Online conformal inference for multi-step time series forecasting

Xiaoqian Wang
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: xiaoqian.wang@monash.edu
Corresponding author

Rob J Hyndman
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: rob.hyndman@monash.edu

13 October 2024

JEL classification: C53,C22,C14

Online conformal inference for multi-step time series forecasting

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 ensures theoretical long-run coverage guarantees for multi-step prediction intervals, though we note that increased forecasting horizons may exacerbate deviations from the target coverage, particularly in the context of limited sample sizes. 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.

1 Introduction

Conformal prediction (**vovk2005**) is a powerful and flexible tool for uncertainty quantification, distinguished by its simplicity, generality, and ease of implementation. 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 applicable for quantifying the uncertainty of predictions produced by any black-box machine learning model (**shafer2008**; **papadopoulos2008**; **barber2021**) or non-parametric model (**lei2014**).

Three key classes of conformal prediction methods are widely used for constructing distribution-free prediction intervals: split conformal prediction (**vovk2005**), full conformal prediction (**vovk2005**), and jackknife+ (**barber2021**). The split conformal method, which relies on a holdout set, offers computational efficiency but sacrifices some statistical efficiency due to data splitting. In contrast, full conformal prediction avoids data splitting, providing higher accuracy at the cost of increased computational complexity. Jackknife+ strikes a balance between these methods, offering a compromise between statistical precision and computational demands. All three methods guarantee coverage at the target level under the assumption of data exchangeability.

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 accommodate non-exchangeable data. Notable examples include methods for handling covariate shift (**tibshirani2019**; **lei2021**; **yang2024**), online distribution shift (**gibbs2021**; **zaffran2022**; **bastani2022**), label shift (**podkopaev2021**), time series data (**chernozhukov2018**; **gibbs2021**; **xu2021**; **xu2023**; **zaffran2022**), and spatial prediction (**mao2024**), and methods based on certain distributional assumptions of the data rather than exchangeability (**oliveira2024**; **xu2021**; **xu2023**). Additionally, some methods propose weighting the nonconformity scores (for example, prediction errors) differently, either using non-data-dependent weights (**barber2023**) or weights based on observed feature values (**tibshirani2019**; **guan2023**; **hore2023**).

Recently, several attempts have been made to enable conformal prediction on time series data, 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, **chernozhukov2018** and **yu2022** 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, **xu2021** and **xu2023** extend conformal prediction methods under classical nonparametric assumptions to achieve asymptotic valid conditional coverage for time series. **barber2023** use weighted residual distributions to provide robustness against distribution drift. Additionally, **oliveira2024** 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, **gibbs2021** develop adaptive conformal inference 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. Follow-up work has refined this idea by introducing time-dependent step sizes to respond to arbitrary distribution shifts, as seen in studies by **bastani2022**, **zaffran2022**, and **gibbs2024**. However, these methods may produce prediction intervals that are either infinite or null. To address this issue, recent research has proposed a generalized updating process that tracks the quantile of the nonconformity score sequence, as discussed by **bhatnagar2023**, **angelopoulos2024**, and **angelopoulos2024online**.

Existing conformal prediction methods for time series primarily focus on single-step forecasting. However, many applications require forecasts for multiple future time steps, not just one. Related research into multi-step time series forecasting is limited and does not account for the temporal dependencies inherent in multi-step forecasts. For example, **stankeviciute2021** integrate conformal prediction with recurrent neural networks for multi-step forecasting and then apply Bonferroni correction to achieve the desired coverage rate. This approach, however, assumes data independence, which is not often unrealistic for time series data. **yang2024ts** propose Bellman conformal inference, an extension of adaptive conformal prediction, to control multi-step miscoverage rates simultaneously at each time point t by minimizing a loss function that balances the average interval length across forecast horizons with miscoverage. While this method considers multi-step intervals, it does not account for their temporal dependencies and may be computationally intensive when solving the associated optimization problems at each time step. Additionally, several extensions to multivariate targets have been explored, see, e.g., **schlembach2022** and **sun2022**.

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 distribution of $\{(\boldsymbol{x}_t, y_t)\}_{t\geq 1} \subseteq \mathbb{R}^p \times \mathbb{R}$ is allowed to vary over time, to accommodate 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 the output $\hat{\mathscr{C}}_{t+h|t}$ to be a *valid* prediction interval so that y_{t+h} falls within the prediction interval $\hat{\mathscr{C}}_{t+h|t}$ at least $100(1-\alpha)\%$ of the time.

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 accommodate multi-step forecasting scenarios. Second, we provide theoretical proofs demonstrating that the forecast errors of optimal h-step-ahead forecasts approximate a linear combination of at most its lag $(h-1)$ with respect to forecast horizon when we assume a general non-stationary autoregressive data generating process. Third, we introduce the autocorrelated multi-step conformal prediction method, which accounts for the autocorrelations of multi-step forecast errors. Our method is proven to achieve long-run coverage guarantees without making any assumptions on data distribution shifts. We also highlight that for $t \ll \infty$, increasing the forecast horizon h generally leads to greater deviations from the target coverage, which aligns with our expectations. Finally, we illustrate the practical utility of these proposed methods through two simulations and two applications to electricity demand and eating-out expenditure forecasting.

We developed the `conformalForecast` package for R, available at https://github.com/xqnwang/conformalForecast, to implement the proposed multi-step conformal prediction methods. 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, we have a forecasting model $\hat{f}_t$ trained using the historical data $z_{1:t}$. Throughout the paper, we assume that the predictors are known into the future. In this way, we perform ex-post forecasting and there is no additional uncertainty introduced from forecasting the exogenous predictors. 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

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

The task is to employ conformal inference to build H-step prediction intervals, $\{\hat{\mathscr{C}}^{\alpha}_{t+h|t}(z_{1:t}, \boldsymbol{x}_{t+1:h})\}_{h\in[H]}$, 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. Instead, 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* $\mathscr{D}_{\text{tr}} \subset \{1,\dots,t_r\}$ and a *calibration set* $\mathscr{D}_{\text{cal}} \subset \{t_r+1,\dots,t\}$, where $t_c = t - t_r \gg H$, $\mathscr{D}_{\text{tr}} \cup \mathscr{D}_{\text{cal}} = \{1,\dots,t\}$, 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 (**papadopoulos2007**). 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 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.

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.

