Sequential Predictive Conformal Inference for Time Series

Chen Xu [1] **Yao Xie** [1]

Abstract

We present a new distribution-free conformal prediction algorithm for sequential data (e.g., time series), called the *sequential predictive conformal inference* (SPCI). We specifically account for the nature that time series data are non-exchangeable, and thus many existing conformal prediction algorithms are not applicable. The main idea is to adaptively re-estimate the conditional quantile of non-conformity scores (e.g., prediction residuals), upon exploiting the temporal dependence among them. More precisely, we cast the problem of conformal prediction interval as predicting the quantile of a future residual, given a user-specified point prediction algorithm. Theoretically, we establish asymptotic valid conditional coverage upon extending consistency analyses in quantile regression. Using simulation and real-data experiments, we demonstrate a significant reduction in interval width of SPCI compared to other existing methods under the desired empirical coverage.

1. Introduction

Uncertainty quantification for prediction algorithms is essential for statistical and machine learning models. Sequential prediction or time-series prediction aims to predict the subsequent outcome based on past observations. Uncertainty quantification in the form of prediction intervals is of particular interest for high-stake domains such as finance, energy systems, healthcare, and so on (Harries et al., 1999; Díaz-González et al., 2012; Cochran et al., 2015). Classic approaches for prediction interval are typically based on strong parametric assumptions of time-series models such as autoregressive and moving average (ARMA) models (Brockwell et al., 1991), which impose strong distribution assumptions on the data-generating process. There need to be principled ways to perform uncertainty quantification for complex prediction models such as random forests (Breiman, 2001) and neural networks (Lathuilière et al., 2019).

Conformal prediction (CP) has become a popular distribution-free technique to perform uncertainty quantification for complex machine learning algorithms. However, conformal prediction for time series has been a challenging case because such data do not satisfy the exchangeability assumption in conformal inference, and thus we need to adjust existing or even develop new *sequential CP* algorithms with theoretical guarantees. The challenges also arise in real-world applications where time series data tend to have significant stochastic variations and strong correlations. These challenges are illustrated via a real-data example for solar energy prediction, as shown in Figure 1, where the prediction residuals (using random forest as prediction algorithm) are still highly correlated. Besides the temporal correlation in the prediction residuals (or conformity scores in general), we observe that a notable feature of sequential conformal prediction is that the prediction residuals can be obtained as "feedback" to the algorithm. For instance, for one-step ahead prediction, the prediction accuracy of the prediction algorithm is revealed immediately after one-time step. Thus, the recent prediction residuals reveal whether or not the predictive algorithm is performing well for that segment of data. Such feedback structure is illustrated in Figure 2, which highlights the conceptual difference between traditional conformal and sequential conformal prediction methods. We specifically exploited such feedback structure in designing the sequential conformal prediction algorithms.

More precisely, both the traditional conformal inference and the sequential conformal inference considered in this paper are general-purpose wrappers that can be used around any predictive model for any data and proceed by defining "non-conformity scores". However, there are also significant differences: Traditional conformal prediction assumes exchangeable training and test data to obtain performance guarantees, which leads to exchangeable non-conformity scores, and cannot receive feedback during prediction. In contrast, sequential CP observes non-exchangeable data sequences and leverages feedback during prediction.

In this work, we propose a *sequential predictive confor-*

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. Correspondence to: Yao Xie <yao.xie@isye.gatech.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

mal inference (SPCI) framework for time series with scalar outputs. The idea is to utilize the feedback structure of prediction residuals in the sequential prediction problem to obtain desired coverage. We specifically exploit the serial dependence across prediction residuals (conformity score) by performing quantile regression using past residuals for the future prediction intervals.; thus, the most recent past residuals contain information about the immediate future ones. Similar to most existing conformal prediction literature, we make no assumptions about the data-generating process or the quality of estimation by the point estimator. Our main contributions are

- The main novelty of SPCI is the time-adaptive re-estimation of residual quantiles over time, upon leveraging the temporal dependency among residuals. We use Random Forest for quantile regression here, but SPCI is applicable to other quantile regression methods.

- Theoretically, we obtain asymptotic conditional coverage of the constructed intervals for dependent data, based on prior results for random forest quantile regression. When data are exchangeable, we show that SPCI enjoys the same finite-sample and distribution-free marginal coverage guarantee as traditional conformal prediction methods.

- Experimentally, we demonstrate competitive and/or improved empirical performance against baseline CP methods on sequential data. In particular, SPCI can obtain significantly narrower intervals on real data without coverage loss. We further demonstrate the benefit of SPCI in multi-step predictive inference.

1.1. Literature review

Conformal prediction has been an increasingly popular framework for distribution-free uncertainty quantification. Initially proposed in (Shafer & Vovk, 2008), CP methods generally proceed as follows. First, one designs a type of "non-conformity score" based on the given point estimator $\hat{f}$, where the score measures how different a potential value of the response variable Y is to existing observations. A common choice for such scores in regression problems is the prediction residual. Second, one computes these scores on a *hold-out* set not used to train the estimator $\hat{f}$. Third, the prediction interval is defined as all potential values of Y whose non-conformity score is less than $1-\alpha$ fraction of these scores over the hold-out set. Many existing works such as (Papadopoulos et al., 2007; Gupta et al., 2021; Angelopoulos et al., 2021; Romano et al., 2020) utilize this idea for uncertainty quantification in regression or classification problems. Comprehensive surveys and tutorials can be found in (Fontana et al., 2023; Angelopoulos & Bates, 2021). CP framework are distribution-free and model-free: they require neither distributional assumptions on data nor special classes of prediction functions, hence being particularly attractive in practice. Nevertheless, the desired performance guarantee of CP methods relies on *exchangeability* (e.g., the simplest case is when data are i.i.d.), which hardly holds for time series.

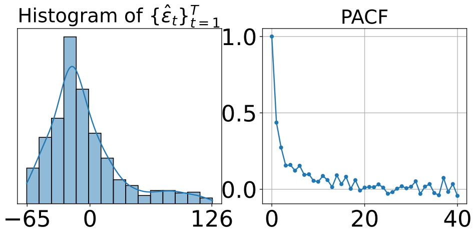


Figure 1: Solar power radiation prediction for downtown Atlanta, Georgia, USA (further explanation in Section 5.2). We use random forest for one-step-ahead prediction. The histogram of prediction residuals (left) shows that residual distribution is highly skewed, and the partial auto-correlation between residuals (right) shows a significant serial correlation among residuals. Thus, it is essential to consider serial dependency when constructing prediction intervals: the serial dependence means that the most recent past residuals contain information about the immediate future ones.

Recently, significant efforts have been made to extend CP methods beyond exchangeable data; several are towards building sequential conformal prediction methods. They typically do so via updating non-conformity scores (e.g., prediction residuals) (Xu & Xie, 2021a;b) and/or adjust significance level α based on rolling coverage of Y_t. This include (Gibbs & Candes, 2021; Zaffran et al., 2022; Feldman et al., 2022; Lin et al., 2022) and specifically, the AdaptCI algorithm, which adjusts the significance level α based on real-time coverage status during prediction—the significance level is lower when the prediction interval at time t fails to contain the actual observation Y_t. The prediction intervals thus have adaptive width based on the updated significance levels and maintain coverage on stock market data in practice. Furthermore, (Barber et al., 2022) proves the coverage gap for non-exchangeable data based on the total variation (TV) distance between the non-conformity scores. The work then proposes NEX-CP, a general re-weighting scheme for non-exchangeable data, where the weights should ideally be chosen to be inversely proportional to the TV distances. The authors demonstrate the robustness of NEX-CP on datasets with change points and/or distribution shifts. For sequential data, (Xu & Xie, 2021b) proposes EnbPI, which updates residuals of ensemble predictors during prediction to more accurately calibrate prediction intervals. In practice, EnbPI can maintain desired $1-\alpha$ coverage for different types of time series. Despite the existing efforts, these sequential CP methods have not exploited serial correlation among non-conformity scores (cf. Figure 1)—they only use empirical quantiles (possibly with fixed weights) of past residuals to compute intervals, which is a drastic difference from SPCI.

