ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

Online conformal inference for multi-step time series forecasting

Xiaoqian Wang, Rob J Hyndman

September 2024

Working Paper no/yr

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

10 September 2024

JEL classification: C10,C14,C32

Online conformal inference for multi-step time series forecasting

Abstract
A brief summary

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

1 Introduction

Motivation: Why conformal prediction?

- Machine Learning, non-parametric models.

Background: Development and popularity of conformal prediction

Application to time series data

Challenges and limitations

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 obviously allowed to vary over time in time series context. 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 observations 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.

Contribution: Overview of proposed methods

Some notes such as we use split conformal prediction throughout this paper, and uniform notation.

Paper structure.

2 Related work

2.1 Conformal prediction for regression

In this section, we focus on the regression setting, which stands as one of the primary areas where conformal prediction methods have seen substantial development. Suppose we have n data points $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$. The aim of conformal prediction is to construct a prediction interval

$\hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1})$ ensuring that the unseen response Y_{n+1} falls within $\hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1})$ at least $100(1-\alpha)\%$ of the time.

2.1.1 Split conformal prediction

Split conformal prediction (SCP, also called inductive conformal prediction, Papadopoulos et al. (2002); Vovk, Gammerman & Shafer (2005); Lei et al. (2018)), is a holdout method for building prediction intervals using a pre-trained model.

In regression setting, SCP randomly separates the available n data points into a *proper training set* $\mathscr{D}_{\text{tr}}$ of size n_t and a *calibration set* $\mathscr{D}_{\text{cal}}$ of size n_c. Given a regression model $\hat{\mu}: \mathbb{R}^d \to \mathbb{R}$ that is fitted on the training set and a score function $\mathscr{S}$, a *nonconformity score* $s_i = \mathscr{S}(X_i, Y_i)$, $i \in \mathscr{D}_{\text{cal}}$, is computed on every data point in $\mathscr{D}_{\text{cal}}$ to measure the nonconformity between the calibration's response values and the predicted values obtained from the fitted model $\hat{\mu}$. One commonly used nonconformity score function $\mathscr{S}$ in regression is the absolute residual, i.e. $s_i = |Y_i - \hat{\mu}(X_i)|$. Then SCP computes the prediction interval for the test data Y_{n+1} using

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

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

Theorem 2.1 (SCP, Vovk, Gammerman & Shafer (2005); Lei et al. (2018))**.** *Assume that the data points (X_i, Y_i), $i = 1, \ldots, n+1$, are i.i.d. (or more generally, exchangeable) from any distribution. For any score function $\mathscr{S}$, and any $\alpha \in (0,1)$, the split conformal prediction interval defined in Equation 1 satisfies*

$$\mathbb{P}\left\{ Y_{n+1} \in \hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1}) \right\} \geq 1 - \alpha.$$

Moreover, if we assume additionally that the nonconformity scores on $\mathscr{D}_{cal}$ are distinct with probability one, then we also have

$$\mathbb{P}\left\{ Y_{n+1} \in \hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1}) \right\} < 1 - \alpha + \frac{1}{n_c+1}.$$

2.1.2 Nonexchangeable conformal prediction

Barber et al. (2023) propose **nonexchangeable conformal prediction** (NexCP) that generalizes the SCP method to allow for some sources of nonexchangeability. For split conformal, the NexCP method can be considered as simply using weighted quantiles to obtain robust inference. Note that NexCP assumes the weights are fixed and data-independent. The intuition 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.

Given weights $w_i \in [0,1]$, $i \in \mathscr{D}_{\text{cal}}$, the prediction interval of the NexCP method for the test data Y_{n+1} is given by

$$\hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1}) = \left\{ y \in \mathbb{R} : \mathscr{S}(X_{n+1}, y) \leq Q_{1-\alpha}\left(\sum_{i \in \mathscr{D}_{\text{cal}}} \tilde{w}_i \cdot \delta_{s_i} + \tilde{w}_{n+1} \cdot \delta_{+\infty} \right) \right\}, \tag{2}$$

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

$$\tilde{w}_i = \frac{w_i}{\sum_{i \in \mathscr{D}_{\text{cal}}} w_i + 1}, \text{ for } i \in \mathscr{D}_{\text{cal}} \quad \text{and} \quad \tilde{w}_{n+1} = \frac{1}{\sum_{i \in \mathscr{D}_{\text{cal}}} w_i + 1}.$$

By placing different prespecified weights on data points, NexCP is able to deal with data that are not exchangeable and provide robustness against distribution drift. Barber et al. (2023) suggests using an exponential weighting scheme for time series data, where the weights decrease exponentially as data points come from the further in the past.

Theorem 2.2 (NexCP, Barber et al. (2023)). *Let $\mathscr{S}(Z)$ denote a vector of nonconformity scores for data points in the calibration and test sets, and $\mathscr{S}(Z^i)$ denote a vector of nonconformity scores after swapping the test point with the ith data point in the calibration set.*

For any score function $\mathscr{S}$, and any $\alpha \in (0,1)$, the nonexchangeable split conformal prediction interval defined in Equation 2 satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1})\right\} \geq 1 - \alpha - \sum_{i \in \mathscr{D}_{cal}} \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(\mathscr{S}(Z), \mathscr{S}\left(Z^i\right)\right),$$

without the assumption of exchangeability of the data, where d_{TV} denotes the total variation distance between two distributions.

Moreover, if we assume additionally that the nonconformity scores on $\mathscr{D}_{cal}$ are distinct with probability one, then the probability also has the upper bound:

$$\mathbb{P}\left\{Y_{n+1} \in \hat{\mathscr{C}}^{\alpha}_{n+1}(X_{n+1})\right\} < 1 - \alpha + \tilde{w}_{n+1} + \sum_{i \in \mathscr{D}_{cal}} \tilde{w}_i \cdot \mathrm{d}_{\mathrm{TV}}\left(\mathscr{S}(Z), \mathscr{S}\left(Z^i\right)\right).$$

2.1.3 Adaptive conformal prediction

The **adaptive conformal prediction** (ACP) proposed by Gibbs & Candès (2021) accounts for nonexchangeability by updating the quantile level in an online manner. Specifically, it treats α as a tunable parameter and estimates it recursively based on the historical performance. ACP can be used to deal with arbitrary online distribution shifts.

Similar to SCP, the initial step involves a random split on the observed data, yielding a training set for fitting a regression model and a withheld calibration set. ACP assumes that there exists an optimal value α_t^* to achieve the desired miscoverage rate α at each time t. To deal with cases where the data generating distribution is shifting over time, ACP recursively estimates the parameter α_t^* on the test points, using the updating equation

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t), \tag{3}$$

rather than consistently using the target miscoverage rate α. Here, $\gamma > 0$ is a fixed step size parameter, α_1 is the initial estimate typically set as $\alpha_1 = \alpha$, and $\mathrm{err}_t = \mathbb{1}\left\{Y_t \notin \hat{\mathscr{C}}^{\alpha_t}_t(X_t)\right\}$, where $\hat{\mathscr{C}}^{\alpha_t}_t(X_t)$ denotes the prediction set obtained using the $1 - \alpha_t$ quantile for the nonconformity scores available up to and including time t.

This update process adapts the estimation of α_t^* based on the historical frequency of miscoverage in the prediction sets. Specifically, it adjusts upwards (or downwards) the estimate of α_t^* if the prediction sets have shown over-coverage (or under-coverage) of the actual outcomes.

Selecting the parameter γ is pivotal yet challenging. Gibbs & Candès (2021) suggests 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 the multivalid prediction algorithm in which the prediction set is established by selecting a threshold from a sequence of candidate thresholds. However, Gibbs & Candès (2024) have empirically shown that both previous methods fail to promptly adapt to the local changes. To address this limitation, Gibbs & Candès (2024) propose 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.

Theorem 2.3 (ACP, Gibbs & Candès (2021))**.** *Assume that, with probability one, $Q_{\tau,t}$ is continuous and nondecreasing so that $Q_{0,t} = -\infty$ and $Q_{1,t} = \infty$. Then for any $\alpha \in (0,1)$, $\gamma > 0$, and any $T \geq 1$ test points, the prediction sets given by ACP satisfy*

$$|\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t - \alpha| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{\gamma T}.$$

In particular, this means that the prediction intervals obtained by ACP yield long-run coverage, i.e. $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t = \alpha$.

Remark. Theorem 2.3 suggests that a larger value for γ generally results in less deviation from the target coverage. As there is no restriction on α_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.

2.1.4 Conformal PID control

Instead of iteratively updating the miscoverage rate α as in Gibbs & Candès (2021), Angelopoulos, Candès & Tibshirani (2024) draw inspiration from control theory and directly update the quantile estimate q_t in an online fashion to achieve long-run coverage. This method treats the system for producing prediction sets as a **proportional-integral-derivative** controller, later referred to as PID.

The iteration of the PID method is given by

$$q_{t+1} = \underbrace{q_t + \eta(\mathrm{err}_t - \alpha)}_{\mathrm{P}} + \underbrace{r_t\left(\sum_{i=1}^{t}(\mathrm{err}_i - \alpha)\right)}_{\mathrm{I}} + \underbrace{g_t'}_{\mathrm{D}}.$$

The PID method integrates three modules, namely quantile tracking (P control), error integration (I control), and scorecasting (D control).