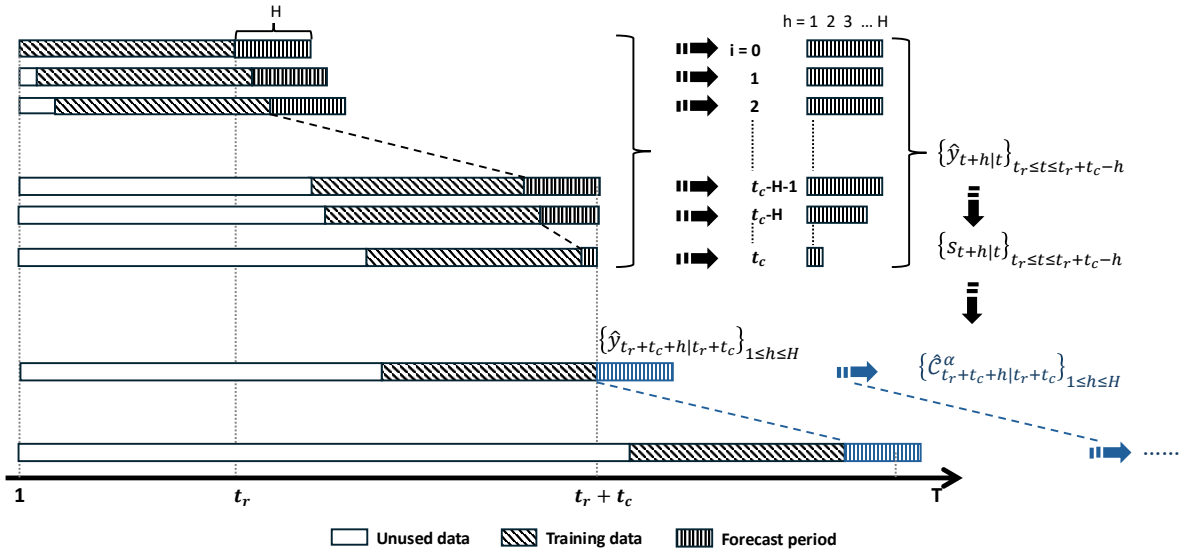

Figure 1: *Diagram of the online learning framework with sequential splits.*

3 Extending conformal prediction methods to multi-step forecasting

In this section, we apply the online learning framework outlined in Section 2 to adapt several popular conformal prediction methods in order to allow for multi-step forecasting problems. In time series forecasting, when a model is properly specified and well-trained, the forecasts that minimize the mean squared error (MSE) can be considered optimal in the sense that they achieve the lowest possible expected squared deviation between the forecasts and actual values. One of the key properties of optimal forecast errors, which holds generally in linear models, is that the variance of the forecast error $e_{t+h|t}$ is non-decreasing in h (**tong1990**; **diebold1996**; **patton2007**). Therefore, we need to apply a separate conformal prediction procedure for each $h \in [H]$.

3.1 Online multi-step split conformal prediction

Split conformal prediction (**papadopoulos2002**; **vovk2005**; **lei2018**), is a holdout method for building prediction intervals using a pre-trained model in regression settings. A key advantage of SCP is its ability to guarantee coverage by assuming data exchangeability. 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. Instead of assuming exchangeability, 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:

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

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, $Q_{\tau}(\cdot)$ denotes the τ-quantile of its argument, and δ_a denotes the point mass at a.

3.2 Online multi-step weighted conformal prediction

barber2023 propose a nonexchangeable conformal prediction procedure (NexCP) that generalizes the SCP method to allow for some sources of nonexchangeability. The core idea is that a higher weight should be assigned to a data point that is believed to originate from the same distribution as the test data. Note that NexCP assumes the weights are fixed and data-independent. When the data are exchangeable, NexCP offers the same coverage guarantees as SCP, while the coverage gap is characterized by the total variation between the swapped nonconformity score vectors when exchangeability is violated. Thus the coverage gap may be quite large in time series contexts.

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.

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 (**hyndman2021**). 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 prediction (ACP) method proposed by **gibbs2021** 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 realized 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, we raise the estimate of $\alpha^{*}_{t+h|t}$ used for quantile estimation at time $t+h$ if $\hat{\mathscr{C}}^{\alpha_{t|t-h}}_{t|t-h}$ covered y_t, whereas we lower the estimate if $\hat{\mathscr{C}}^{\alpha_{t|t-h}}_{t|t-h}$ did not cover y_t. Thus 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. **gibbs2021** 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, **zaffran2022** use an adaptive aggregation of multiple ACPs with a set of candidate values for γ , determining weights based on their historical performance. **bastani2022** propose a multivalid prediction algorithm in which the prediction set is established by selecting a threshold from a sequence of candidate thresholds. However, both previous methods fail to promptly adapt to the local changes. To address this limitation, **gibbs2024** suggest adaptively tuning the step size parameter γ in an online setting, choosing an "optimal" value for γ from a candidate set of values by assessing their historical performance.

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 (**angelopoulos2024**), originally developed for one-step-ahead forecasting. For each individual forecast horizon $h \in [H]$, the iteration of the h-step-ahead quantile estimate is given by

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

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

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

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 Proposition 4.5.

The P control in MPID shows a delayed correction of the quantile estimate for a length of h periods. 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 accounts for the cumulative historical coverage errors associated with h-step-ahead prediction intervals during the update process, 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") trained using the h-step-ahead nonconformity scores available up to and including time t. This module provides additional benefits

only when the scorecaster is well-designed and there are predictable patterns in the nonconformity scores; otherwise, it may increase variability in the coverage and prediction intervals.

4 Autocorrelated multi-step conformal prediction

In the PID method proposed by **angelopoulos2024**, 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 $e_{t+h|t}$ will be correlated with the forecast errors from the past $h-1$ steps, as forecast errors accumulate 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.

4.1 Properties of multi-step forecast errors

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

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

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.

Proposition 4.1 (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. Therefore, the optimal h-step-ahead forecast errors are at most serially correlated to lag $(h-1)$.

Proposition 4.1 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 4.1, provided in Section A.1, 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 (**harvey1997**; **diebold2024**), a property that can be derived using Wold's representation theorem. Here, we extend the investigation to time series generated by a general non-stationary autoregressive process, and provide the corresponding proposition. The proof of this proposition is provided in Section A.2.

Proposition 4.2 (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. Then 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.