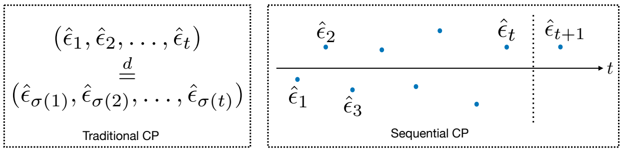

$(\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_t)$
$\overset{d}{=}$
$(\hat{\epsilon}_{\sigma(1)}, \hat{\epsilon}_{\sigma(2)}, \dots, \hat{\epsilon}_{\sigma(t)})$
Traditional CP
$\hat{\epsilon}_2$
$\hat{\epsilon}_t$
$\hat{\epsilon}_{t+1}$
t
$\hat{\epsilon}_1$
$\hat{\epsilon}_3$
Sequential CP

Figure 2: Differences between traditional and sequential Conformal Prediction (CP) methods. In traditional CP, residuals are exchangeable, and the same set of residuals is used throughout the prediction. In contrast, sequential CP assumes an ordering of the potentially non-exchangeable residuals; residuals are available feedback to the prediction algorithms: past residuals are updated to include the new prediction residual $\hat{\epsilon}_{t+1}$ during prediction.

Besides conformal prediction, probabilistic forecasting approaches have also been widely used when building predicting intervals. These approaches typically train a single model to minimize the pinball loss, including the MQ-CNN (Wen et al., 2017), DeepAR (Salinas et al., 2020), Temporal Fusion Transformer (TFT) (Lim et al., 2021), etc. However, comparing to SPCI and related CP works, these approaches have two major limitations. First, they are not “model-free”: special designs of the predictive model and hyper-parameter tuning are required for satisfactory performances. Second, they are not “distribution-free”: distributional assumptions on time-series are often imposed, such as Gaussianity (Salinas et al., 2020). Corresponding theoretical guarantees on constructed prediction intervals are also often lacking. In our experiments, we demonstrate the improved performance of SPCI against DeepAR and TFT.

1.2. Connection with related works

Through theoretical analysis, we find that when using random forest quantile regression, SPCI can be viewed as adaptively learning the (data-dependent) weights of the prediction residuals/non-conformity scores when constructing the prediction intervals using weighted quantile values. Hence, it has an interesting connection to the recent work (Barber et al., 2022), which develops a general conformal prediction framework for non-exchangeable data. In that work, weights are pre-determined and non-adaptive (such as geometrically decaying weights), and the authors also pointed out that “how to choose weights optimally ... is an interesting and important question that we leave for future work” and “leave a more detailed investigation of data dependent weights for future work” (Barber et al., 2022). So our work is a step towards this direction.

We further remark on several key differences of SPCI with prior works. Method-wise, our prediction intervals are constructed using conditional quantile regression functions on *non-conformity scores* (e.g., residuals). In contrast, existing quantile-regression-based conformal prediction methods (Romano et al., 2019; Gupta et al., 2021) directly fit conditional quantile functions on the *response variables* Y, after which the intervals are constructed using *empirical quantiles* of non-conformity scores. Theory-wise, we obtain similar asymptotic conditional coverage for dependent residuals as in (Xu & Xie, 2021b). However, different from that work, we do not assume a particular functional form of the conditional distribution of the scalar output given feature variables.

2. Problem setup

Assume a sequence of observations (X_t, Y_t), $t = 1, 2, \dots$, where Y_t are continuous scalar variables and $X_t \in \mathbb{R}^d$ denote features, which may either be the history of Y_t or contain exogenous variables helpful in predicting the value of Y_t. We can allow observations to be highly correlated under an unknown conditional distribution $Y_t|X_t, \dots, X_1$, and do not assume a particular functional form of the conditional distribution $Y_t|X_t, \dots, X_1$. Let the first T samples $\{(X_t, Y_t)\}_{t=1}^T$ be the training data.

Our goal is to construct prediction intervals sequentially starting from time $T + 1$ such that the prediction intervals will contain the true outcome with a pre-specified high probability $1 - \alpha$ while the prediction interval is as narrow as possible. Here the *significance level* α is user-specified. The prediction intervals $\widehat{C}_{t-1}(X_t)$, which depend on α, are around point predictions $\widehat{Y}_t := \hat{f}(X_t)$ for a given predictive model $\hat{f}$. A commonly used conformity score is the prediction residual:

$$\hat{\epsilon}_t = Y_t - \widehat{Y}_t.$$

We emphasize that our algorithm provides prediction intervals for an arbitrary user-chosen predictive algorithm. Here the subscript $_{t-1}$ indicates the interval is constructed using previous up to $t - 1$ many observations.

There are two types of coverage guarantees to be satisfied by $\widehat{C}_{t-1}(X_t)$. The first is the weaker *marginal* coverage:

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)) \geq 1 - \alpha, \forall t, \tag{1}$$

while the second is the stronger *conditional* coverage:

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)|X_t) \geq 1 - \alpha, \forall t. \tag{2}$$

If $\widehat{C}_{t-1}(X_t)$ satisfies (1) or (2), it is called marginally or conditionally valid, respectively. In terms of the interval width, to avoid vacuous prediction interval $\widehat{C}_{t-1}(X_t)$ (in the extreme case, if one chooses the entire real line for all t, it will always contain the true outcome Y_t with high probability), we should construct intervals with width $|\widehat{C}_{t-1}(X_t)|$ as narrow as possible.

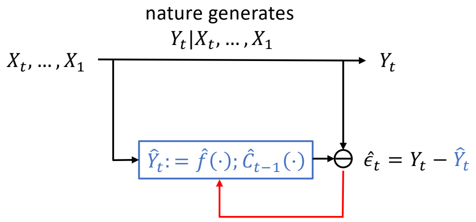

nature generates
$Y_t|X_t, \dots, X_1$
$X_t, \dots, X_1$
Y_t
$\hat{Y}_t := \hat{f}(\cdot); \hat{C}_{t-1}(\cdot)$
$\hat{\epsilon}_t = Y_t - \hat{Y}_t$

Figure 3: Unlike traditional CP methods, sequential CP methods leverage feedback (in red arrow) during prediction. In this work, we use prediction residual $\hat{\epsilon}_t = Y_t - \widehat{Y}_t$ as an example of the non-conformity score.

A natural approach in developing sequential CP methods is constructing sequential prediction intervals using the most recent feedback in predicting Y_t, as shown in Figure 3. However, using the empirical distribution of updated residuals may not fully exploit the temporal dependence across the residuals. Indeed, when residuals are temporally correlated, the past residuals contain information about the distribution of future residuals and can be used to perform "predictive" conformal inference. More precisely, we should use the past residuals to predict the tail probability of the new residual, as doing so may allow certain adaptivity. The above is the main idea of our proposed `SPCI` algorithm.