The P control module updates the quantile iteratively with a constant learning rate $\eta > 0$. The underlying intuition is similar to that of ACP: it increases (or decreases) the quantile estimate if the prediction set at time t miscovered (or covered) the corresponding realization. ACP 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.

Theorem 2.4 (The P control, Angelopoulos, Candès & Tibshirani (2024))**.** *Assume that the nonconformity scores are bounded within $[-b, b]$, for $0 < b < \infty$. Then for any $\alpha \in (0,1)$, $\eta > 0$, and any $T \geq 1$ the P control iteration satisfies*

$$|\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t - \alpha| \leq \frac{b+\eta}{\eta T}.$$

In particular, this means that the prediction intervals obtained by the P control iteration yield long-run coverage, i.e. $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathrm{err}_t = \alpha$.

The I control involves the sum of historical coverage errors $\sum_{i=1}^{t}(\mathrm{err}_i - \alpha)$ in a saturation function r_t when updating the quantile estimate, further stabilizing the coverage.

Theorem 2.5 (The I control, Angelopoulos, Candès & Tibshirani (2024)). *Assume that the nonconformity scores are bounded within* $[-b, b]$*, for* $b > 0$*, and that the saturation function* r_t *satisfies*

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

for positive constants b, c *and an admissible function* g*. Then for any* $\alpha \in (0,1)$*, and any* $T \geq 1$ *the I control iteration satisfies*

$$|\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t - \alpha| \leq \frac{c \cdot g(T) + 1}{T}.$$

In particular, this means that the prediction intervals obtained by the I control iteration yield long-run coverage, i.e. $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathrm{err}_t = \alpha$.

Finally, the D control g'_t is the forecast of q_{t+1} produced by a scorecaster (forecasting model) fitted using the nonconformity scores available up to and including time t. Instead of reacting to the past miscoverage events, this module looks forward and identifies the leftover signal not captured by the regression model $\hat{\mu}$. However, it may introduce variability and result in wider prediction sets if the scorecaster is aggressive or if there is not much leftover signal in the nonconformity scores.

2.2 Conformal prediction for time series

- Brief literature review of conformal prediction methods or applications on time series data.
- Multi-step conformal prediction.

3 Multi-step conformal prediction for time series

We now consider multi-step time series forecasting problems. In the following sections, we first introduce the setup for time series forecasting problems, and then generalize the existing conformal prediction methods described in Section 2.1 to deal with multi-step time series forecasting. Finally, we explore the properties of forecast errors for optimal multi-step forecasts, and then propose the **autocorrelated multi-step conformal prediction** (AcMCP) method to account for the serial correlation of multi-step forecast errors.

3.1 Setup

Let $z_t = (\boldsymbol{x}_t, y_t)$ denote the data point (including the response and possibly predictors) 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. 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 prediction interval produced using the information up to time t.

Sequential split. In time series context, it is infeasible to perform random splitting in split conformal due to the temporal dependency present in the data. Instead, throughout the conformal prediction methods in this section, 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 $t - t_r \gg h$.

Online learning. Here we consider a generic online learning framework to adapt to all conformal prediction methods we will discuss in subsequent sections. The framework adopts a standard rolling window evaluation strategy. Let the length of the training set be denoted as t_r and the calibration set length as t_c. The online learning framework consists of the following steps.

1. Initialization. Train a forecasting model on the initial proper training set $z_{t-t_r+1:t}$, setting $t = t_r$. Then generate H-step point forecasts $\{\hat{y}_{t_r+h|t_r}\}_{h\in[H]}$ and compute the corresponding nonconformity scores $\{s_{t_r+h|t_r} = \mathscr{S}(z_{1:t_r}, y_{t_r+h})\}_{h\in[H]}$ based on the true values H time steps ahead, i.e. $\{y_{t_r+h}\}_{h\in[H]}$.
2. Recurring procedure. Roll the training set forward by one data point by setting $t \to t+1$. Then repeat the 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 prediction intervals on the test set.
4. Online updating. Recursively roll the training set and calibration set forward by one data point to update the nonconformity scores for calibration, and then repeat the step 3 until prediction intervals for the entire test set are obtained, i.e., $\{\hat{\mathscr{C}}^{\alpha}_{t+h|t}\}_{t_r+t_c\le t\le T-h}$ for $h \in [H]$, where $T - t_r - t_c$ is the length of the test set. Therefore, our goal is to achieve long-run coverage in time.

For simplicity, so far we have only presented the nonconformity score defined as the (signed) forecast error

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

which is the most commonly used accuracy measure in the context of time series forecasting. We also note that the online learning setting can also be easily adjusted to work with expanding windows for the training and calibration sets.

Remark. With sequential splitting, multiple H-step forecasts and their respective nonconformity scores can be computed on the calibration set. These nonconformity scores have diverse forecast horizons, ranging from 1 to H, i.e., the number of periods between the forecast origin and the time at which nonconformity scores are evaluated. Thus, we can not uniformly treat these nonconformity scores and generate H-step prediction intervals of identical width.

3.2 Related methods extensions to multi-step forecasting

One of the key properties of optimal forecast errors is that the variance of the forecast error $e_{t+h|t}$ is non-decreasing in h (Diebold & Lopez 1996; Patton & Timmermann 2007). Therefore, instead of uniformly treating H-step nonconformity scores and generating H-step prediction intervals of identical width, we consider a setting wherein a separate conformal prediction procedure is applied for each $h \in [H]$ in an online manner.

3.2.1 Online multi-step split conformal prediction

We introduce online **multi-step split conformal prediction** (MSCP) as a generalization of SCP to recursively update all H-step prediction intervals over time. Specifically, for each $h \in [H]$, we consider the following simple online update to construct prediction intervals on the test set:

$$\hat{\mathscr{C}}^{\alpha}_{t+h|t} = \left\{ y \in \mathbb{R} : s^{y}_{t+h|t} \leq Q_{1-\alpha}\left(\sum_{i=t-t_c+1}^{t} \frac{1}{t_c+1} \cdot \delta_{s_{i|i-h}} + \frac{1}{t_c+1} \cdot \delta_{+\infty} \right) \right\}, \tag{5}$$

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

- Quantile estimation.
- Theorem?

3.2.2 Online multi-step weighted conformal prediction

The online **multi-step weighted conformal prediction** (MWCP) method adapts the NexCP method to the online setting. 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 order, akin to the rationale behind the exponential smoothing method in time series forecasting. Then for each $h \in [H]$, MWCP consider the online update for 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

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

- Weighted quantile estimation.
- Theorem?

3.2.3 Multi-step adaptive conformal prediction

In the online learning framework outlined in Section 3.1, we extend the ACP method to address multi-step time series forecasting, introducing the **multi-step adaptive conformal prediction** (MACP) method. Specifically, for each $h \in [H]$, we iteratively estimate α^* (treated as a tunable parameter) using the update equation

$$\alpha_{t+h|t} := \alpha_{t+h-1|t-1} + \gamma\left(\alpha - \mathrm{err}_{t|t-h}\right), \tag{6}$$

and compute the h-step-ahead prediction interval using Equation 5 by setting $\alpha = \alpha_{t+h|t}$. Here, $\gamma > 0$ denotes a fixed step size parameter, $\alpha_{h+1|1}$ denotes the initial estimate typically set to α, and $\mathrm{err}_{t|t-h}$ denotes the miscoverage event $\mathrm{err}_{t|t-h} = \mathbb{1}\left\{ y_t \notin \hat{\mathscr{C}}^{\alpha_{t|t-h}}_{t|t-h} \right\}$.

Equation 5 indicates that the correction to the estimation of α at time $t+h$ is determined by the historical miscoverage frequency up to time t. At each iteration, we raise the estimate of α used for quantile estimation at time $t+h$ if $\hat{\mathscr{C}}_{t|t-h}^{\alpha_{t|t-h}}$ covered y_t, whereas we lower the estimate if $\hat{\mathscr{C}}_{t|t-h}^{\alpha_{t|t-h}}$ miscovered y_t. Thus the miscoverage event has a delayed impact on the estimation of α over h periods, indicating that the correction of the α estimate becomes less prompt with increasing values of h. Particularly, Equation 6 reduces to the update for ACP as given by Equation 3 for $h=1$.

It should be noted that the update equation $\alpha_{t+1|t-h+1} := \alpha_{t|t-h} + \gamma\left(\alpha - \mathrm{err}_{t|t-h}\right)$ is not utilized in this context. This arises from its limitation that the available information at time t is insufficient to derive the α estimate for forecasting h steps ahead.

- Theorem?

3.2.4 Multi-step conformal PID control

We propose **multi-step conformal PID control** method (referred to as MPID hereafter), a generalization of the PID method to deal with multi-step forecasting.

For each individual forecast horizon $h \in [H]$, the iteration of the h-step-ahead quantile estimate is given by

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

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

$$x \geq c \cdot g(t-h) \Longrightarrow r_t(x) \geq b, \quad \text{and} \quad x \leq -c \cdot g(t-h) \Longrightarrow r_t(x) \leq -b, \tag{7}$$

for constant $b, c > 0$, and an admissible function g that is sublinear, nonnegative, nondecreasing. Here, we define $\hat{s}_{t+h|t}$ as the h-step-ahead forecast of the nonconformity score (defined as the forecast error), produced by any suitable scorecaster (forecasting model) trained using the h-step-ahead nonconformity scores available up to and including time t. With this updating equation, we can obtain all required h-step-ahead prediction intervals using information available up to time t.

