processes, *in* 'Handbook of statistics', Vol. 30, Elsevier, pp. 351–413.

Diebold, F. X. & Lopez, J. A. (1996), Forecast evaluation and combination, *in* 'Statistical Methods in Finance', Vol. 14 of *Handbook of Statistics*, Elsevier, pp. 241–268.

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

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

Guan, L. (2023), 'Localized conformal prediction: A generalized inference framework for conformal prediction', *Biometrika* **110**(1), 33–50.

Gupta, C., Kuchibhotla, A. K. & Ramdas, A. (2022), 'Nested conformal prediction and quantile out-of-bag ensemble methods', *Pattern Recognition* **127**, 108496.

Harvey, D., Leybourne, S. & Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**(2), 281–291.

Hyndman, R. J. & Athanasopoulos, G. (2021), *Forecasting: principles and practice*, 3rd edn, OTexts, Melbourne, Australia. https://OTexts.com/fpp3.

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

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: the forecast package for R', *Journal of Statistical Software* **26**(3), 1–22.

Lei, J. & Wasserman, L. (2014), 'Distribution-free prediction bands for non-parametric regression', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**(1), 71–96.

Oliveira, R. I., Orenstein, P., Ramos, T. & Romano, J. V. (2024), 'Split conformal prediction and non-exchangeable data', *Journal of Machine Learning Research* **25**(225), 1–38.

Papadopoulos, H. (2008), Inductive conformal prediction: Theory and application to neural networks, *in* P. Fritzsche, ed., 'Tools in Artificial Intelligence', InTech, chapter 18, pp. 315–330.

Papadopoulos, H., Proedrou, K., Vovk, V. & Gammerman, A. (2002), Inductive confidence machines for regression, *in* 'European conference on machine learning', Springer, pp. 345–356.

Papadopoulos, H., Vovk, V. & Gammerman, A. (2007), Conformal prediction with neural networks, *in* '19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)', Vol. 2, IEEE, pp. 388–395.

Patton, A. J. & Timmermann, A. (2007), 'Properties of optimal forecasts under asymmetric loss and nonlinearity', *Journal of Econometrics* **140**(2), 884–918.

Sale, Y., Javanmardi, A. & Hüllermeier, E. (2025), Aleatoric and epistemic uncertainty in conformal prediction, *in* 'Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications', Vol. 266 of *Proceedings of Machine Learning Research*, PMLR, pp. 784–786.

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

Shafer, G. & Vovk, V. (2008), 'A tutorial on conformal prediction.', *Journal of Machine Learning Research* **9**(3).

Sommer, B. (2023), Forecasting and decision-making for empty container repositioning, PhD thesis, Technical University of Denmark.

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

Tibshirani, R. J., Foygel Barber, R., Candes, E. & Ramdas, A. (2019), 'Conformal prediction under covariate shift', *Advances in Neural Information Processing Systems* **32**.

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

Wisniewski, W., Lindsay, D. & Lindsay, S. (2020), Application of conformal prediction interval estimations to market makers' net positions, *in* 'Conformal and probabilistic prediction and applications', PMLR, pp. 285–301.

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

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

Yang, Y., Kuchibhotla, A. K. & Tchetgen Tchetgen, E. (2024), 'Doubly robust calibration of prediction sets under covariate shift', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86**(4), 943–965.

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

Yu, X., Yao, J. & Xue, L. (2022), 'Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors', *Journal of Business & Economic Statistics* **40**(1), 342–354.

Zaffran, M., Féron, O., Goude, Y., Josse, J. & Dieuleveut, A. (2022), Adaptive conformal predictions for time series, *in* 'International Conference on Machine Learning', Vol. 162, PMLR, pp. 25834–25866.