3. Algorithms

Below, we first consider a simple split conformal prediction as a vanilla baseline approach based on traditional CP, which constructs prediction intervals without considering feedback during prediction. Then, we present the `EnbPI` (Xu & Xie, 2021b) method in sequential CP as a refined approach and illustrate its limitation in using empirical quantile of past residuals. Finally, we introduce the proposed `SPCI` as an improved algorithm for sequential CP for time series data.

3.1. Vanilla split conformal

One of the most commonly used conformal prediction methods is *split conformal* (Papadopoulos et al., 2007), so we describe it as a prototypical example. First, split the indices of training data $[T] := \{1, \dots, T\}$ into two halves $\mathcal{I}_1$ and $\mathcal{I}_2$. Second, fit the prediction model $\hat{f}$ on $\{(X_t, Y_t), t \in \mathcal{I}_1\}$ to make point predictions $\widehat{Y}_t = \hat{f}(X_t), t \neq \mathcal{I}_1$. Third, compute *non-conformity score* on $\mathcal{I}_2$, where a typical choice is the residual. Lastly, let $\mathcal{E}[\mathcal{I}_2] = \{\hat{\epsilon}_j\}_{j \in \mathcal{I}_2}$ and define the prediction interval $\widehat{C}_{t-1}(X_t)$ for $t > T$ as

$$[\hat{f}(X_t) + q_{\alpha/2}(\mathcal{E}[\mathcal{I}_2]), \hat{f}(X_t) + q_{1-\alpha/2}(\mathcal{E}[\mathcal{I}_2])], \tag{3}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile function over a set of values. In particular, the set of non-conformity scores $\{\hat{\epsilon}_j\}_{j \in \mathcal{I}_2}$ is fixed during prediction. When (X_t, Y_t) are exchangeable (i.e., we can shuffle the order of these random variables without affecting the joint distribution), split conformal intervals in (3) reaches exact finite-sample marginal coverage defined in (3). However, without further distribution assumptions, split conformal intervals cannot reach valid conditional coverage in (2) (Foygel Barber et al., 2021).

3.2. EnbPI: Ensemble version using empirical residuals

Compared to split conformal in the previous section, `EnbPI` involves no data-splitting, trains ensemble predictors that make more accurate point predictions and utilizes feedback during prediction on test data. Thus, `EnbPI` is more suitable than split conformal for sequential prediction interval construction. `EnbPI` has the following three steps. First, it leverages training data as much as possible by fitting "leave-one-out" (LOO) ensemble prediction models $\hat{f}_t(X_t) := \phi(\{\hat{f}_b(X_t) : t \notin S_b\})$, where ϕ denotes an arbitrary aggregation function (e.g., mean, median, etc.) over a set of scalars, and $S_b \subset [T]$ is the bootstrap index set used to train the b-th bootstrap estimator $\hat{f}_b$. The point predictor on test data is defined as $\hat{f}(X_t) := \phi(\{\hat{f}_b(X_t)\})$, which aggregates all bootstrap predictions. Second, we obtain residuals using the LOO models $\hat{\epsilon}_t := Y_t - \hat{f}_t(X_t)$. Third, it updates the past residuals during predictions so that the prediction intervals have adaptive width. For a fixed $w \geq 1$, define $\mathcal{E}_t^w := \{\hat{\epsilon}_{t-1}, \dots, \hat{\epsilon}_{t-w}\}$. Then, `EnbPI` intervals $\widehat{C}_{t-1}(X_t)$ have the form:

$$[\hat{f}(X_t) + q_{\alpha/2}(\mathcal{E}_t^T), \hat{f}(X_t) + q_{1-\alpha/2}(\mathcal{E}_t^T)], \tag{4}$$

which utilize the past $w = T$ residuals and greatly resemble traditional CP intervals in (3) due to the use of empirical quantile function $q_{1-\alpha/2}$ to compute interval width.

However, `EnbPI` intervals in (4) can have limitations under dependent residuals. Note that dependent residuals lead to non-equivalence between conditional and marginal distributions of $\hat{\epsilon}_t$, namely $\hat{\epsilon}_t|\mathcal{E}_t^w \neq \hat{\epsilon}_t$ in distribution. More precisely, let $F(z|\mathcal{E}_t^w) := \mathbb{P}(\hat{\epsilon}_t \leq z|\mathcal{E}_t^w)$ be the unknown conditional distribution function of the residual $\hat{\epsilon}_t$, where we implicitly assume the conditional distribution function is invariant over time (i.e., residuals have identical conditional distributions). Based on (4),

$$\begin{aligned}
&\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)|X_t) && (5)\\
&= \mathbb{P}(\hat{\epsilon}_t \in [q_{\alpha/2}(\mathcal{E}_t^T), q_{1-\alpha/2}(\mathcal{E}_t^T)]|X_t)\\
&= F(q_{1-\alpha/2}(\mathcal{E}_t^T)|\mathcal{E}_t^w) - F(q_{\alpha/2}(\mathcal{E}_t^T)|\mathcal{E}_t^w). && (6)
\end{aligned}$$

However, the distribution function F evaluated at the empirical quantiles may not yield the desired coverage. More precisely, define

$$Q_t(p) := \inf\{e^* \in \mathbb{R} : F(e^*|\mathcal{E}_t^w) \geq p\}, \tag{7}$$

which is the p-th quantile of the residual $\hat{\epsilon}_t$. By definition,

$$F(Q_t(1-\alpha/2)|\mathcal{E}_t^w) - F(Q_t(\alpha/2)|\mathcal{E}_t^w) = 1 - \alpha. \tag{8}$$

Thus, in order for EnbPI intervals in (4) to have the desired $1-\alpha$ coverage asymptotically, the empirical quantile must uniformly converge to the actual quantile value, namely:

$$\sup_{p\in[0,1]} |q_p(\mathcal{E}_t^T) - Q_t(p)| \to 0 \text{ as } T \to \infty. \tag{9}$$

However, the condition (9) requires strong assumptions: (Xu & Xie, 2021b) assumes a particular linear functional form of $Y_t|X_t$ (i.e., $Y_t = f(X_t) + \epsilon_t$), which further needs to be consistently estimated as sample size approaches infinity. Such assumptions can impose limitations in practice.

3.3. Proposed SPCI algorithm

Due to the limitations above by split conformal and EnbPI, we propose SPCI in Algorithm 1 as a more general framework than both approaches. In particular, SPCI directly leverages the dependency of $\hat{\epsilon}_t$ on the past residuals when constructing the prediction intervals. Based on the equivalence in (6) and the coverage property in (8), SPCI replaces the empirical quantile with an estimate by a conditional quantile estimator. Specifically, let $\widehat{Q}_t(p)$ be an estimator of the true quantile $Q_t(p)$ in (7) and let $\hat{f}$ be a pre-trained point predictor, SPCI intervals $\widehat{C}_{t-1}(X_t)$ are defined as

$$[\hat{f}(X_t) + \widehat{Q}_t(\hat{\beta}), \hat{f}(X_t) + \widehat{Q}_t(1-\alpha+\hat{\beta})], \tag{10}$$

where $\hat{\beta}$ minimizes interval width:

$$\hat{\beta} = \arg\min_{\beta\in[0,\alpha]} (\widehat{Q}_t(1-\alpha+\beta) - \widehat{Q}_t(\beta)). \tag{11}$$

In particular, SPCI is more general than both EnbPI and split conformal. If we train LOO point predictors, choose the quantile estimator $\widehat{Q}_t(\cdot)$ as the empirical quantile, and use $\hat{\beta} = \alpha/2$, SPCI in (10) reduces to EnbPI in (4). If we follow split conformal prediction to train the point predictor $\hat{f}$, train quantile predictor $\widehat{Q}_t$ on residuals from calibration set, and do no update residuals during prediction, SPCI intervals reduce to the split conformal intervals in (3).