The P control in MPID shows a delayed correction of the α estimate for a length of h periods. 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. Moreover, the D control performs h-step-ahead forecasting, which tends to result in increased forecast variance for larger forecast horizon h.

- Theorem?

3.3 Autocorrelated multi-step conformal prediction

In the PID method proposed by Angelopoulos, Candès & Tibshirani (2024), a notable feature is the inclusion of a scorecaster, a model trained on the score sequence, to forecast the next score. The rationale behind it is to residualize out any leftover signal in the score distribution not captured by the base forecasting model, such as trend and seasonality. However, in the context of time series forecasting, good forecasts are essential for making good decisions. We naturally expect to use a good forecasting model and ensure there is no useful signal in forecast errors (defined as nonconformity scores in this paper). If the forecasts are not optimal, the forecasting model should be improved to enhance its performance. Hence, we typically assume the use of a good forecasting model, and therefore, relying on another model to predict forecast errors to capture leftover information is not a

commonly used solution. Moreover, the inclusion of a scorecaster often only introduces variance to the quantile estimate, resulting in wider prediction intervals.

On the other hand, in our general setup outlined in Section 1 and Section 3.1, the DGP of a time series may depend on its own past, along with other exogenous predictors and their histories. Consequently, the h-step-ahead forecast errors $e_{t+h|t}$ may depend on the forecast errors from the past $h-1$ steps, i.e. $e_{t+1|t}, \ldots, e_{t+h-1|t}$, and forecast errors may 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 considers the autocorrelations of multi-step forecast errors.

3.3.1 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:

$$y_t = f_t\left(y_{(t-d):(t-1)}, \boldsymbol{x}_{(t-k):t}\right) + \epsilon_t, \tag{8}$$

where f_t is considered a nonlinear function in d lagged values of y_t (i.e. $y_{(t-d):(t-1)}$) and the current value along with the preceding k values of the exogenous predictors (i.e. $\boldsymbol{x}_{(t-k):t}$), and ϵ_t is white noise.

It is well-established in the forecasting literature that, for optimal h-step-ahead forecasts, the sequence of forecast errors is *at most* an MA($h-1$) process (Harvey, Leybourne & Newbold 1997; Diebold 2017). We now present the property under the assumption of a non-stationary autoregressive DGP, and provide its proof in Section B.1 based on the proof of Proposition 3.2 that we will introduce later.

Proposition 3.1 (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 8. Assume that the exogenous predictors are known into the future if applicable. The forecast errors of optimal h-step-ahead forecasts follow an approximate MA($h-1$) process:*

$$e_{t+h|t} = c + \epsilon_{t+h} + \theta_1 \epsilon_{t+h-1} + \cdots + \theta_{h-1}\epsilon_{t+1},$$

where $c = 0$, motivated by the property that optimal forecasts are unbiased.

We proceed by exploring the autocorrelations of multi-step forecast errors for optimal forecasts.

Proposition 3.2 (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 8. Assume that the exogenous predictors are known into the future if applicable. The forecast errors of optimal h-step-ahead forecasts are <u>at most</u> an approximate AR($h-1$) process given by:*

$$e_{t+h|t} = c + \epsilon_{t+h} + \phi_1 e_{t+h-1|t} + \cdots + \phi_{h-1} e_{t+1|t},$$

where $e_{t+h|t}$ is the h-step-ahead forecast error with variance non-decreasing in h, and the intercept $c = 0$, given the property that optimal forecasts are unbiased.

Proof. Let $\hat{y}_{t+h|t}$ be the optimal h-step-ahead point forecast, and $e_{t+h|t}$ be the h-step-ahead forecast error. Denote $\boldsymbol{u}_{t+h} = \boldsymbol{x}_{(t-k+h):(t+h)}$. Then we have

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

For $h = 1$, we simply have $e_{t+1|t} = \epsilon_{t+1}$.

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

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

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

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

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

Therefore, the forecast errors of optimal h-step-ahead forecasts are at most an approximate AR($h-1$) process. □

Proposition 3.2 can be viewed as an extension of Proposition 3.1. It suggests that the 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, we note that the autocorrelation among errors associated with optimal forecasts can not be used to improve forecasting performance, as it does not incorporate any information available when the forecast was made. It is reasonable because if we could forecast the forecast error, we could improve the forecast, indicating that the initial forecast couldn't have been optimal.

The proof of Proposition 3.2 suggests that, if f_t is a linear autoregressive model, then the AR coefficients are the linear coefficients of the optimal forecasting model. However, when f_t takes on a more complex nonlinear structure, the AR coefficients become complicated functions of observed data and unobserved model coefficients.

3.3.2 The AcMCP method

Inspired by the properties of multi-step forecast errors discussed in Section 3.3.1, we now propose the **autocorrelated multi-step conformal prediction** (AcMCP) method. Unlike extensions of existing conformal prediction methods that treat multi-step forecasting as independent events (see Section 3.2), the AcMCP method integrates the autocorrelations inherent in multi-step forecast errors, thereby making the output multi-step prediction intervals more logically structured.

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

$$q_{t+h|t} = q_{t+h-1|t-1} + \eta\left(\mathrm{err}_{t|t-h} - \alpha\right) + r_t\left(\sum_{i=h+1}^{t}\left(\mathrm{err}_{i|i-h} - \alpha\right)\right) + \tilde{e}_{t+h|t}, \tag{9}$$

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 forward. Instead, we can make multi-step forecasting, constructing distribution-free prediction intervals for steps $t+1, \ldots, t+h$ based on available information up to time t. This is highly important in the field of time series forecasting.

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 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 an AR$(h-1)$ model (with respect to h instead of t) 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 in this paper), but rather to incorporate autocorrelations present in multi-step forecast errors within the resulting multi-step prediction intervals. As previously explained, in the context of time series forecasting, we typically assume the use of a good base forecasting model, making it unnecessary to train an additional model to predict forecast errors in order to capture leftover information. If the forecasts are not optimal, the base forecasting model should be improved to enhance its performance.

3.3.3 Coverage guarantees

Proposition 3.3. *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 7, for an admissible function g. Then the iteration $q_{t+h|t} = r_t\left(\sum_{i=h+1}^{t}\left(\mathrm{err}_{i|i-h} - \alpha\right)\right)$ satisfies*

$$\left|\frac{1}{T-h}\sum_{t=h+1}^{T}\left(\mathrm{err}_{t|t-h} - \alpha\right)\right| \leq \frac{c \cdot g(T-h) + h}{T-h}, \tag{10}$$

for any $T \geq h+1$, where $c > 0$ is the constant in Equation 7.

In particular, this means the prediction intervals obtained by the iteration yield 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. Let $E_T = \sum_{t=h+1}^{T}\left(\mathrm{err}_{t|t-h} - \alpha\right)$. The inequality given by Equation 10 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 $T = h+1, \ldots, 2h$, $E_T = \sum_{t=h+1}^{T}(\mathrm{err}_{t|t-h} - \alpha) \leq (T-h) - (T-h)\alpha \leq T-h \leq h \leq cg(T-h) + h$ as $c > 0$, $h \geq 1$, g is nonnegative, and $\mathrm{err}_{t|t-h} \leq 1$. Thus, Equation 10 holds for $T = h+1, \ldots, 2h$.

Now, assuming Equation 10 is true up to T. We partition the argument into $h+1$ cases:

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

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 7. 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 $s_{T+h|T} \le q_{T+h|T}$ 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 $\text{err}_{T+1|T-h+1} \le 1$, this means that

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

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 $E_T \le cg(T-h)$, and simply using $\text{err}_{T+h-i|T-i} \le 1$ for $i = 0, \dots, h-1$, we have

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

Therefore, we can deduce the desired outcome at any $T \ge h+1$. This completes the proof for the first part of Proposition 3.3.

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. □

The quantile iteration $q_{t+h|t} = q_{t+h-1|t-1} + \eta\left(\text{err}_{t|t-h} - \alpha\right)$ can be seen as a particular instance of the iteration outlined in Proposition 3.3 if we set $q_{2h|h} = 0$ without losing generality. Thus, its coverage bounds can be easily derived as a result of Proposition 3.3.

Proposition 3.4. *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 $q_{t+h|t} = q_{t+h-1|t-1} + \eta\left(\text{err}_{t|t-h} - \alpha\right)$ satisfies*

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

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

In particular, this means the prediction intervals obtained by the iteration yield 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. We set $q_{2h|h}=0$ without losing generality, the iteration $q_{t+h|t}=q_{t+h-1|t-1}+\eta\left(\mathrm{err}_{t|t-h}-\alpha\right)$ simplifies to $q_{t+h|t}=\eta\sum_{i=h+1}^{t}\left(\mathrm{err}_{i|i-h}-\alpha\right)$. Let $r_t(x)=\eta x$ and the admissible function $g(t-h)=b$, Equation 7 holds for $c=\frac{1}{\eta}$. Then Proposition 3.3 applies and we can easily derive the desired result. □

More importantly, Proposition 3.3 is also adequate for establishing the coverage guarantee of the proposed AcMCP method given by Equation 9. Following the idea of Angelopoulos, Candès & Tibshirani (2024), we first reformulate Equation 9 as

$$q_{t+h|t}=\hat{q}_{t+h|t}+r_t\left(\sum_{i=h+1}^{t}\left(\mathrm{err}_{i|i-h}-\alpha\right)\right), \tag{11}$$

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 $\hat{q}_{t+h|t}=q_{t+h-1|t-1}+\eta\left(\mathrm{err}_{t|t-h}-\alpha\right)+\tilde{e}_{t+h|t}$ will recover Equation 9. 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 11.