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 is given by

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

for $h \in [H]$. Obviously, the AcMCP method can be viewed as a further extension of the PID method. 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. $e_{t+h-1|t},\ldots,e_{t+1|t}$). Thus, we perform multi-step conformal prediction recursively, contrasting with the independent approach employed in MPID. Moreover, the inclusion of $\tilde{e}_{t+h|t}$ is not intended to forecast the nonconformity scores (i.e., forecast errors), but rather to incorporate autocorrelations present in multi-step forecast errors within the resulting multi-step prediction intervals.

4.3 Coverage guarantees

Proposition 4.3. *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., $\lim_{T\to\infty}\frac{1}{T-h}\sum_{t=h+1}^{T}\mathrm{err}_{t|t-h} = \alpha$.

Proposition 4.3 indicates that, for $t \ll \infty$, increasing the forecast horizon h tends to amplify 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. This is consistent with expectations, as extending the forecast horizon generally increases forecast uncertainty. As predictions extend further into the future, more factors contribute to variability and uncertainty. In this case, conformal prediction intervals may not scale perfectly with the increasing uncertainty, leading to a larger discrepancy between the desired and actual 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 4.3 if we set $q_{2h|h} = 0$ without losing generality. Thus, its coverage bounds can be easily derived as a result of Proposition 4.3.

Proposition 4.4. *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., $\lim_{T\to\infty}\frac{1}{T-h}\sum_{t=h+1}^{T}\mathrm{err}_{t|t-h} = \alpha$.

More importantly, Proposition 4.3 is also adequate for establishing the coverage guarantee of the proposed AcMCP method given by Equation 8. Following **angelopoulos2024**, 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 $\{(\boldsymbol{x}_i, y_i)\}_{1\leq i\leq t}$ 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 coverage guarantee for AcMCP given by Equation 10.

Proposition 4.5. *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., $\lim_{T\to\infty}\frac{1}{T-h}\sum_{t=h+1}^{T}\mathrm{err}_{t|t-h} = \alpha$.*

5 Experiments

We examine the empirical performance of the proposed multi-step conformal prediction methods using two simulated data settings and two real data examples.

Throughout the experiments, we adhere to the following parameter settings:

- we focus on the target coverage level $1-\alpha = 0.9$;
- for the MWCP method, we use $b = 0.99$ as per **barber2023**;
- following **angelopoulos2024**, we use a step size parameter of $\gamma = 0.005$ for the MACP method, a Theta model as the scorecaster in the MPID method, and a learning rate of $\eta = 0.01\hat{B}_t$ for quantile tracking in the MPID and AcMCP methods, 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 and the window length Δ is set to be same as the length of the calibration set;
- we adopt a nonlinear saturation function defined as $r_t(x) = K_1 \tan(x \log(t)/(tC_{\text{sat}}))$, where $\tan(x) = \text{sign}(x) \cdot \infty$ for $x \notin [-\pi/2, \pi/2]$, and constants $C_{\text{sat}}, K_1 > 0$ are chosen heuristically, as suggested by **angelopoulos2024**;
- we consider a clipped version of MACP by imputing infinite intervals with the largest score seen so far.

5.1 Simulated linear autoregressive process

We first consider a simulated stationary time series which is generated from a simple 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 $\mathcal{D}_{\text{tr}}$ and calibration sets $\mathcal{D}_{\text{cal}}$, 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 (**hyndman2024**). 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 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. The trajectories of the rolling mean and median of the interval widths for each method are largely consistent. AcMCP constructs narrower prediction intervals than MPID, despite both methods achieving similar coverage. Moreover, we see that AcMCP tends to offer adaptive prediction intervals, 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. However, MPID achieves tight coverage but at the cost of constructing wider prediction intervals. This is due 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.

The inclusion of the last term $\tilde{e}_{t+h|t}$ in AcMCP should only result in a slight difference compared to

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 black dashed line indicates the target level of* $1-\alpha = 0.9$.

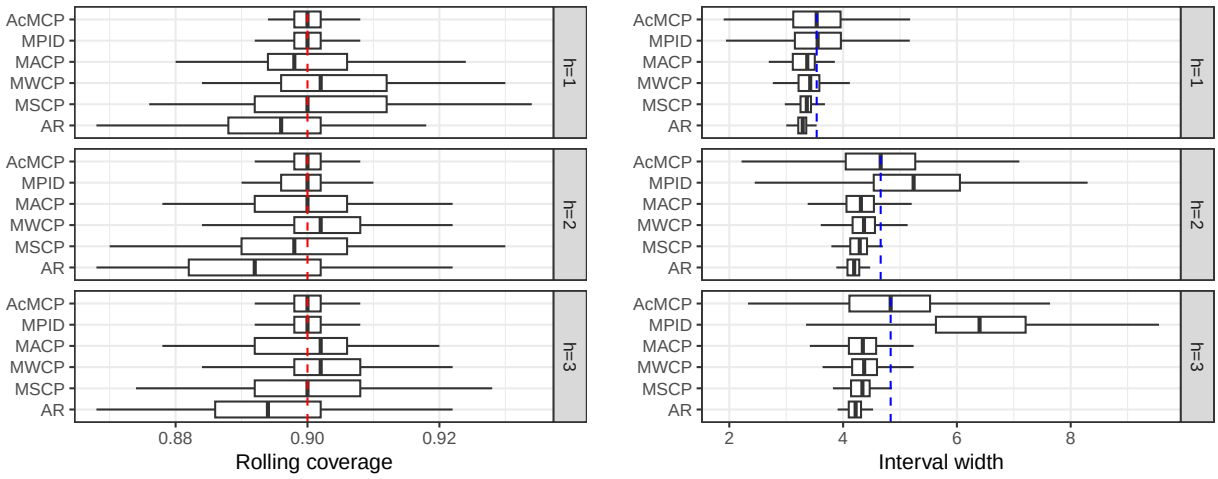

Figure 3: *AR(2) simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

the 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. 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.