We particularly comment on the computational aspect of fitting conditional quantile estimators $\widehat{Q}_t$, the essential step of SPCI. To train $\widehat{Q}_t$, one minimizes the pinball loss

$$\mathcal{L}(x,\alpha) = \begin{cases} \alpha x & \text{if } x \geq 0, \\ (\alpha-1)x & \text{if } x < 0, \end{cases} \tag{12}$$

which depends on the significance level α. Because SPCI aims to produce intervals as narrow as possible and refits the quantile regression models at each t, it is important to choose quantile regression algorithms that are efficient enough in this sequential setting. In this work, we will use quantile random forest (QRF) (Meinshausen, 2006) to train $\widehat{Q}_t$ and establish coverage guarantees.

We train QRF *auto-regressively* in SPCI to leverage the dependency in residuals. In short, we use the past $w \geq 1$ residuals to predict the conditional quantile of the future (unobserved) residual. More precisely, suppose we have T past residuals $\mathcal{E}_t^T$ available at prediction index t. Let $\tilde{T} := T - w$. For $t' = 1, \ldots, \tilde{T}$, define

$$\tilde{X}_{t'} := [\hat{\epsilon}_{t'+w-1}, \ldots, \hat{\epsilon}_{t'}], \tilde{Y}_{t'} := \hat{\epsilon}_{t'+w}. \tag{13}$$

Thus, feature $\tilde{X}_{t'}$ contains w residuals useful for predicting the conditional quantile of $\tilde{Y}_{t'}$, which is the residual at index $t'+w$. We use the feature $\tilde{X}_{\tilde{T}+1}$ to predict the conditional quantile of $\tilde{Y}_{\tilde{T}+1}$. As a result, the QRF is trained using $\tilde{T}$ training data $(\tilde{X}_{t'}, \tilde{Y}_{t'}), t' = 1, \ldots, \tilde{T}$. When re-fitting the QRF at each prediction index, we re-design these $\tilde{T}$ training data using a sliding window of most recent T residuals. In our experiments, we use the Python implementation of QRF by (Roebroek, 2022).

Algorithm 1 Sequential Predictive Conformal Inference (SPCI)

```
Require: Training data $\{(X_t, Y_t)\}_{t=1}^T$, prediction algorithm $\mathcal{A}$, significance level $\alpha$, quantile regression algorithm $\mathcal{Q}$ .
Output: Prediction intervals $\widehat{C}_{t-1}(X_t), t > T$
Obtain $\hat{f}$ and prediction residuals $\widehat{\boldsymbol{\epsilon}}$ with $\mathcal{A}$ and $\{(X_t, Y_t)\}_{t=1}^T$
for $t > T$ do
    Use quantile regression to obtain $\widehat{Q}_t \leftarrow \mathcal{Q}(\widehat{\boldsymbol{\epsilon}})$
    Obtain prediction interval $\widehat{C}_{t-1}(X_t)$ as in (10)
    Obtain new residual $\hat{\epsilon}_t$
    Update residuals $\widehat{\boldsymbol{\epsilon}}$ by sliding one index forward (i.e., add $\hat{\epsilon}_t$ and remove the oldest one)
end for
```

Remark 1 (SPCI vs. Quantile regression). We further highlight the essential difference of SPCI against quantile regression approaches. In general, quantile regression algorithms rely on minimizing the pinball loss for specific regression algorithms. Doing so can often lead to inaccurate results and require special hyper-parameter tuning. In general, these algorithms can also be computationally expensive to train for multiple significance levels, as the pinball loss depends on α. In contrast, SPCI is compatible with any user-specified point prediction model, remains distribution-free, and provides coverage guarantees (see Theorem 2). Hence, SPCI inherits the main benefits of CP methods. In addition, SPCI leverages the dependency of non-conformity scores by fitting a QRF model of the quantiles (one can use general quantile regression models if desired). Computationally, SPCI is also efficient in test time as the fitting of the QRF model does not rely on the significance level alpha. In practice, we find such a hybrid approach to outperform quantile regression models using deep neural networks (see Table 3).

4. Theory

We first show that when data are exchangeable, one can reach exact marginal coverage when using the empirical quantile function as the quantile regression predictor. We then establish asymptotic coverage upon considering the dependency of estimated residuals. For dependent residuals, we adapt the proof in (Meinshausen, 2006) for independent observations, where we replace the independence assumption with stationary and decaying dependence assumptions. Most proofs and additional theoretical details appear in Appendix A.

4.1. Under exchangeability

We show below that SPCI maintains marginal coverage when data are exchangeable. The proof is standard based on showing the marginal coverage of split conformal prediction

Proposition 1 (Finite-sample marginal coverage under exchangeability (Papadopoulos et al., 2007))**.** *Suppose the data $(X_t, Y_t), t \geq 1$ are exchangeable (e.g., independent and identically distributed). Prediction intervals obtained via Algorithm 2 (i.e., a special version of SPCI) satisfy*

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)) \geq 1 - \alpha.$$

4.2. Beyond exchangeability

The primary theoretical contribution of our work is to show the asymptotic conditional validity of SPCI intervals when the quantile random forest (Meinshausen, 2006) is used as the conditional quantile estimator. Specifically, we have

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t) | X_t) \to 1 - \alpha \text{ as } T \to \infty,$$

which by (6) and (7), is equivalent to proving

$$\sup_{p\in[0,1]} |\widehat{Q}_t(p) - Q_t(p)| \to 0 \text{ as } T \to \infty, \tag{14}$$

where $\widehat{Q}_t(p)$ is the QRF estimator. More precisely, we want to estimate the conditional quantile values of $\tilde{Y}_{\tilde{T}+1}$ given $\tilde{X}_{\tilde{T}+1}$, both of which are defined in (13). Note that (14) for i.i.d. observations has been proven in (Meinshausen, 2006, Theorem 1), so that our analysis also extends the original statement therein to observations with dependency.