Proposition 3.5. *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 7, for an admissible function g. Then the prediction intervals obtained by the AcMCP iteration given by Equation 11 yield 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. Let $q^*_{t+h|t}=q_{t+h|t}-\hat{q}_{t+h|t}$, then Equation 11 transforms into an update process $q^*_{t+h|t}=r_t\left(\sum_{i=h+1}^{t}\left(\mathrm{err}_{i|i-h}-\alpha\right)\right)$, 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 3.4 can be directly applied to establish the long-run coverage achieved by the AcMCP method. □

- Remark on propositions.
- Choosing the learning rate? Analysis of the parameter effect.

4 Experiments

In this section, we examine the empirical performance of the previously 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 Barber et al. (2023); following Angelopoulos, Candès & Tibshirani (2024), 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 $\hat{B}_t=\max\{s_{t-\Delta+1|t-\Delta-h+1},\cdots,s_{t|t-h}\}$ 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 consider a clipped version of MACP by imputing infinite intervals with the largest score seen so far.

4.1 Simulated examples

4.1.1 Linear autoregressive process

We first consider a simulated stationary time series which is generated from an AR(2) process

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

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. $H = 3$) with the automated algorithm implemented in the **forecast** R package (Hyndman et al. 2023). 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 1 presents the rolling coverage and interval width of each method for each forecast horizon, with metrics computed over a rolling window of size 500. In terms of coverage, 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 large troughs and peaks in coverage as time changes. Turning to the prediction interval width, 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 fact that the inclusion of a second model (scorecaster) is likely to introduce large variance into the generated prediction intervals. The results can be further elucidated with Figure 2, which presents boxplots of rolling coverage and interval width for each method and each forecast horizon.

We note that 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 interval over a truncated period of length 500, as shown in Figure 3. 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 may 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 3 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.

4.1.2 Nonlinear autoregressive process

We then consider the case of a nonlinear data generation process, which happens in many practical time series applications. We specify the true data generation process as

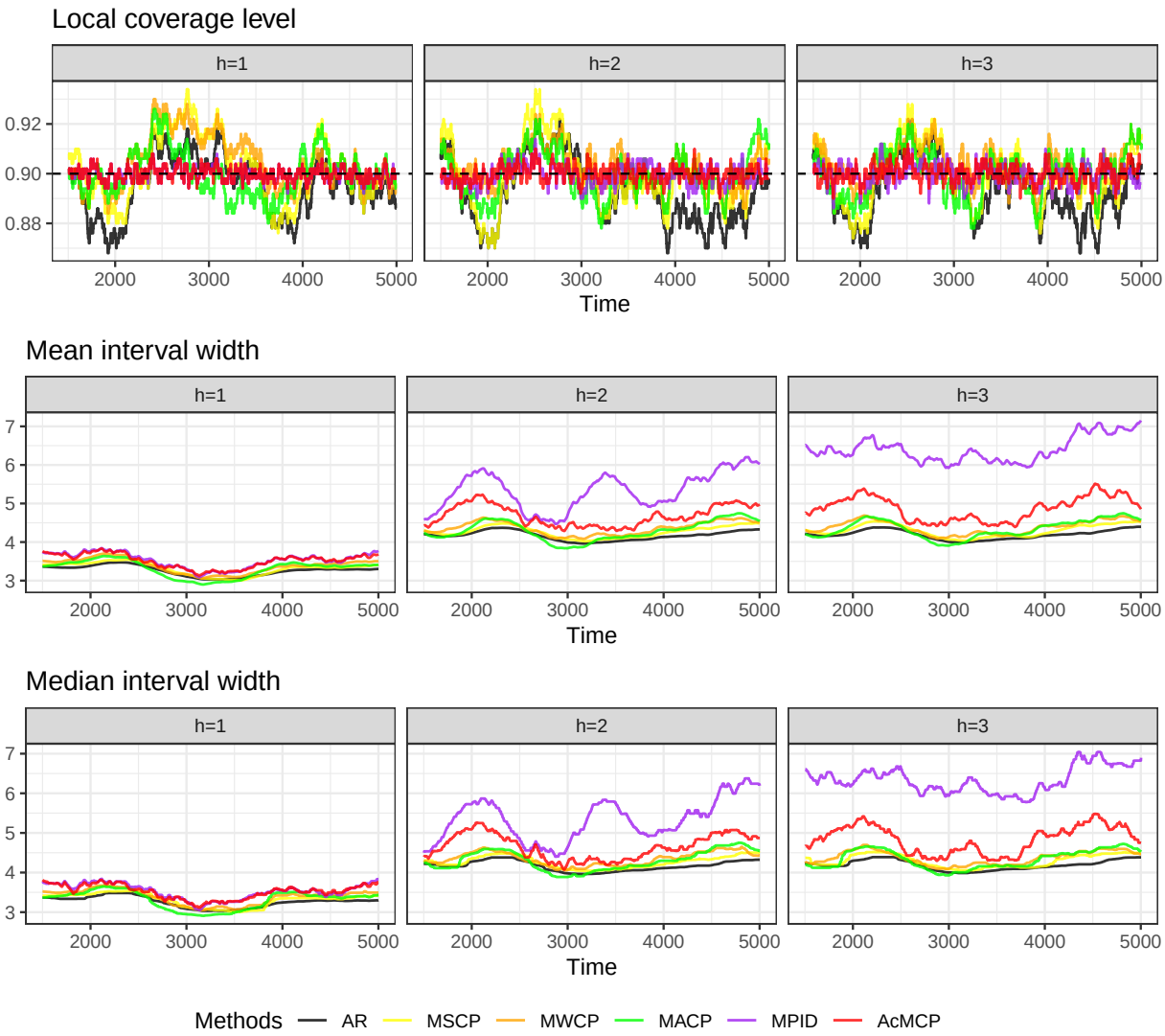

Figure 1: *AR(2) simulation results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 500*. The black dashed line indicates the target level of* $1-\alpha=0.9$*.*