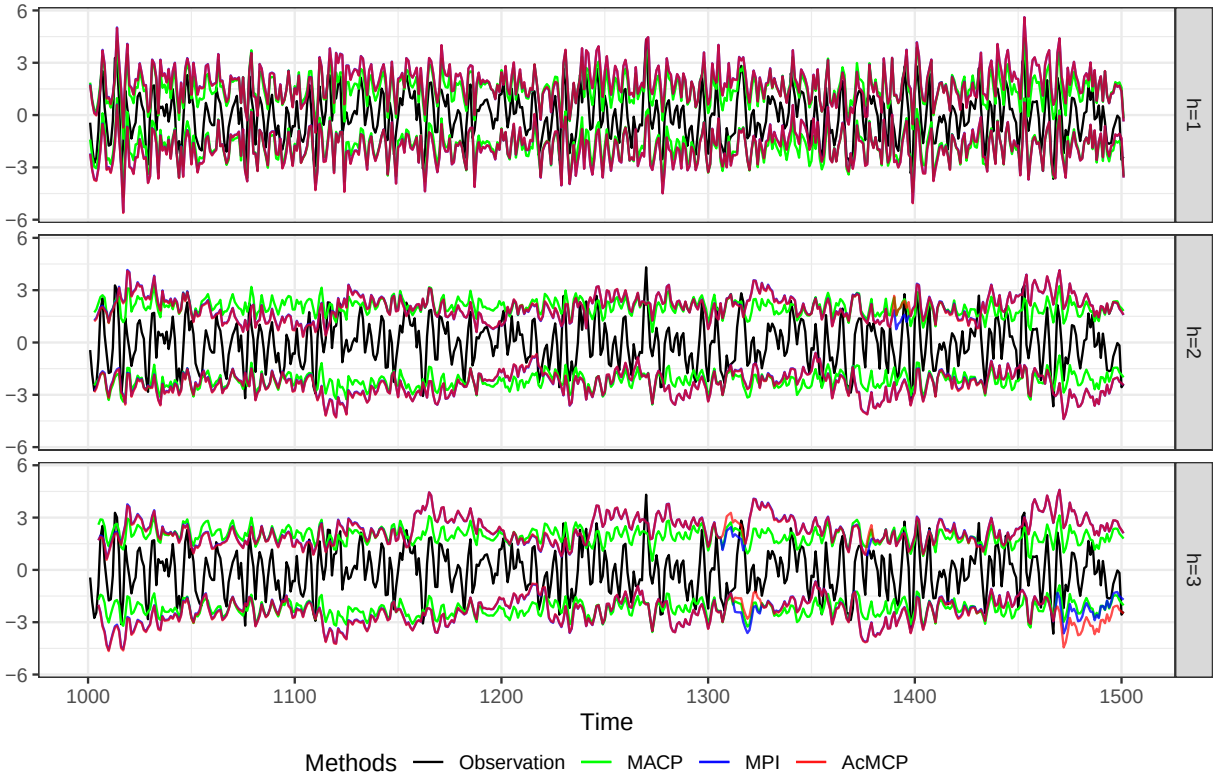

Figure 4: *AR(2) simulation results showing the prediction interval bounds for the MACP, MPI, and AcMCP methods over a truncated period of length 500.*

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 error variance $\sigma^2 = 0.1$. Thus, the time series y_t nonlinearly depends on its lagged values y_{t-1} and y_{t-2}, as well as 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, we create training sets $\mathscr{D}_{\text{tr}}$ and calibration sets $\mathscr{D}_{\text{cal}}$, each with a length of 500. Given the nonlinear structure of the DGP, we use feed-forward neural networks with a single

hidden layer and lagged inputs to generate 1- to 3-step-ahead point forecasts (i.e. $H = 3$), using the `nnetar()` function from the `forecast` R package (**hyndman2024**). Note that it is not possible to derive analytic intervals for neural networks, thus we do not include neural network models when presenting the results.

Figure 5 illustrates the rolling coverage and interval width of each method, with calculations based on a rolling window of size 100. We see that MPID and AcMCP are able to maintain minor fluctuations around the target coverage of 90% across all time indices, contrasting with MSCP, MWCP, and MACP, which struggle to sustain the target coverage and display pronounced fluctuations over time. Moreover, all conformal prediction methods, except for MSCP, construct adaptive prediction intervals. They widen intervals in response to undercoverage and narrow them when overcoverage occurs. Notably, MPID and AcMCP demonstrate greater adaptability, displaying higher variability in interval widths compared to competing methods in order to uphold the desired coverage. AcMCP intervals are evidently narrower than MPID intervals for 2-step-ahead forecasting but wider for 3-step-ahead forecasting. AcMCP intervals appear to be more reasonable, as they tend to widen with increasing forecast horizons.