We follow the notation in (Meinshausen, 2006) to introduce QRF. For the feature $\tilde{X}_t, t \geq 1$, assume its support $\text{Supp}(\tilde{X}_t) \subset \mathbb{B} \subset \mathbb{R}^p$. We grow the tree $T(\theta)$ with parameter θ as follows: every leaf $l = 1, \ldots, L$ of a tree $T(\theta)$ is associated with a rectangular subspace $R_l \subset \mathbb{B}$. In particular, they are disjoint and cover the entire space $\mathbb{B}$: for every $x \in \mathbb{B}$, there is *one and only one* leaf l, thus denoted as $l(x, \theta)$, such that $x \in R_{l(x,\theta)}$. If we grow K trees, let each of them have separate parameter θ_k. Now, for a given $x \in \mathbb{B}$ and $\tilde{T}$ observed features $\tilde{X}_1, \ldots, \tilde{X}_{\tilde{T}}$, we define the following weights:

$$k_\theta(l) := \#\{j \in \{1, \ldots, \tilde{T}\} : \tilde{X}_j \in R_{l(x,\theta)}\} \tag{15}$$

$$w_t(x, \theta) := \frac{\mathbb{1}(\tilde{X}_t \in R_{l(x,\theta)})}{k_\theta(l)} \tag{16}$$

$$w_t(x) := K^{-1} \sum_{k=1}^{K} w_t(x, \theta_k) \tag{17}$$

For interpretation, (15) counts the "node size" of the leaf $l(x, \theta)$, (16) weighs the i-th observation using whether $\tilde{X}_t$ belongs to this leaf and its node size, and (17) weighs such weights from K trees. Based on weights in (17), the estimated conditional distribution function $\hat{F}(z|x)= \hat{F}(z|\tilde{X}_{\tilde{T}+1} = x)$ is defined as

$$\hat{F}(z|x) := \sum_{t=1}^{\tilde{T}} w_t(x) \mathbb{1}(\tilde{Y}_t \leq z). \tag{18}$$

In retrospect, the estimation in (18) is similar to that under fixed weights by (Barber et al., 2022). The key difference is that (18) uses data-adaptive weights as it exploits the temporal autocorrelation of residuals. In contrast, (Barber et al., 2022) uses fixed and non-adaptive weights.

To show the convergence of the estimated QRF quantile to the true value, we first have the following lemma relating the convergence of quantile estimates to the convergence of corresponding distribution functions.

Lemma 1. *For random variable $\hat{\epsilon}_t$ (i.e., residual in our setup), let $F(z|x)$ be its conditional distribution function and $Q(p) := \inf\{z \in \mathbb{R} : F(z|x) \geq p\}$ be the p-th quantile, which is assumed to be unique. Let $\hat{F}(z|x)$ be an estimator trained on $\tilde{T}$ samples $\{(\tilde{X}_t, \tilde{Y}_t)\}_{t=1}^{\tilde{T}}$. If for all z and x it holds that*

$$\hat{F}(z|x) \to F(z|x) \textit{ in probability as } \tilde{T} \to \infty, \tag{19}$$

then $\widehat{Q}(p) := \inf\{z \in \mathbb{R} : \hat{F}(z|x) \geq p\}$ satisfies $\widehat{Q}(p) \to Q(p)$ in probability for every $p \in (0, 1)$ and x.

Thus, the crux of the remaining analyses relies on showing the point-wise convergence in (19) for the QRF in (18). The case where all data are independent and identically distributed has been addressed in (Meinshausen, 2006, Theorem 1). We address the more general case for dependent observations in Proposition 2.

Proposition 2. *If Assumptions 1—4 defined in Appendix A hold, we obtain the point-wise convergence in* (19) *for QRF.*

We briefly explain and discuss the necessary theoretical assumptions 1—4 used in proving Proposition 2:

- Assumption 1: This assumption states two things. First, the dependency of the covariance of the indicator random variables (defined over the residual quantiles) only depends on the difference in index (see Eq. (22)). Such assumption on residual dependency resembles the weak or wide-sense stationarity assumption. Second, the value of covariances can be uniformly bounded over the conditioning variables by a function $\tilde{g}$, and there is a growth order constraint on $\tilde{g}$ (see Eq. (24)). This condition is imposed to avoid strong dependency among the residuals, which prevents asymptotic consistency of the QRF estimator.

- Assumption 2: This assumption requires that the weights $w_t(x)$ in QRF decay linearly with respect to the number of training samples for QRF. In practice, we often found that the weights decay at such an order.

- Assumption 3 and 4: These distributional assumptions on the conditional quantile function follow those in QRF (Meinshausen, 2006), and they are reasonably mild. In particular, we are not assuming a particular parametric of the conditional quantile function so the results are distribution-free.

We finally obtain the asymptotic guarantee on interval coverage.

Theorem 2 (Asymptotic conditional coverage beyond exchangeability)**.** *Under the same assumptions as Lemma 1 and Proposition 2, as the sample size $T \to \infty$, we have for any $\alpha \in (0, 1)$*

$$|\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)|X_t) - (1-\alpha)| \xrightarrow{p} 0 \tag{20}$$

4.3. Implications of results

We discuss several implications of the results: (1) how the results are distribution-free and model-free; (2) challenges in obtaining interval convergence; (3) the generality of proving guarantees for QRF; (4) convergence analyses beyond using QRF.

Distribution-free & model-free guarantees. Note that our coverage guarantee makes no explicit distributional assumptions of the residuals (e.g., density function has certain parametric form). Instead, our assumptions are on the dependency among the residuals and the regularity of the density functions of the residuals. On the other hand, our results also make no assumptions on the underlying data generation process of Y_t given X_t, in contrast to the linear assumption $Y_t = f(X_t) + \epsilon_t$ in EnbPI (Xu & Xie, 2021b). One can also use arbitrary predictive model to obtain the residuals, rather than relying on special deep neural network architectures (Salinas et al., 2020; Lim et al., 2021).

Interval convergence. Ideally, we wish SPCI intervals in (10) to converge in width to the oracle interval defined by $Y_t|X_t$. However, doing so requires assumptions on the inverse CDF of $Y_t|X_t$, which deviate from our focus on model-free interval construction. Even though such theoretical analyses are lacking, experiments in Section 5 demonstrate that SPCI improves over recent sequential conformal prediction models in many cases.

Generality of QRF. Note that decision trees are simple functions, thus satisfying the assumptions of the *Simple Function Approximation Theorem* (Royden & Fitzpatrick, 1988). In other words, the QRF estimates can theoretically approximate those of any other quantile estimates. As a result, this can be useful if one analyzes the convergence of QRF quantile estimates for residuals with a more general dependency.

Convergence beyond using QRF. The convergence of quantile estimates has been a long-standing question in statistics. In our case, we are particularly interested in the quantile estimates under time-series data. In the past, several lines of work have established such results for different estimators under various assumptions on dependency. (Cai, 2002) studied weighted Nadaraya-Watson quantile estimates for α-mixing sequences. (Biau & Patra, 2011) proposes a nearest-neighbor strategy for stationary and ergodic data. (Zhou & Wu, 2009) analyzed local linear quantile estimators for locally stationary time series. More analyses appear in the survey (Xiao, 2012).

5. Experiments

We empirically demonstrate the improved performance of SPCI over competing sequential CP methods and probabilistic forecasting methods in terms of interval coverage and width. The CP baselines are EnbPI (Xu & Xie, 2021b), AdaptiveCI (Gibbs & Candes, 2021), and NEX-CP (Barber et al., 2022), whose details are in Appendix B. The two probabilistic forecasting methods are DeepAR (Salinas et al., 2020) and TFT (Lim et al., 2021). In all experiments, we obtain LOO point predictors $\hat{f}$ and prediction residuals $\widehat{\epsilon}$ as in EnbPI. Official implementation can be found at https://github.com/hamrel-cxu/SPCI-code.

5.1. Simulation

We first compare SPCI with EnbPI on non-stationary and/or heteroskedastic time-series. We then compare SPCI with NEX-CP on data with distribution drifts and change-points under the setting described in (Barber et al., 2022). Details on data simulation are in Appendix B.1.

(1) Comparison with EnbPI. Given a feature X_t, we specify the true data-generating process as $Y_t = f(X_t) + \epsilon_t$. We

Table 1: Simulation: `EnbPI` vs. `SPCI` on simulated time-series with $\alpha = 0.1$. `SPCI` outperforms `EnbPI` in terms of interval width without sacrificing valid coverage.