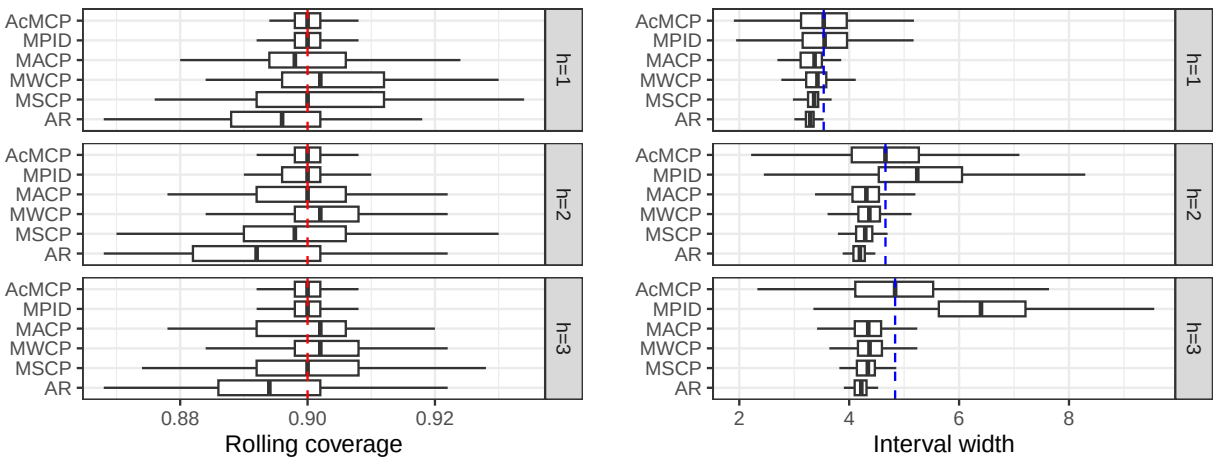

Figure 2: *AR(2) simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

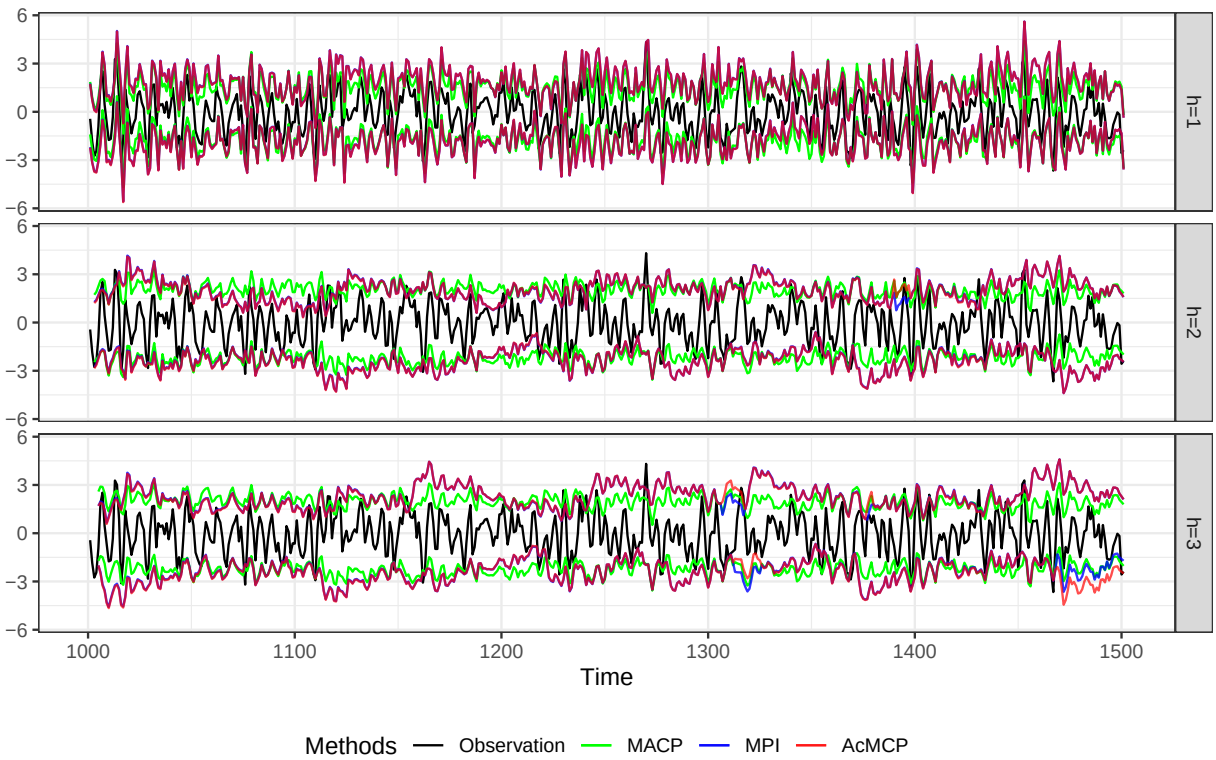

Figure 3: *AR(2) simulation results showing the prediction interval bounds for the MACP, MPI, and AcMCP methods over a truncated period of length 500.*

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

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 data generation process, 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. $H = 3$) with the automated algorithm implemented in the **forecast** R package. Note that it is not straightforward for neural networks to derive interval forecasts, thus we do not include neural network models when presenting the results.

Figure 4 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 fluctionations 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. Lastly, 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.

The most basic requirement for a prediction interval is to cover the actual value with the desired

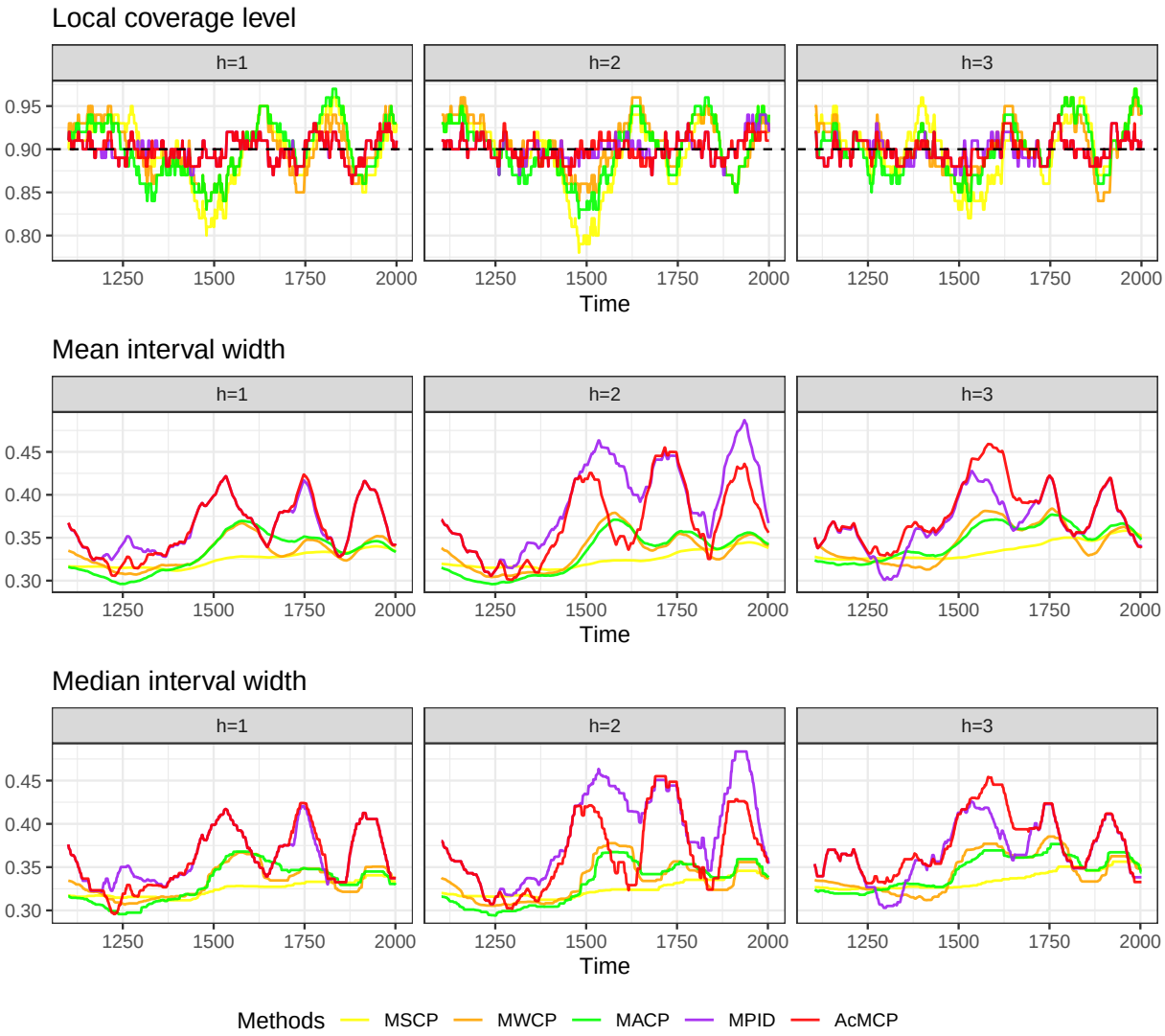

Figure 4: *Nonlinear simulation results showing rolling coverage, mean and median interval width for each forecast horizon. The displayed curves are smoothed over a rolling window of size* 100. *The black dashed line indicates the target level of* $1-\alpha = 0.9$.

probability. Evaluating the efficiency of forecast intervals is relevant only when they demonstrate valid coverage. As discussed, MSCP, MWCP and MACP have coverage that deviate a bit far from the desired coverage, thus we now focus our comparison on MPID and AcMCP, the two methods that offer more accurate coverage. Table 1 shows the resulting coverage gap (the difference between actual and nominal coverage), forecast interval width, and Winkler score (Winkler 1972), averaged over all test sets, for the MPI, MPID, and AcMCP methods. The Winkler score is proposed by Winkler (1972) to evaluate a prediction interval, and is defined as the length of the interval plus a penalty if the observation is outside the interval. A smaller Winkler score indicates better performance. The results indicate that, in this nonlinear data generation process setting, there is essentially no difference among these three methods in terms of overall coverage gap, average interval width, and Winkler score.

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 5. We see 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.

Table 1: *Nonlinear simulation results showing coverage gap, interval width, and Winkler score, averaged over all test sets.*

	h=1			h=2			h=3		
Methods	Coverage gap	Mean width	Winkler score	Coverage gap	Mean width	Winkler score	Coverage gap	Mean width	Winkler score
MPI	0	0.361	0.436	0.003	0.373	0.468	0.001	0.373	0.459
MPID	0	0.362	0.436	0.003	0.388	0.476	0.001	0.367	0.460
AcMCP	0	0.361	0.435	0.002	0.369	0.467	0.001	0.376	0.459

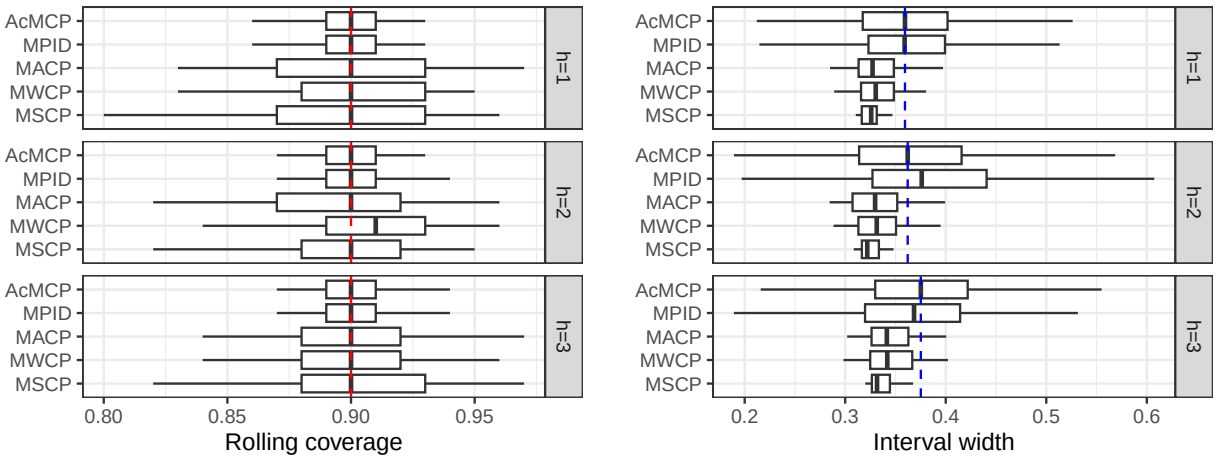

Figure 5: *Nonlinear simulation results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

4.2 Real data examples

4.2.1 Electricity demand data

Now we examine empirical performance of the conformal prediction methods using an electricity demand data set. The data set tracks 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. Figure 6 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).