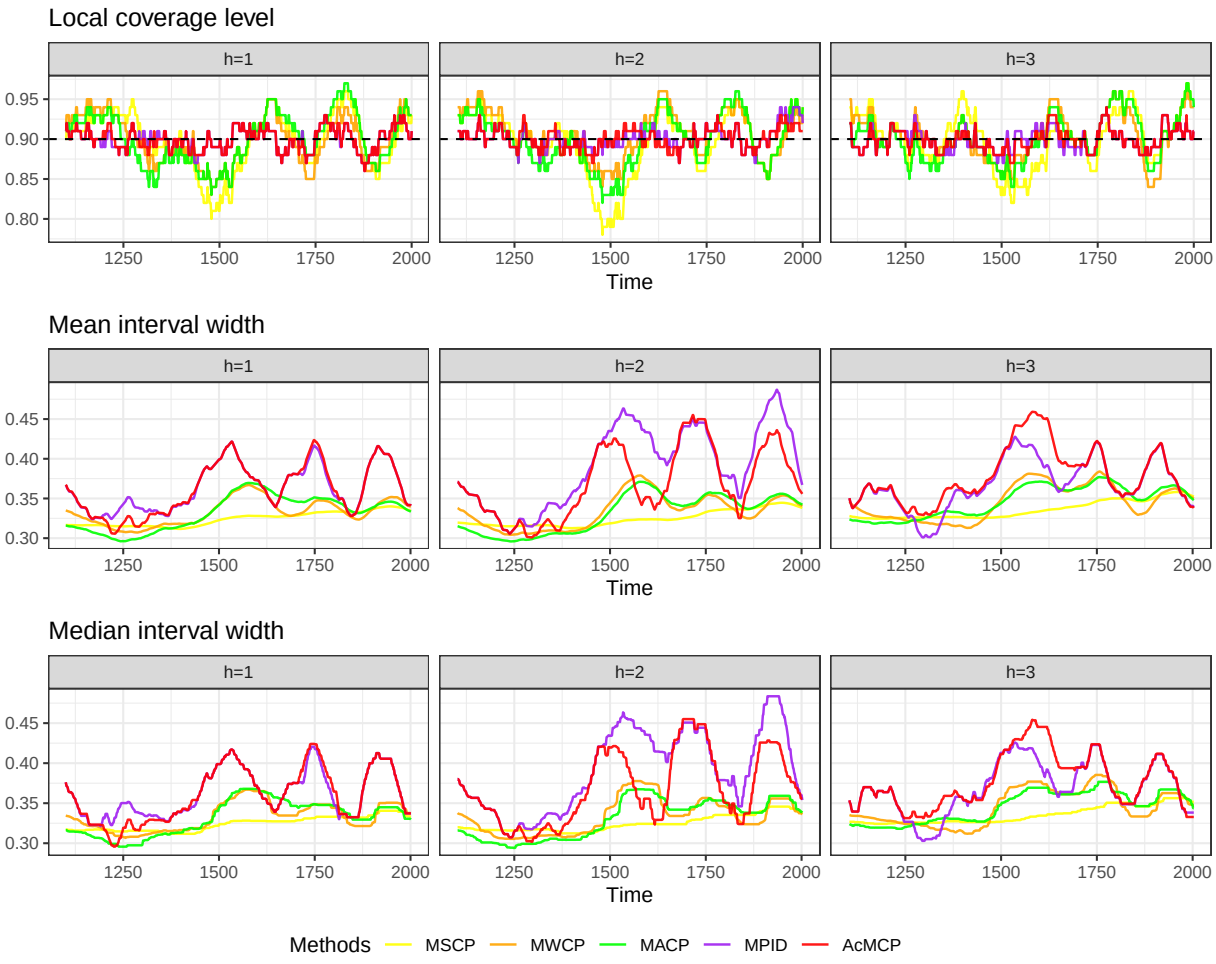

Figure 5: *Nonlinear simulation results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 100. *The black dashed line indicates the target level of* $1 - \alpha = 0.9$.

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.

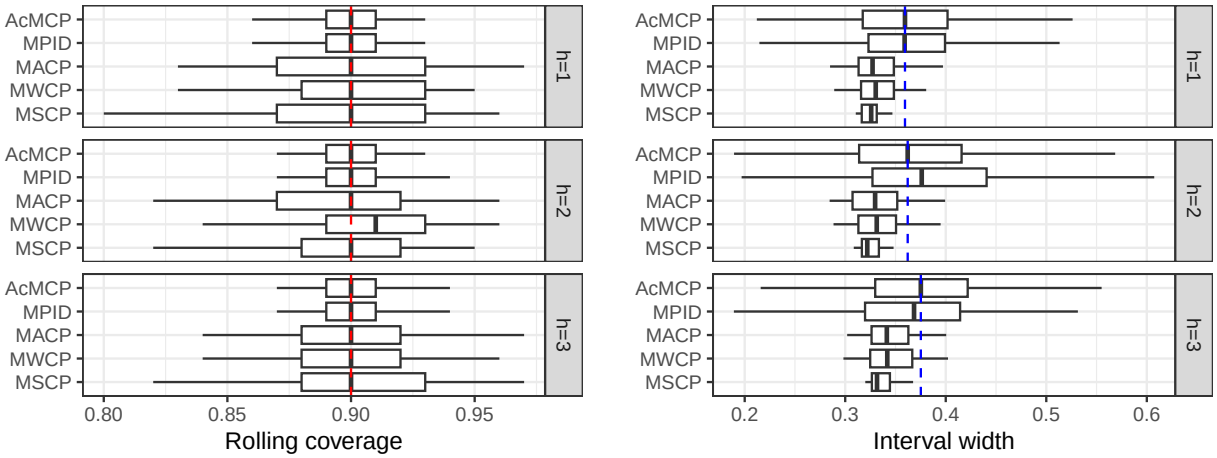

Figure 6: *Nonlinear simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

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 the state of Victoria, Australia, spanning a three-year period from 2012 to 2014. Temperatures are taken from the Melbourne Regional Office weather station in the capital city of Victoria. The left panel of Figure 7 displays the daily electricity demand during 2012–2014, along with temperatures. The right panel shows a nonlinear relationship between electricity 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 in the right panel correspond to working days and non-working days.

Our response variable is `Demand`, and we use two covariates: `Temperature`, and `Workday` (an indicator variable for if the day was a working day or not). Following **hyndman2021**, we will fit a dynamic regression model with a piecewise linear function of temperature (containing a knot at 18 degrees) to generate 1- to 7-step-ahead point forecasts (i.e. $H = 7$). The error series in the regression

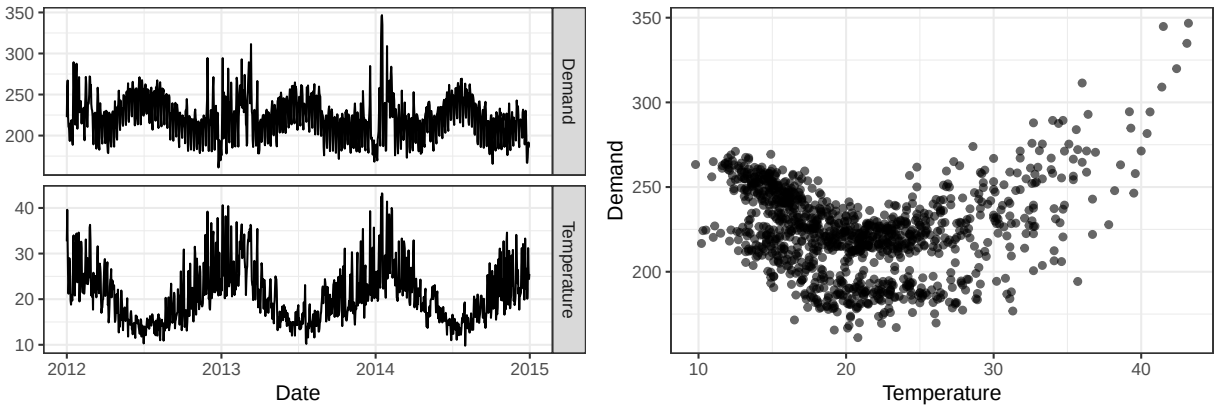

Figure 7: *Daily electricity demand and corresponding daily maximum temperatures in 2012–2014, Victoria, Australia.*

is assumed to follow an ARIMA model to contain autocorrelation. We use two years of data as training sets to fit dynamic regression models, and use 100 data points for calibration sets.