	Nstat coverage	Nstat width	Hetero coverage	Hetero width
SPCI	0.94 (2.04e-3)	11.23 (3.37e-2)	0.89 (9.43e-3)	24.09 (8.27e-1)
EnbPI	0.91 (1.11e-3)	25.22 (2.84e-2)	0.92 (1.18e-2)	25.84 (3.47e-1)

simulate two types of time-series data. The first considers non-stationary (Nstat) time-series. The second considers heteroskedastic (Hetero) time-series in which the variance of ϵ_t depends on X_t.

Table 1 compares `EnbPI` with `SPCI`, where both use the random forest regression model to fit the point estimator $\hat{f}$. We see clear improvement of `SPCI`. We suspect the improvement lies in the more adaptive and accurate calibration of quantile values of residual distributions in prediction.

(2) Comparison with NEX-CP. We consider data with distribution drift and changepoints, where data are simulated according to examples in (Barber et al., 2022).

Table 2 shows competitive results of both methods. We notice slight under-coverage by `SPCI` under both settings, despite the much narrower intervals by `SPCI`. When we slightly lower the significance level α, which is held constant when constructing all intervals, `SPCI` maintains valid coverage with comparable interval widths as NEX-CP. Figure A.1 visualizes rolling coverage and width after a burn-in period, with a rolling window of 50 samples. The results are similar to the best model in (Barber et al., 2022, Figure 2). In Appendix B.1, we further explain why `SPCI` tends to under-cover in these settings before α adjustment.

5.2. Real-data examples

We primarily consider three real time-series in this section, whose details are in Appendix B. We first compare the marginal coverage and width of `SPCI` against baseline methods. We then examine the rolling coverage and width of each method to assess their stability during prediction.

Table 2: Simulation: NEX-CP vs. `SPCI` on simulated time-series with 90% target coverage. Entries in the bracket indicate standard deviation over ten trials where data are re-generated. The symbol * denotes results from (Barber et al., 2022, Table 1). Results from the second row are based on $\alpha = 0.09$ (dist. shift) and $\alpha = 0.075$ (change-point).

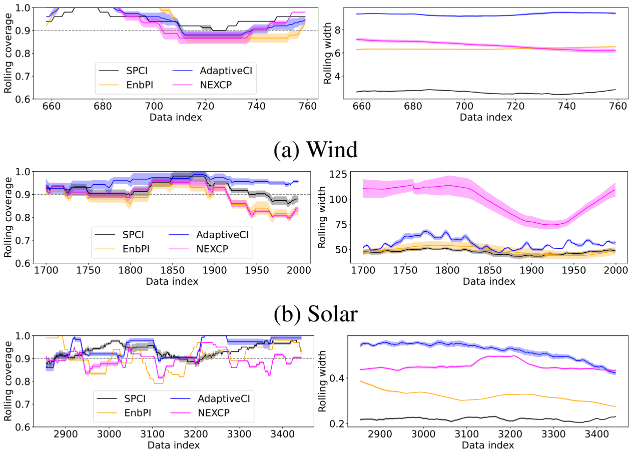
	Drift coverage	Drift width	Change coverage	Change width
SPCI	0.89 (5.04e-3)	3.33 (4.17e-2)	0.87 (2.75e-3)	3.85 (4.12e-2)
SPCI, adjusted α	0.90 (4.63e-3)	3.43 (4.43e-2)	0.90 (3.71e-3)	4.18 (4.89e-2)
NEX-CP*	0.91	3.45	0.91	4.13

We lastly apply `SPCI` on a more challenging multi-step ahead inference case to illustrates its usefulness. We fix $\alpha = 0.1$ and use the first 80% (resp. rest 20%) data for training (resp. testing). For `SPCI` and `EnbPI`, we use the random forest regression model with 25 bootstrap models.

(1) Marginal coverage and width. Table 3 shows the marginal coverage and width of all methods on the three time series. While all methods nearly maintain validity at $\alpha = 0.1$, `SPCI` yields significantly narrower intervals, especially on the wind speed prediction data. Such results illustrate the advantages of fitting conditional quantile regression on residuals for width calibration and training LOO regression predictors for point prediction.

In the appendix, Table A.2 further compares `SPCI` against AdaptiveCI on stock market return data, which are similar to ones used in (Gibbs & Candes, 2021). We show that `SPCI` always maintains valid $1 - \alpha$ coverage and yields narrower intervals than Adaptive CI.

(2) Rolling coverage and width. Besides the marginal metric, we provide further insights into the dynamics of prediction intervals. Figure 4 visualizes the rolling coverage and width of each method, where the metric is computed over a rolling window of size 100 (resp. 50) for the solar and electricity (resp. wind) datasets. The results first show that `SPCI` barely loses rolling coverage when competing methods (e.g., `EnbPI`) can fail to do so. Secondly, `SPCI` intervals are adaptive: they are wider or narrower depending on the data index, which likely reflects higher or less uncertainty in test



(a) Wind

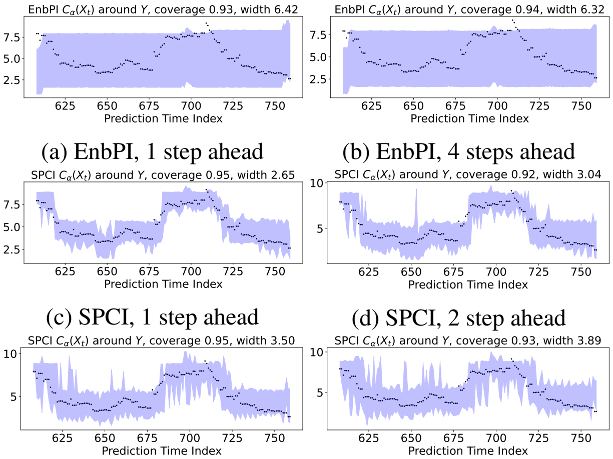
(b) Solar

(c) Electric

Figure 4: Rolling coverage and interval width over three real time series by different methods. `SPCI` in black not only yields valid rolling coverage but also consistently yields the narrowest prediction intervals. Furthermore, the variance of `SPCI` results over trials is also small, as shown by the shaded regions over coverage and width results.

Table 3: Marginal coverage and width by all methods on three real time series. The target coverage is 0.9, and entries in the bracket indicate standard deviation over three independent trials. SPCI outperforms competitors with a much narrower interval width and does not lose coverage.