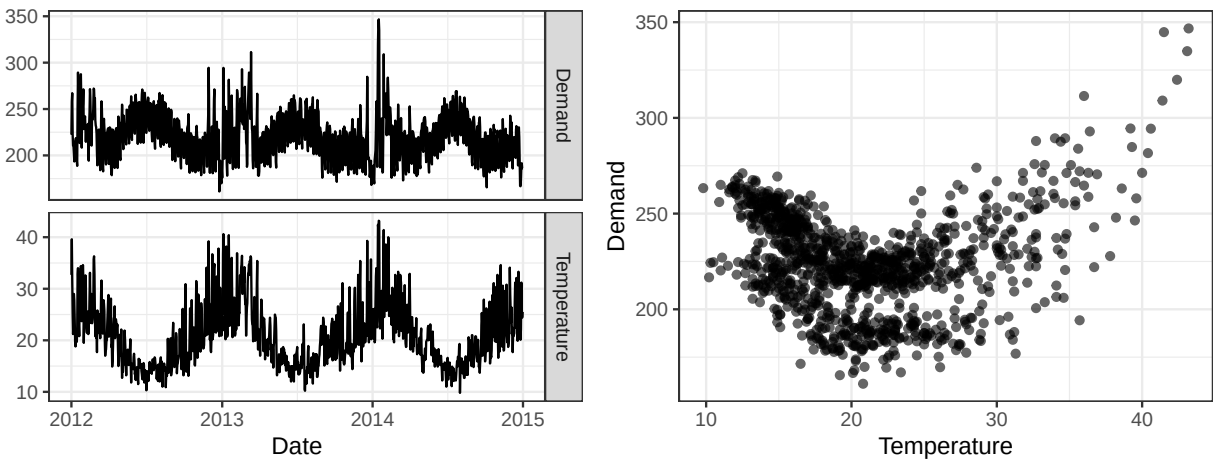

Figure 6: *Daily electricity demand and corresponding daily maximum temperatures in 2012–2014, Victoria, Australia.*

Our response variable is `Demand`, and we use two covariates: `Temperature`, and `Workday` which is an indicator variable for if the day was a working day or not. Following Hyndman & Athanasopoulos (2021), 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 is assumed to follow an ARIMA model to contain autocorrelation. Under the sequential split and online learning settings, we use two years of data as training sets to fit dynamic regression models, and use 100 data points for calibration sets.

We present the results in Figure 7 and Figure 8, comparing the rolling coverage and interval width of each method. These computations are based on a rolling window of size 100. First, we observe that DR (dynamic regression) consistently achieves a significantly higher coverage than the 90% target coverage, resulting in much wider intervals than other methods for $h = 1, 2, 3, 4$. Secondly, 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. Thirdly, 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.

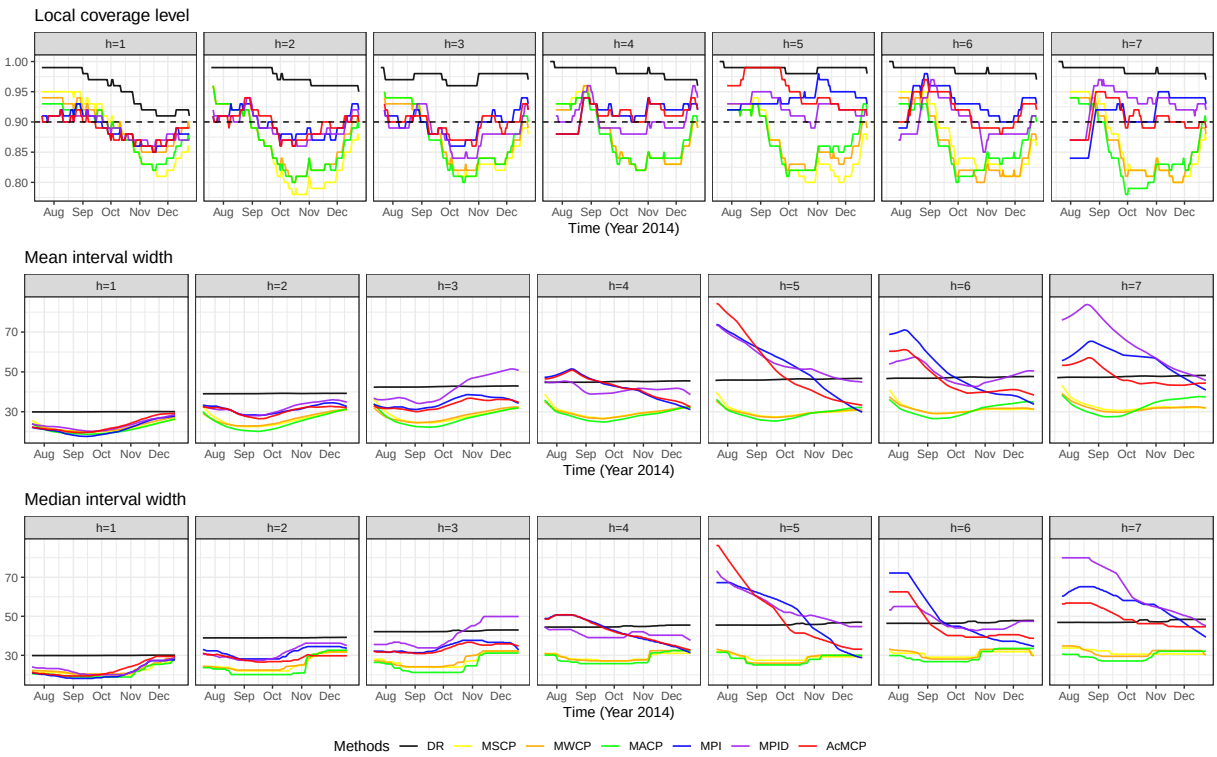

Figure 7: *Electricity demand data results showing rolling coverage, mean and median interval width for each forecast horizon. The dynamic regression model is abbreviated to DR. The displayed curves are smoothed over a rolling window of size* 100*. The black dashed line indicates the target level of* $1 - \alpha = 0.9$.

We present the forecast interval bounds for MACP, MPI, and AcMCP in Figure 9. The plot shows that MACP intervals are too narrow to adequately hug the true values, particularly from September to November of 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 of 2014, the upper bounds of MPI intervals struggle to