Appendix

A1 Proof of Proposition 1

Proof. Considering the time series $\{y_t\}_{t\geq 1}$ generated by a locally stationary autoregressive process as defined in Equation (5). Let $\hat{y}_{t+h|t}$ be the optimal h-step-ahead point forecast generated by a well-trained model $\hat{f}$, using information available up to time t, and $e_{t+h|t}$ be the corresponding optimal h-step-ahead forecast error. Denote that $\boldsymbol{u}_{t+h} = \boldsymbol{x}_{(t-k+h):(t+h)}$. Then we have

$$\hat{y}_{t+h|t} = \begin{cases} \hat{f}(y_t, \ldots, y_{t-d+1}, \boldsymbol{u}_{t+1}) & \text{if } h = 1, \\ \hat{f}(\hat{y}_{t+h-1|t}, \ldots, \hat{y}_{t+1|t}, y_t, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h}) & \text{if } 1 < h \leq d, \\ \hat{f}(\hat{y}_{t+h-1|t}, \ldots, \hat{y}_{t+h-d|t}, \boldsymbol{u}_{t+h}) & \text{if } h > d. \end{cases}$$

For $h = 1$, we simply have $e_{t+1|t} = \omega_{t+1}$, where ω_t is a white noise series. This follows from the well-established property that optimal forecasts have 1-step-ahead errors that are white noise.

For $1 < h \leq d$, applying the first order Taylor series expansion, we can write

$$\begin{aligned} y_{t+h} &= \hat{f}(y_{t+h-1}, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h}) + \omega_{t+h} \\ &= \hat{f}(\hat{y}_{t+h-1|t} + e_{t+h-1|t}, \ldots, \hat{y}_{t+1|t} + e_{t+1|t}, y_t, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h}) + \omega_{t+h} \\ &\underset{\text{te}}{\approx} \hat{f}(\boldsymbol{a}) + \mathrm{D}\hat{f}(\boldsymbol{a})(\boldsymbol{v} - \boldsymbol{a}) + \omega_{t+h} \\ &= \hat{f}(\hat{y}_{t+h-1|t}, \ldots, \hat{y}_{t+1|t}, y_t, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h}) \\ &\quad + e_{t+h-1|t}\frac{\partial \hat{f}(\boldsymbol{a})}{\partial v_1} + \cdots + e_{t+2|t}\frac{\partial \hat{f}(\boldsymbol{a})}{\partial v_{h-2}} + e_{t+1|t}\frac{\partial \hat{f}(\boldsymbol{a})}{\partial v_{h-1}} + \omega_{t+h} \\ &= \hat{y}_{t+h|t} + e_{t+h|t}, \end{aligned}$$

where $\boldsymbol{v} = (y_{t+h-1}, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h})$, $\boldsymbol{a} = (\hat{y}_{t+h-1|t}, \ldots, \hat{y}_{t+1|t}, y_t, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h})$, $\mathrm{D}\hat{f}(\boldsymbol{a})$ denotes the matrix of partial derivative of $\hat{f}(\boldsymbol{v})$ at $\boldsymbol{v} = \boldsymbol{a}$, and $\frac{\partial}{\partial v_i}$ denotes the partial derivative with respect to the ith component in $\hat{f}$.

Similarly, for $h > d$, we can write

$$
\begin{aligned}
y_{t+h} &= \hat{f}(y_{t+h-1}, \ldots, y_{t+h-d}, \boldsymbol{u}_{t+h}) + \omega_{t+h} \\
&= \hat{f}(\hat{y}_{t+h-1|t} + e_{t+h-1|t}, \ldots, \hat{y}_{t+h-d|t} + e_{t+h-d|t}, \boldsymbol{u}_{t+h}) + \omega_{t+h} \\
&\underset{\text{te}}{\approx} \hat{f}(\boldsymbol{a}) + \mathrm{D}\hat{f}(\boldsymbol{a})(\boldsymbol{v} - \boldsymbol{a}) + \omega_{t+h} \\
&= \hat{f}(\hat{y}_{t+h-1|t}, \ldots, \hat{y}_{t+h-d|t}, \boldsymbol{u}_{t+h}) \\
&\qquad + e_{t+h-1|t}\frac{\partial \hat{f}(\boldsymbol{a})}{\partial v_1} + e_{t+h-d|t}\frac{\partial \hat{f}(\boldsymbol{a})}{\partial v_d} + \omega_{t+h} \\
&= \hat{y}_{t+h|t} + e_{t+h|t},
\end{aligned}
$$