	Wind coverage	Wind width	Electric coverage	Electric width	Solar coverage	Solar width
SPCI	0.95 (1.50e-2)	2.65 (1.60e-2)	0.93 (4.79e-3)	0.22 (1.68e-3)	0.91 (1.12e-2)	47.61 (1.33e+0)
EnbPI	0.93 (6.20e-3)	6.38 (3.01e-2)	0.91 (6.84e-4)	0.32 (9.11e-4)	0.88 (4.25e-3)	48.95 (3.38e+0)
AdaptiveCI	0.95 (5.37e-3)	9.34 (3.56e-2)	0.95 (1.81e-3)	0.51 (7.25e-3)	0.96 (1.39e-2)	56.34 (1.15e+0)
NEX-CP	0.96 (8.21e-3)	6.68 (7.73e-2)	0.90 (2.05e-3)	0.45 (2.16e-3)	0.90 (7.73e-3)	102.80 (5.25e+0)
DeepAR	0.95 (5.32e-3)	6.86 (7.86e-3)	0.91 (3.45e-3)	0.62 (2.56e0-3)	0.92 (5.35e-3)	80.23 (4.94e+0)
TFT	0.92 (6.34e-2)	7.56 (5.34e-3)	0.95 (2.34e-2)	0.66 (2.34e-3)	0.93 (2.84e-3)	74.82 (4.23e+0)



(a) EnbPI, 1 step ahead (b) EnbPI, 4 steps ahead

(c) SPCI, 1 step ahead (d) SPCI, 2 step ahead

(e) SPCI, 3 step ahead (f) SPCI, 4 step ahead

Figure 5: Multi-step ahead prediction interval construction by SPCI and EnbPI on wind speed data. Compared to EnbPI results in subfigures (a) and (b), SPCI intervals are much narrower and more adaptive—SPCI intervals follow the trajectory of the time-series whereas EnbPI ones are overly conservative. In addition, SPCI interval increase in width as the predictive horizon increases, reflecting the existence of more uncertainty in long horizons.

data. Thirdly, SPCI intervals are evidently narrower than those by competing methods. Lastly, SPCI rolling results have less variance than others such as NEX-CP.

(3) Multi-step predictive inference. In practice, it is often desirable and important to construct $S > 1$ prediction intervals at once. This is a challenging problem for SPCI since it involves estimating the conditional *joint* distribution of S residuals ahead. We thus modify SPCI to tackle this problem through a "divide-and-conquer" approach. Specifically, we apply SPCI S times on lagged training data $(X_t, Y_{t+s}), s = 0, \ldots, S-1$, so that we obtain S fitted QRF estimators to compute the S prediction intervals simultaneously. Additional details including the motivation and algorithm appear in Appendix B.3.

Figure 5 compares SPCI with EnbPI on the wind dataset in terms of multi-step ahead coverage and width. We compare with EnbPI because it supports multi-step ahead prediction in the algorithm, although each batch of $S-$step ahead intervals have the same width by construction. We first note that EnbPI intervals are too wide and non-adaptive, as 4-step ahead intervals may even be narrower than 1-step ahead ones. In contrast, SPCI intervals closely follow the trajectory of actual data and are more adaptive: $S-$step ahead intervals with larger S yield wider intervals on average. This increase in width is expected because there are greater uncertainty when predicting more prediction intervals simultaneously.

6. Conclusions

In this work, we propose SPCI, a general framework for constructing prediction intervals for time series. Similar to existing conformal prediction methods, SPCI is model-free and distribution-free, making it applicable to any time series with arbitrary predictive models. Unlike existing CP methods, SPCI fits quantile regression models on *residuals* to utilize temporal dependency among residuals to achieve more adaptive confidence intervals and better coverage. Theoretical analyses verify the asymptotic valid conditional coverage by SPCI. Experimental results consistently show improved performance by SPCI over existing sequential CP methods.

In the future, we aim to extend SPCI for constructing confidence regions for multi-variate time-series, by further exploiting the dependency among individual uni-variate time-series and designing non-conformity scores that enable efficient interval construction. How to develop the multi-step SPCI in Algorithm 3 to more precisely capture the joint distribution of future residuals is also a promising direction.

Acknowledgement

This work is partially supported by an NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Biau, G. and Patra, B. Sequential quantile prediction of time series. *IEEE Transactions on Information Theory*, 57(3): 1664–1674, 2011.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Brockwell, P. J., Davis, R. A., and Fienberg, S. E. Time series: theory and methods: theory and methods. *Springer Science & Business Media*, 1991.

Cai, Z. Regression quantiles for time series. *Econometric theory*, 18(1):169–192, 2002.

Cochran, J., Denholm, P., Speer, B., and Miller, M. Grid integration and the carrying capacity of the us grid to incorporate variable renewable energy. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2015.

Cody, W. J. and Thacher, H. C. Chebyshev approximations for the exponential integral. *Mathematics of Computation*, 23:289–303, 1969.

Díaz-González, F., Sumper, A., Gomis-Bellmunt, O., and Villafáfila-Robles, R. A review of energy storage technologies for wind power applications. *Renewable and sustainable energy reviews*, 16(4):2154–2171, 2012.

Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

Feldman, S., Bates, S., and Romano, Y. Conformalized online learning: Online calibration without a holdout set. *arXiv preprint arXiv:2205.09095*, 2022.

Fontana, M., Zeni, G., and Vantini, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, pp. 108496, 2021.

Harries, M., of New South Wales. School of Computer Science, U., and Engineering. *Splice-2 Comparative Evaluation: Electricity Pricing*. PANDORA electronic collection. University of New South Wales, School of Computer Science and Engineering, 1999.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., and Horaud, R. A comprehensive analysis of deep regression. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=PM5gVmG2Jj.

Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, 2006.

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

Ridler-Rowe, C. J. A graduate course in probability. *Journal of the Royal Statistical Society. Series A (General)*, 131 (2):230–231, 1968. ISSN 00359238. URL http://www.jstor.org/stable/2343845.

Roebroek, J. Sklearn-quantile, 2022. URL https://github.com/jasperroebroek/sklearn-quantile. (visited on 2023-01-11).

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pp. 3543–3553, 2019.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Royden, H. L. and Fitzpatrick, P. *Real analysis*, volume 32. Macmillan New York, 1988.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

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

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wen, R., Torkkola, K., Narayanaswamy, B. M., and Madeka, D. A multi-horizon quantile recurrent forecaster. In *NeurIPS 2017*, 2017. URL https://www.amazon.science/publications/a-multi-horizon-quantile-recurrent-forecaster.

Xiao, Z. Time series quantile regressions. In *Handbook of statistics*, volume 30, pp. 213–257. Elsevier, 2012.

Xu, C. and Xie, Y. Conformal anomaly detection on spatio-temporal observations with missing data. In *ICML 2021 Workshop on Distribution-free Uncertainty Quantification*, 2021a.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021b.

Zaffran, M., Dieuleveut, A., F'eron, O., Goude, Y., and Josse, J. Adaptive conformal predictions for time series. In *ICML*, 2022.

Zhou, Z. and Wu, W. B. Local linear quantile estimation for nonstationary time series. *The Annals of Statistics*, 37 (5B):2696–2729, 2009.

Zhu, S., Zhang, H., Xie, Y., and Van Hentenryck, P. Multi-resolution spatio-temporal prediction with application to wind power generation. In *2022 INFORMS Workshop on Data Science*, 2021.