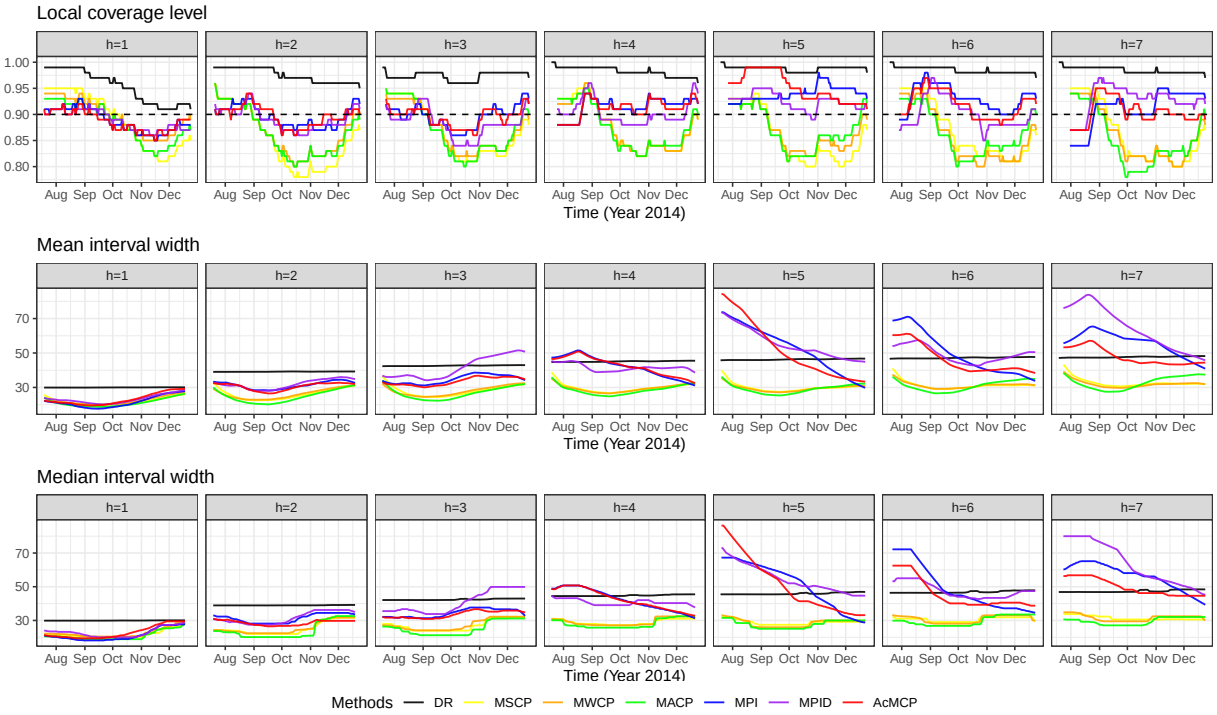

Figure 8: *Electricity demand data results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 100*. The black dashed line indicates the target level of* $1 - \alpha = 0.9$.

We present the results in Figure 8 and Figure 9, comparing the rolling coverage and interval width of each 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, we observe that DR consistently achieves a significantly higher coverage than the 90% target coverage, resulting in much wider intervals than other methods for $h = 1, 2, 3, 4$. Second, MSCP, MWCP, and MACP fail to sustain the target coverage and noticeably undercover after September 2014 for all forecast horizons, thus leading to narrower intervals than others. Third, MPI, MPID, and AcMCP offer prediction intervals that are wider than those of other conformal prediction methods, effectively mitigating or avoiding the undercoverage issue observed after September 2014. Additionally, we notice that MPID performs slightly worse than MPI and AcMCP in terms of coverage for $h = 3, 4, 7$, despite leading to wider intervals. Finally, MPI and AcMCP coverage display similar pattern, but AcMCP is capable of constructing narrower intervals than MPI.

We present the forecast interval bounds for MACP, MPI, and AcMCP in Figure 10. The plot shows that MACP intervals are too narrow to adequately hug the true values, particularly from September to November 2014. In contrast, MPI and AcMCP perform better by widening their intervals, resulting in narrower swings away from the 90% target level. The differences between MPI and AcMCP prediction intervals are most pronounced in the 5-, 6-, and 7-step-ahead forecast results. For 5-step-ahead forecasting, from May to July 2014, the upper bounds of MPI intervals struggle to capture the true values. AcMCP addresses this undercoverage by construct larger upper bounds. In August and September, AcMCP constructs smaller upper bounds than MPI while still capturing the true values effectively. For 6-step-ahead forecasting from May to July, AcMCP offers smaller upper bounds than