Therefore, the forecast errors of optimal h-step-ahead forecasts follow an approximate AR(p) process, where $p = \min\{d, h-1\}$. This implies that the optimal h-step-ahead forecast errors are at most serially correlated to lag $(h-1)$. □

A2 Proof of Proposition 2

Proof. Here, we give the proof of Proposition 2 based on Proposition 1.

Based on Proposition 1 and its proof, we have

$$
\begin{aligned}
e_{t+1|t} &= \omega_{t+1} \\
e_{t+2|t} &= \omega_{t+2} + \phi_1^{(2)} e_{t+1|t} \\
e_{t+3|t} &= \omega_{t+3} + \phi_1^{(3)} e_{t+2|t} + \phi_2^{(3)} e_{t+1|t} \\
&\vdots \\
e_{t+h|t} &= \omega_{t+h} + \phi_1^{(h)} e_{t+h-1|t} + \cdots + \phi_p^{(h)} e_{t+h-p|t}, \text{ with } p = \min\{d, h-1\},
\end{aligned}
$$

where ω_t is a white noise series, $\phi_i^{(j)}$ denotes the coefficient for the lag i term in the AR model of order $\min\{d, j-1\}$ for the forecast error $e_{t+j|t}$ and here the AR model is applied at the forecast horizon j, rather than at the time index t.

Substituting all equations above into the following equation, we can obtain

$$e_{t+h|t} = \omega_{t+h} + \sum_{i=1}^{h-1} \theta_i \omega_{t+h-i}, \text{ for each } h \in [H],$$

where θ_i is a complex combination of ϕ terms from the previous $h-1$ equations. So we conclude that the h-step-ahead forecast error sequence $\{e_{t+h|t}\}$ follows an approximate MA$(h-1)$ process.

□

A3 Proof of Proposition 3

Proof. Let $E_T = \sum_{t=h+1}^{T} (\text{err}_{t|t-h} - \alpha)$. The inequality given by Equation (9) can be expressed as $|E_T| \leq c \cdot g(T-h) + h$. We will prove one side of the absolute inequality, specifically $E_T \leq c \cdot g(T-h) + h$, with the other side following analogously. We proceed with the proof using induction.

For $T = h+1, \ldots, 2h$, $E_T = \sum_{t=h+1}^{T} (\text{err}_{t|t-h} - \alpha) \leq (T-h) - (T-h)\alpha \leq T-h \leq h \leq cg(T-h) + h$ as $c > 0$, $h \geq 1$, g is nonnegative, and $\text{err}_{t|t-h} \leq 1$. Thus, Equation (9) holds for $T = h+1, \ldots, 2h$.

Now, assuming Equation (9) is true up to T. We partition the argument into $h+1$ cases:

$$\begin{cases} cg(T-h) + h - 1 < E_T \leq cg(T-h) + h, & \text{case (1)} \\ cg(T-h) + h - 2 < E_T \leq cg(T-h) + h - 1, & \text{case (2)} \\ \quad \vdots & \\ cg(T-h) < E_T \leq cg(T-h) + 1, & \text{case (h)} \\ E_T \leq cg(T-h). & \text{case (h+1)} \end{cases}$$

In case (1), we observe that $E_T > cg(T-h) + h - 1 > cg(T-h)$, implying $q_{T+h|T} = r_t(E_T) \geq b$ according to Equation (4). Thus, $s_{T+h|T} \leq q_{T+h|T}$ and $\text{err}_{T+h|T} = 0$. Furthermore, we have $E_{T-1} = E_T - (\text{err}_{T|T-h} - \alpha) > cg(T-h) + h - 2 > cg(T-h-1)$ as g is nondecreasing. This implies $q_{T+h-1|T-1} = r_t(E_{T-1}) \geq b$, hence $s_{T+h-1|T-1} \leq q_{T+h-1|T-1}$ and $\text{err}_{T+h-1|T-1} = 0$. Similarly,