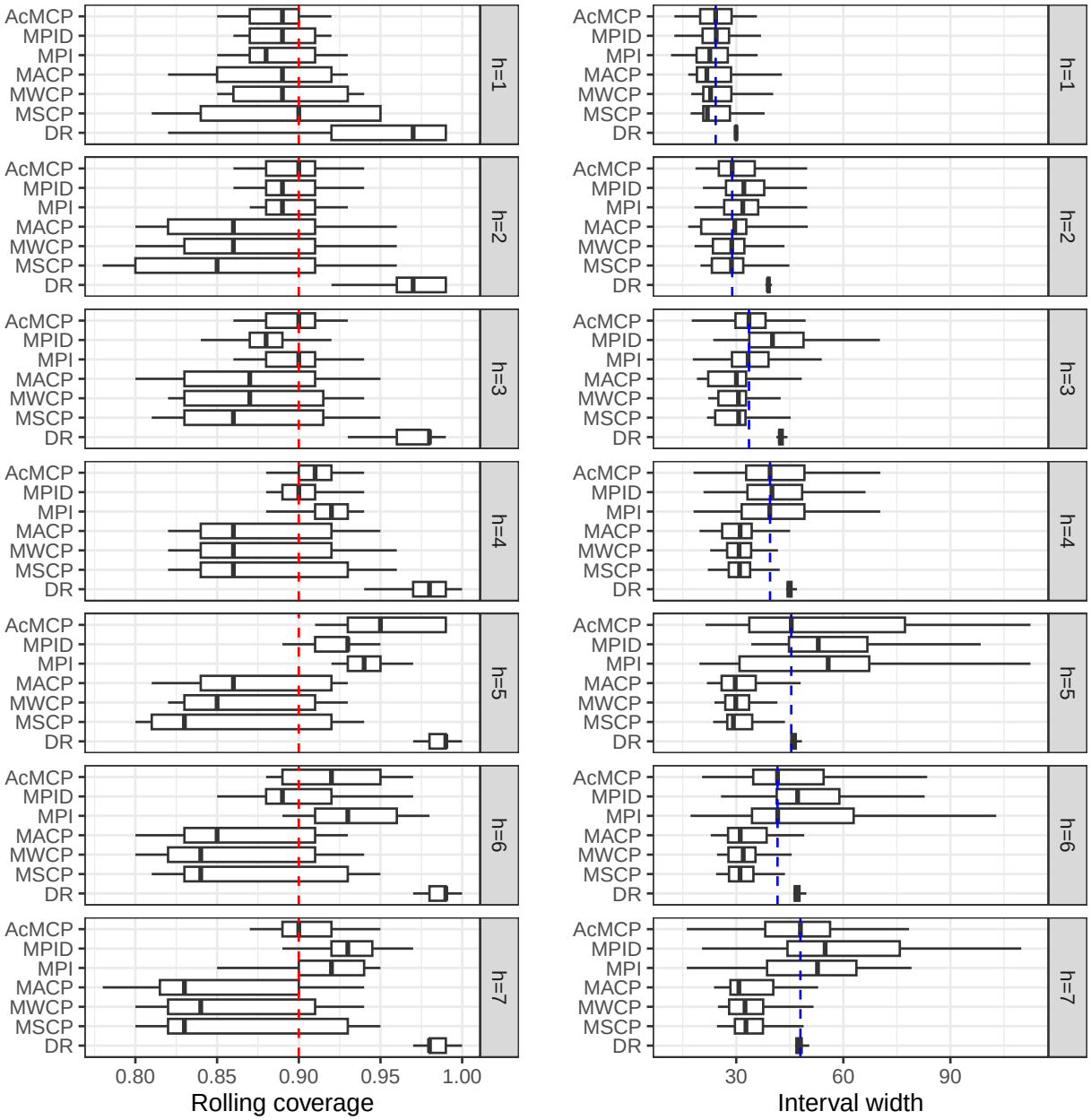

Figure 8: *Electricity demand data results showing boxplots of the rolling coverage and interval width for each method across different forecast horizons. The dynamic regression model is abbreviated to DR. The red dashed lines show the target coverage level, while the blue dashed lines indicate the median interval width of the AcMCP method.*

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 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.

Figure 10 presents Winkler scores of each method across different forecast horizons. DR and MPID consistently perform worse in terms of Winkler scores. For 1-, 2-, and 3-step ahead forecasting, MPI and AcMCP deliver comparable or even better results than MSCP, MWCP, and MACP, while still providing more precise coverage. For 4-, 5-, 6-, and 7-step ahead forecasting, MPI and AcMCP yield much larger Winkler scores than other conformal prediction methods, whereas the competing

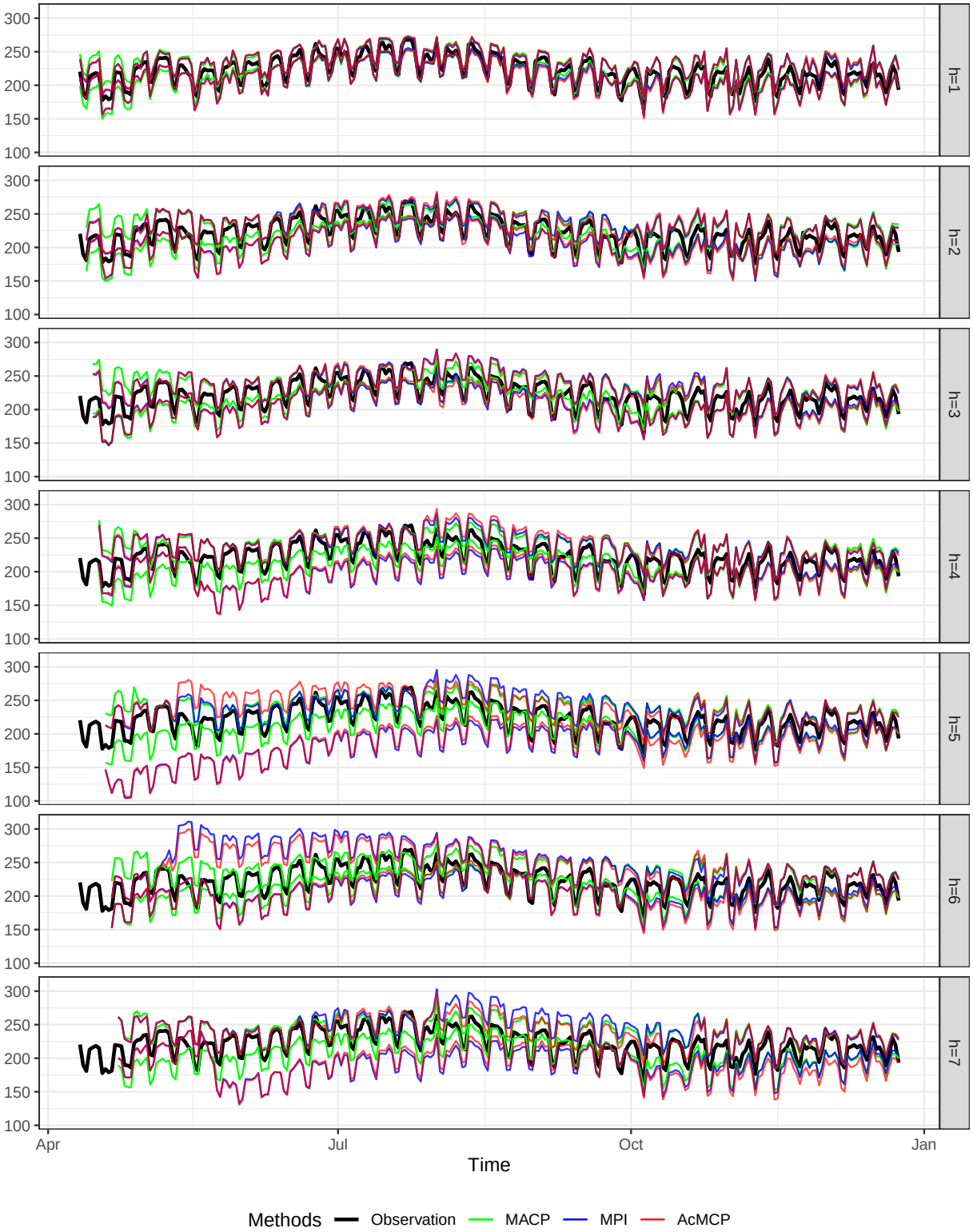

Figure 9: *Electricity demand data results showing the forecast interval bounds for MACP, MPI, and AcMCP over the whole test set.*

methods suffer from severe undercoverage issues. Thus, MPI and AcMCP are able to provide valid prediction intervals. We also notice that AcMCP outperforms MPI for $h = 6$ and $h = 7$, where both methods offer similar coverage.

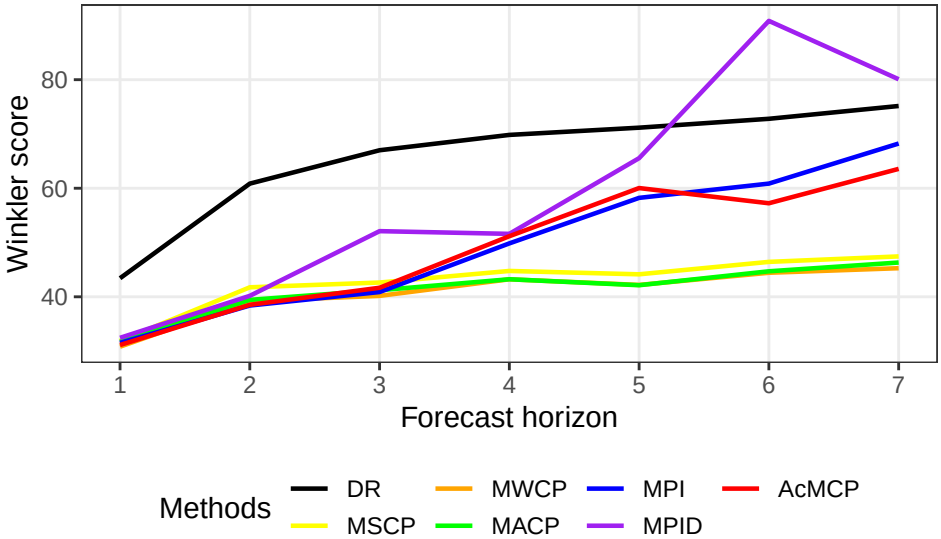

Figure 10: *Electricity demand data results showing Winkler scores of each method across different forecast horizons.*

4.2.2 Eating out expenditure data

Finally, we apply the conformal prediction methods to predict 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 anuual seasonal patterns, and variability proportional to the data level.