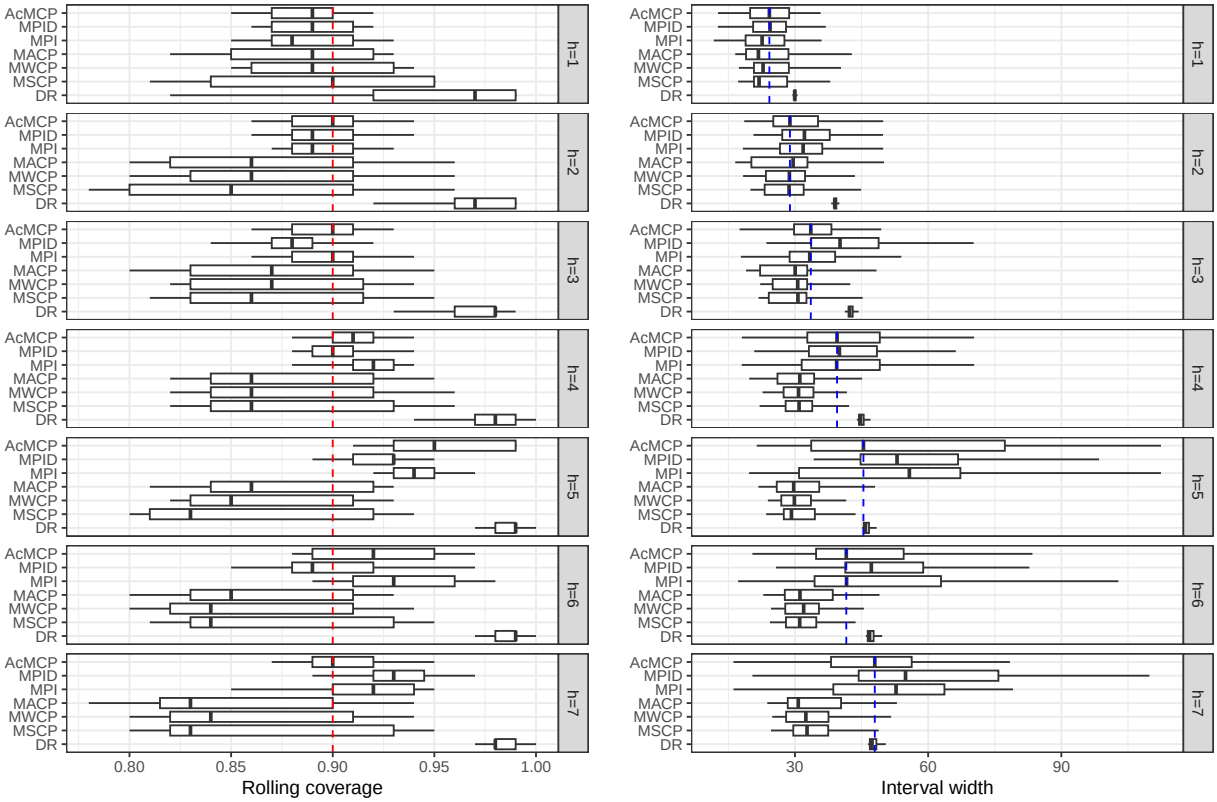

Figure 9: *Electricity demand data results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

MPI, which provides upper bounds that are far away from the truth. Similar reaction is observed for 7-step-ahead forecasting during August and September.

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 set includes monthly expenditure on cafes, restaurants and takeaway food services in Victoria from April 1982 up to December 2019, as shown in Figure 11. The data shows an overall upward trend, obvious annual seasonal patterns, and variability proportional to the data level.

We consider three models: ARIMA with logarithmic transformation, ETS, and STL-ETS (**hyndman2021**), and then output their simple average as final point forecasts. STL-ETS involves forecasting using the STL decomposition method, applying ETS to forecast the seasonally adjusted series. All three models can be automatically trained using the `forecast` R package (**hyndman2024**). Our goal is to forecast 12 months ahead, i.e. $H = 12$. We use 20 years of data for training the models and 5 years of data for calibration sets. The whole test set only has a length of 152 months. Therefore, we will not compute rolling coverage and interval width in this experiment, but rather compute the coverage and interval width averaged over the entire test set.

We summarize the average coverage and interval width for each method and each forecast horizon in Figure 12. 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

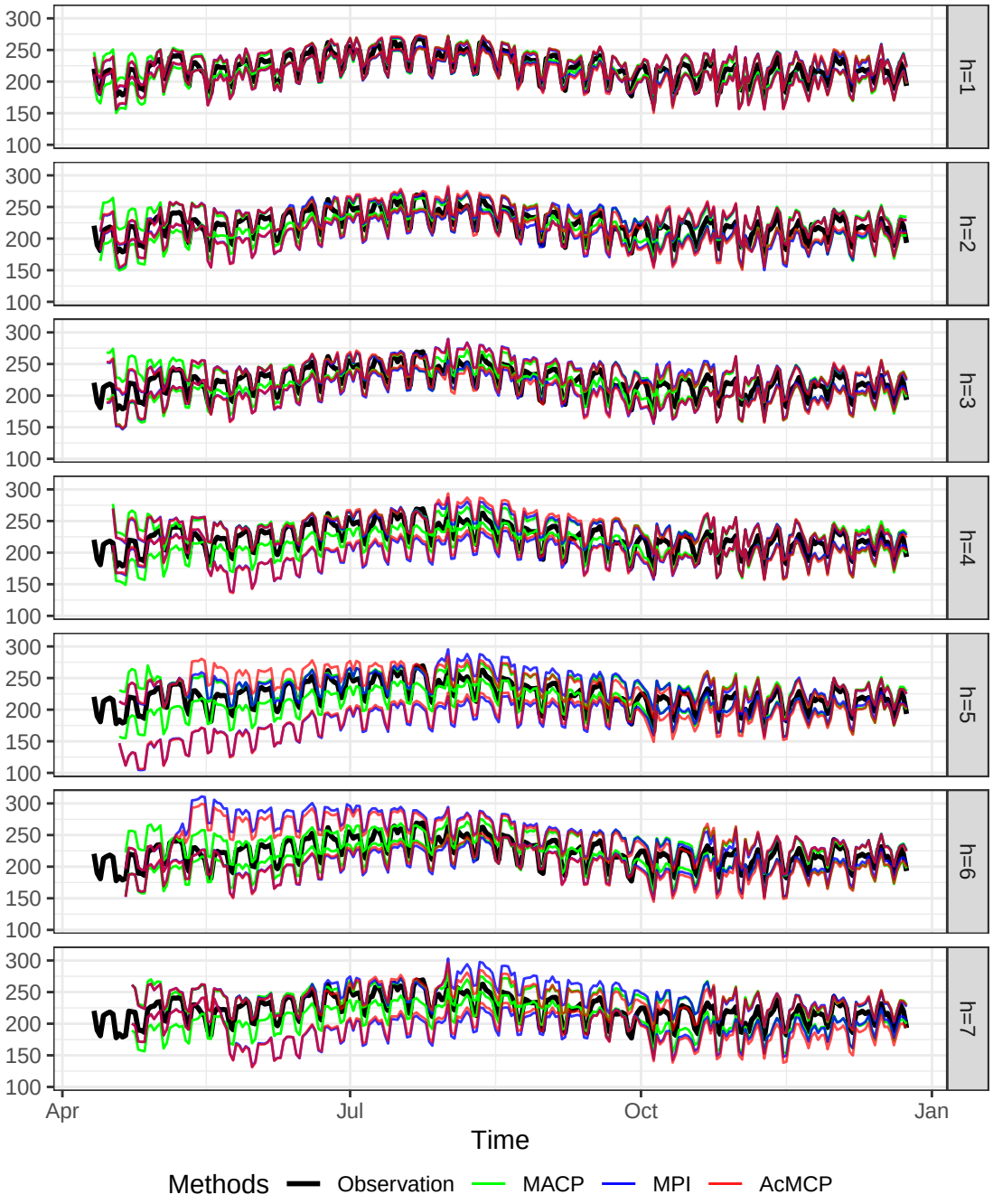

Figure 10: *Electricity demand data results showing the forecast interval bounds for MACP, MPI, and AcMCP over the whole test set.*

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.