$E_{T-2} = E_{T-1} - (\text{err}_{T-1|T-h-1} - \alpha) > cg(T-h) + h - 3 > cg(T-h-2)$, thus $\text{err}_{T+h-2|T-2} = 0$. This process iterates, leading to $\text{err}_{T+h|T} = \text{err}_{T+h-1|T-1} = \cdots = \text{err}_{T+1|T-h+1} = 0$. Consequently,

$$E_{T+h} = E_T + \sum_{t=T+1}^{T+h} (\text{err}_{t|t-h} - \alpha) \leq cg(T-h) + h - h\alpha \leq cg(T) + h,$$

which is the desired result at $T+h$.

In case (2), we observe that $E_T > cg(T-h) + h - 2 > cg(T-h)$, thus $s_{T+h|T} \leq q_{T+h|T}$ and $\text{err}_{T+h|T} = 0$. Moving forward, we have $\text{err}_{T+h|T} = \text{err}_{T+h-1|T-1} = \cdots = \text{err}_{T+2|T-h+2} = 0$. Along with $\text{err}_{T+1|T-h+1} \leq 1$, this means that

$$E_{T+h} = E_T + \sum_{t=T+1}^{T+h} (\text{err}_{t|t-h} - \alpha) \leq cg(T-h) + h - 1 + 1 - h\alpha \leq cg(T) + h,$$

which again gives the desired result at $T+h$.

Similarly, in cases (3)-(h), we can always get the desired result at $T+h$.

In case (h+1), noticing $E_T \leq cg(T-h)$, and simply using $\text{err}_{T+h-i|T-i} \leq 1$ for $i = 0, \ldots, h-1$, we have

$$E_{T+h} = E_T + \sum_{t=T+1}^{T+h} (\text{err}_{t|t-h} - \alpha) \leq cg(T-h) + h - h\alpha \leq cg(T) + h.$$

Therefore, we can deduce the desired outcome at any $T \geq h+1$. This completes the proof for the first part of Proposition 3.

Regarding the second part, $g(t-h)/(t-h) \to 0$ as $t \to \infty$ due to the sublinearity of the admissible function g. Hence, the second part holds trivially. □

A4 Proof of Corollary 1

Proof. We set $q_{2h|h} = 0$ without losing generality, the iteration $q_{t+h|t} = q_{t+h-1|t-1} + \eta(\text{err}_{t|t-h} - \alpha)$ simplifies to $q_{t+h|t} = \eta \sum_{i=h+1}^{t} (\text{err}_{i|i-h} - \alpha)$. Let $r_t(x) = \eta x$ and the admissible function $g(t-h) = b$, Equation (4) holds for $c = \frac{1}{\eta}$. Then Proposition 3 applies and we can easily derive the desired result. □

A5 Proof of Corollary 2

Proof. Let $q^*_{t+h|t} = q_{t+h|t} - \hat{q}_{t+h|t}$, then Equation (10) transforms into an update process $q^*_{t+h|t} = r_t\left(\sum_{i=h+1}^{t}(\mathrm{err}_{i|i-h} - \alpha)\right)$, which is an update with respect to $q^*_{t+h|t}$. Under this new framework, the nonconformity score becomes $s^*_{t+h|t} = s_{t+h|t} - \hat{q}_{t+h|t}$, with values ranging in $[-b, b]$, given the assumption that both $s_{t+h|t}$ and $\hat{q}_{t+h|t}$ fall within $[-\frac{b}{2}, \frac{b}{2}]$. Thus, Proposition 3 can be directly applied to establish the long-run coverage achieved by the AcMCP method. □