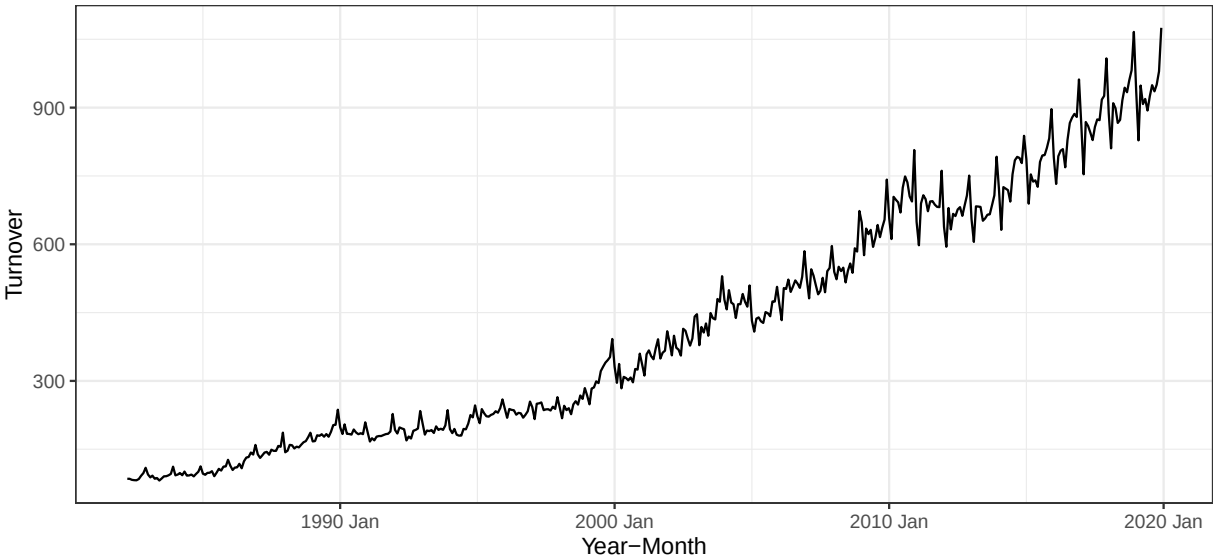

Figure 11: *Monthly expenditure on cafes, restaurants and takeaway food services in Victoria, Australia, from April 1982 to December 2019.*

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 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. Our goal

is to forecast 12 months ahead, i.e. $H = 12$. Under the sequential split and online learning settings, we use 20 years of data for training the models and 5 years of data points 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, interval width, and Winkler score 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 ($h > 5$). Secondly, for $h \leq 5$, MPI and AcMCP can approximately achieve the desired coverage and provide comparable mean interval widths and Winkler scores with other methods, except for MPID. Thirdly, 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 less deviation from the desired coverage across most forecast horizons.

5 Discussion

- Large forecast horizon
- Learning rate determination
- Variability introduced by the forecasts of predictors
- Efficiency of prediction intervals

References

Angelopoulos, AN, EJ Candès & RJ Tibshirani (2024). Conformal PID control for time series prediction. *Advances in Neural Information Processing Systems* **36**.

Barber, RF, EJ Candès, A Ramdas & RJ Tibshirani (1, 2023). Conformal prediction beyond exchangeability. *The Annals of Statistics* **51**(2).

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

Diebold, FX (2017). *Forecasting*. Department of Economics, University of Pennsylvania. http://www.ssc.upenn.edu/~fdiebold/Textbooks.html.

Diebold, FX & JA Lopez (1996). "Forecast evaluation and combination". In: *Statistical Methods in Finance*. Vol. 14. Handbook of Statistics. Elsevier, pp.241–268.

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

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

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

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

Hyndman, RJ, G Athanasopoulos, C Bergmeir, G Caceres, L Chhay, M O'Hara-Wild, F Petropoulos, S Razbash, E Wang & F Yasmeen (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.22.0. https://pkg.robjhyndman.com/forecast/.

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

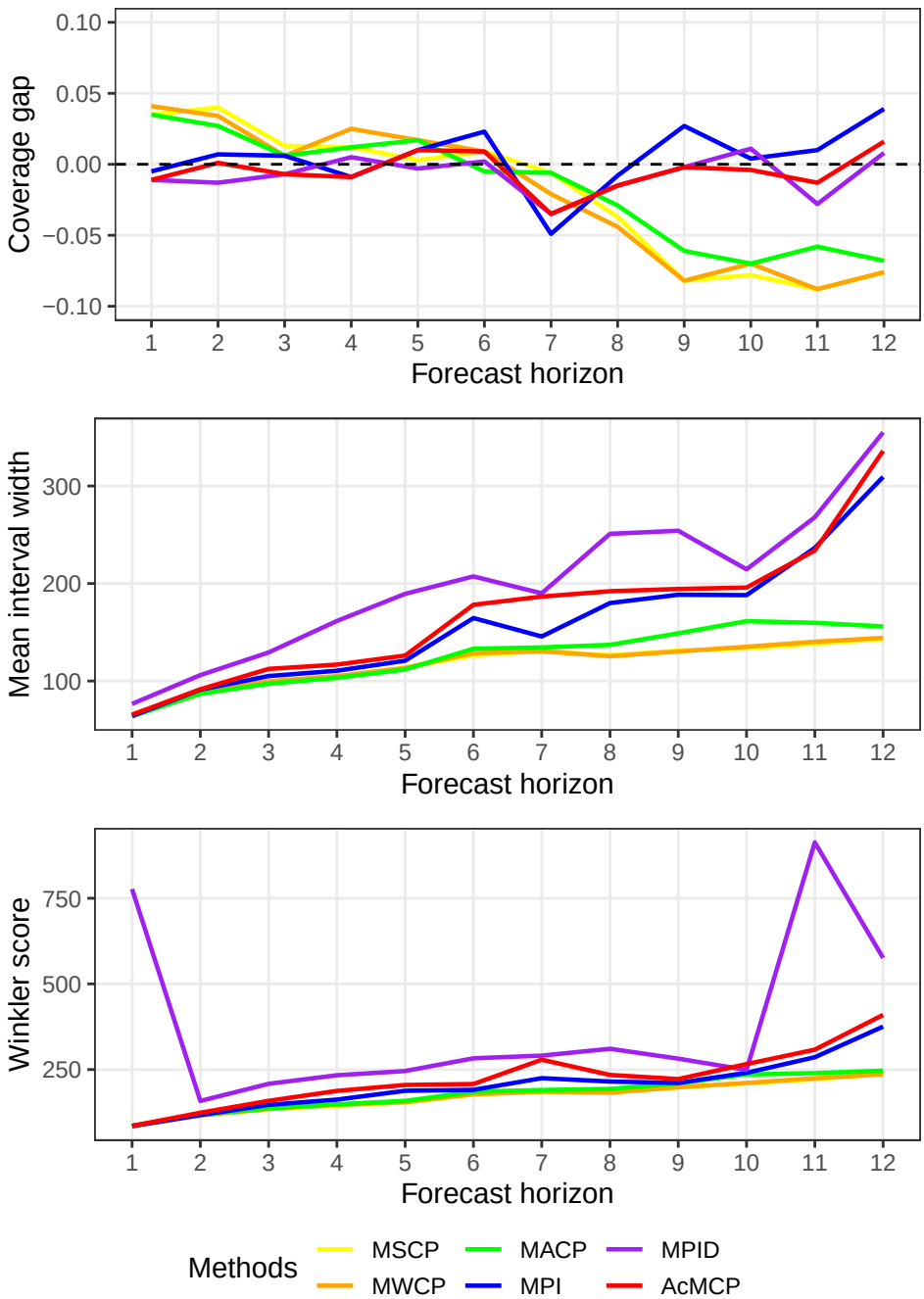

Figure 12: *Eating out expenditure data results showing coverage gap, interval width, and Winkler scores averaged over the entire test set for each forecast horizon. The black dashed line in the top panel indicates no difference from the 90% target level.*

Papadopoulos, H, K Proedrou, V Vovk & A Gammerman (2002). Inductive confidence machines for regression. In: *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*. Springer, pp.345–356.

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

Sommer, B (2023). "Forecasting and decision-making for empty container repositioning". PhD thesis.

Vovk, V, A Gammerman & G Shafer (2005). *Algorithmic Learning in a Random World*. Springer-Verlag. http://dx.doi.org/10.1007/b106715.

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

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

Appendix A Weighted quantile estimation

- median-unbiased estimates

Appendix B Proof

B.1 Proof of Proposition 3.1

Proof. Here, we give the proof of Proposition 3.1 based on Proposition 3.2, and the idea is motivated by Sommer (2023).

For 1-step-ahead optimal forecast, we have

$$y_{t+1} = f_{t+1}(y_{(t-d+1):t}, \boldsymbol{x}_{(t-k+1):(t+1)}) + \epsilon_{t+1},$$

so $e_{t+1|t} = \epsilon_{t+1}$.

Based on Proposition 3.2, we have

$$\begin{aligned}
e_{t+2|t} &= \epsilon_{t+2} + \phi_1^{(2)} e_{t+1|t} \\
e_{t+3|t} &= \epsilon_{t+3} + \phi_1^{(3)} e_{t+2|t} + \phi_2^{(3)} e_{t+1|t} \\
&\cdots \\
e_{t+d|t} &= \epsilon_{t+d} + \phi_1^{(d)} e_{t+d-1|t} + \cdots + \phi_{d-1}^{(d)} e_{t+1|t} \\
e_{t+d+1|t} &= \epsilon_{t+d+1} + \phi_1^{(d+1)} e_{t+d|t} + \cdots + \phi_{d-1}^{(d+1)} e_{t+2|t} + \phi_d^{(d+1)} e_{t+1|t} \\
&\cdots \\
e_{t+H|t} &= \epsilon_{t+H} + \phi_1^{(H)} e_{t+H-1|t} + \cdots + \phi_{d-1}^{(H)} e_{t+H-d+1|t} + \phi_d^{(H)} e_{t+H-d|t}, \quad H > d+1.
\end{aligned}$$

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

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

where θ_i is a complex function derived from the AR coefficients of all AR$(j-1)$ models, for $j = 1, 2, \ldots, h-1$. So we conclude that the h-step-ahead forecast error sequence $\{e_{t+h|t}\}$ follows an MA$(h-1)$ process. □