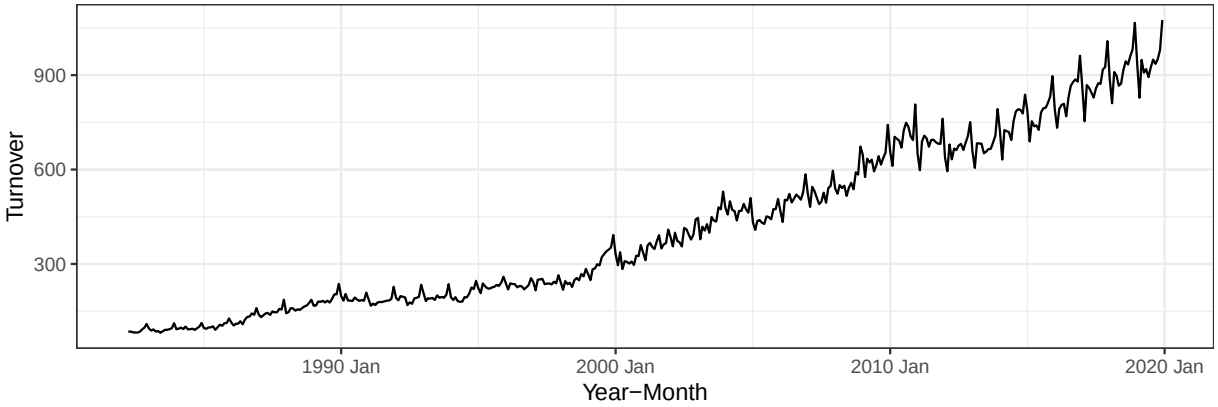

Figure 11: *Monthly expenditure on cafes, restaurants and takeaway food services in Victoria, Australia, from April 1982 to December 2019.*

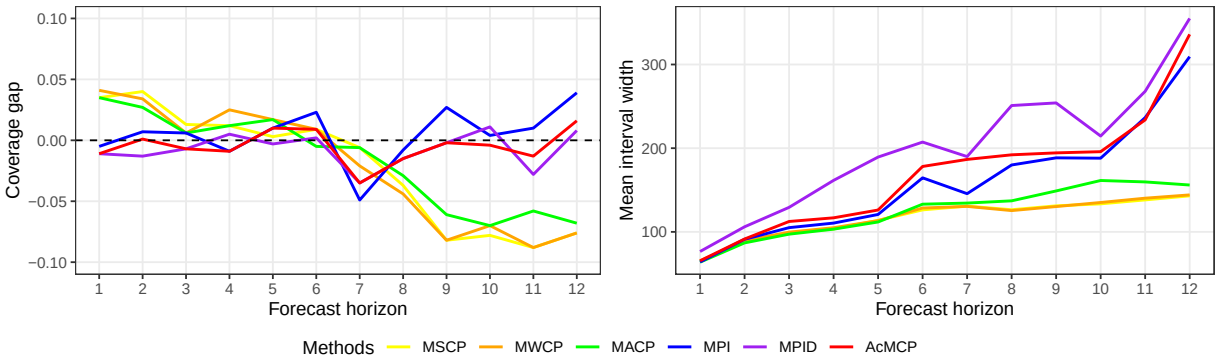

Figure 12: *Eating out expenditure data results showing coverage gap and interval width averaged over the entire test set for each forecast horizon. The black dashed line in the top panel indicates no difference from the* 90% *target level.*

6 Conclusion and discussion

We have introduced a unified notation to formalize the mathematical representation of conformal prediction within the context of time series data, with a particular focus on multi-step time series forecasting in a generic online learning framework.

Several accessible conformal prediction methods have been extended to address the challenges of multi-step forecasting scenarios.

We have shown 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, under the assumption of a general non-stationary autoregressive DGP. Building on this foundation, we introduce a novel method, AcMCP, which accounts for the autocorrelations inherent in multi-step forecast errors. Our method achieves long-run coverage guarantees without imposing assumptions regarding data distribution shifts. In both simulations and applications to data, our proposed method achieves coverage closer to the target within local windows while offering adaptive prediction intervals that adjust effectively to varying conditions.

The methods discussed here are limited to ex-post forecasting, operating under the assumption that actual values of the exogenous predictors during the forecast period are available. In many applications, this may not be true and the methods will need to be adapted to allow forecasts of the predictors to be used, and the resulting uncertainty to be included in the conformal inference.

Additionally, our methodology does not incorporate an algorithmic approach to tuning the learning rate parameter.

These considerations pave the way for numerous avenues for future research. Potential directions include the introduction of a time-dependent learning rate parameter to minimize interval width while maintaining the guaranteed coverage rate, as well as the development of refined methodologies that account for variability introduced by forecasting predictors in ex-ante scenarios.

References

Appendix A Proofs

A.1 Proof of Proposition 4.1

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

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

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 $(h-1)$. □

A.2 Proof of Proposition 4.2

Proof. Here, we give the proof of Proposition 4.2 based on Proposition 4.1; the idea is motivated by **sommer2023**.

Based on Proposition 4.1 and its proof, we have

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

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. □

A.3 Proof of Proposition 4.3

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 $|E_T| \le c \cdot g(T-h) + h$. We will prove one side of the absolute inequality, specifically $E_T \le c \cdot g(T-h) + h$, with the other side following analogously. We proceed with the proof using induction.

For $T = h+1, \ldots, 2h$, $E_T = \sum_{t=h+1}^{T} (\text{err}_{t|t-h} - \alpha) \le (T-h) - (T-h)\alpha \le T-h \le h \le cg(T-h) + h$ as $c > 0$, $h \ge 1$, g is nonnegative, and $\text{err}_{t|t-h} \le 1$. 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:

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

In case (1), we observe that $E_T > cg(T-h) + h - 1 > cg(T-h)$, implying $q_{T+h|T} = r_t(E_T) \ge b$ according to Equation 4. Thus, $s_{T+h|T} \le q_{T+h|T}$ 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 $q_{T+h-1|T-1} = r_t(E_{T-1}) \ge b$, hence $s_{T+h-1|T-1} \le q_{T+h-1|T-1}$ 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,

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

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 4.3.

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. □

A.4 Proof of Proposition 4.4

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 4.3 applies and we can easily derive the desired result. □

A.5 Proof of Proposition 4.5

Proof. Let $q^*_{t+h|t} = q_{t+h|t} - \hat{q}_{t+h|t}$, then Equation 10 transforms into an update process $q^*_{t+h|t} = r_t\left(\sum_{i=h+1}^{t} (\text{err}_{i|i-h} - \alpha)\right)$, 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 4.3 can be directly applied to establish the long-run coverage achieved by the AcMCP method. □