A. Proof

Proof of Proposition 1. The proof is standard in conformal prediction literature based on an exchangeability argument. By (3), we know that

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)) = \mathbb{P}(\hat{\epsilon}_t \in [q_{\alpha/2}(\{\mathcal{E}[\mathcal{I}_2]), q_{1-\alpha/2}(\mathcal{E}[\mathcal{I}_2])]).$$

By exchangeability of the original data and the fact that $\hat{f}$ is trained on $(X_t, Y_t), t \in \mathcal{I}_1$, we have $\mathcal{E}[\mathcal{I}_2] = \{\hat{\epsilon}_j\}_{j\in\mathcal{I}_2}$ and $\hat{\epsilon}_t$ are exchangeable. For $p \in [0,1]$, let $q_p := q_{\alpha/2}(\{\hat{\epsilon}_j\}_{j\in\mathcal{I}_2})$. Thus, by exchangeability, we have

$$\begin{aligned}
&\mathbb{P}(\hat{\epsilon}_t \in [q_{\alpha/2}, q_{1-\alpha/2}]) \\
=&\frac{1}{|\mathcal{I}_2|}\sum_{j\in\mathcal{I}_2}\mathbb{P}(\hat{\epsilon}_j \in [q_{\alpha/2}, q_{1-\alpha/2}]) \\
=&\frac{1}{|\mathcal{I}_2|}\mathbb{E}\left[\sum_{j\in\mathcal{I}_2}\mathbb{1}(\hat{\epsilon}_j \in [q_{\alpha/2}, q_{1-\alpha/2}])\right] = 1-\alpha,
\end{aligned}$$

where the last equality holds by the definition of the interval $[q_{\alpha/2}, q_{1-\alpha/2}]$. □

Proof of Lemma 1. First, by (Ridler-Rowe, 1968, Theorem 1, p.127-128), we know that (19) implies

$$\sup_{z\in\mathbb{R}} |\hat{F}(z|x) - F(z|x)| \to 0 \text{ in probability.} \tag{21}$$

Recall that $Q(p)$ is unique. Thus, for any x, there exists $\epsilon = \epsilon(x) > 0$ such that

$$\delta = \delta(\epsilon) := \min\{p - F(Q(p)-\epsilon|x), F(Q(p)+\epsilon|x) - p\} > 0.$$

Namely, there exists a small pertubation of $Q(p)$ whereby the change in the value of the distribution function is at least positive. Thus, we have that

$$\begin{aligned}
\mathbb{P}(|\widehat{Q}(p) - Q(p)| > \epsilon) &\overset{(i)}{=} \mathbb{P}(|F(\widehat{Q}(p)|x) - p| > \delta) \\
&= \mathbb{P}(|F(\widehat{Q}(p)|x) - \hat{F}(\widehat{Q}(p)|x)| > \delta) \\
&\le \mathbb{P}(\sup_{z\in\mathbb{R}} |F(z|x) - \hat{F}(z|x)| > \delta).
\end{aligned}$$

Note that (i) holds because the event $|\widehat{Q}(p) - Q(p)| > \epsilon$ means that $\widehat{Q}(p)$ is at least ϵ far away from $Q(p)$. By monotonicity of the distribution function F, this event implies the occurrence of the event $|F(\widehat{Q}(p)|x) - p| > \delta$.

Now, (21) implies the convergence of estimated quantile values, hence finishing the proof. □

To prove Proposition 2, we need several assumptions followed by interpretation and examples.

Assumption 1. Define $U_t := F(\tilde{Y}_t|X = \tilde{X}_t)$ as the quantile of observations $\tilde{Y}_t$ conditioning on the observed feature $\tilde{X}_t$, where $U_t \sim \text{Unif}[0,1]$. For a $x \in \mathcal{B} := Supp(\{\tilde{X}_t\}_{t\ge 1})$, define the scalar $z[x] := F(z|X = x)$. Given

$$g(i, j, x_1, x_2) := \text{Cov}(\mathbb{1}(U_i \le z[x_1]), \mathbb{1}(U_j \le z[x_2])),$$

we require that for any pair of $x_1, x_2 \in \mathbb{B}$,

$$g(i, j, x_1, x_2) = g(|i-j|, x_1, x_2) \text{ for } i \ne j. \tag{22}$$

In addition, there exists $\tilde{g}$ such that

$$g(k, x_1, x_2) \le \tilde{g}(k)\ \forall x_1, x_2 \in \mathbb{B}, k \ge 1 \tag{23}$$

$$\lim_{\tilde{T}\to\infty}\left[\int_1^{\tilde{T}}\int_1^x \tilde{g}(u)dudx\right]/\tilde{T}^2 \to 0. \tag{24}$$

In other words, (22) assumes that the covariance of the indicator random variables only depends on the difference in index, where this assumption appears widely in the *weak or wide-sense stationary* processes. The difference is that we do not require constant mean values of the indicator variables. In fact, constant mean is impossible, as $\mathbb{E}[\mathbb{1}(U_t \leq z[x])] = z[x]$, whose value changes depending on the conditioning value x. Meanwhile, there is a function $\tilde{g}(k)$ in (23) bounding the covariance uniformly over pairs of values x_1, x_2, and (24) further assumes a restriction on the order of growth of the function $\tilde{g}(k)$. Below are examples of $\tilde{g}(k)$ for which (24) holds and we can also characterize the decay rate of (24).

Example 1 (Finite memory). For some cutoff index $s \in \mathbb{Z}$ and constants $\{c_1, \ldots, c_s\}$,

$$\tilde{g}(k) = \begin{cases} c_k & k \leq s \\ 0 & k > s \end{cases}$$

Showing $\tilde{g}(k)$ in Example 1 satisfies (24) is trivial, with decay rate $O(1/\tilde{T}^2)$. This example appears in stochastic processes with finite memory.

Example 2 (Linear decay). For every $k \geq 1$, $\tilde{g}(k) = \frac{1}{k^p}, p \geq 1$.

Example 2 is weaker than Example 1. To characterize the decay rate, we see that

$$\begin{aligned} \int_1^{\tilde{T}} \int_1^x \tilde{g}(u) du dx &\leq \int_1^{\tilde{T}} \int_1^x 1/u du dx \\ &= \int_1^{\tilde{T}} \log(x) dx = \tilde{T}(\log \tilde{T} - 1). \end{aligned}$$

Thus, $\tilde{T}^{-2} \int_1^{\tilde{T}} \int_1^x \tilde{g}(u) du dx \leq \frac{\tilde{T}(\log \tilde{T} - 1)}{\tilde{T}^2} = O(\log(\tilde{T})/\tilde{T})$. Hence, (24) is proven for Example 2.

Example 3 (Logarithmic decay). For every $k \geq 1$, $\tilde{g}(k) = \left[\frac{1}{\log(k+1)}\right]^p, p \geq 1$.

Example 3 is weaker than the above two examples as it imposes a weaker decay order on the covariance. Lemma 2 presents the proof of (24) for this example, which decays at the order of $O(\frac{1}{2 \log \tilde{T}})$. In general, we wish to show (24) in this example when $p \in (0, 1)$. However, doing so is difficult as the analysis of the integral $\int_1^{\tilde{T}} \int_1^x [\frac{1}{\log(u+1)}]^p du dx$ is complicated. Furthermore, note that $\log(u+1)^p \to 1$ as $p \to 0$, so this integral tends to $\tilde{T}^2/2$, whereby (24) cannot be obtained for small enough p.

Lemma 2. *For $p \geq 1$, we have*

$$\lim_{\tilde{T} \to \infty} \left[\int_1^{\tilde{T}} \int_2^x \frac{1}{\log(u)^p} du dx\right] / \tilde{T}^2 = O\left(\frac{1}{2 \log \tilde{T}}\right).$$

Proof of Lemma 2. First, consider the case where $p = 1$. Define $li(x)$ as the anti-derivative of $1/\log(x)$. To find the growth order of $li(x)$, we note that $li(x) = Ei(\log x)$, where $Ei(x)$ standards for the *exponential integral* with the form $Ei(x) = \int_{-\infty}^x \frac{e^t}{t} dt$. This can be shown via the change of variable $\log(u) = t$. Note that we have the following asymptotic expansion for $Ei(x)$ (Cody & Thacher, 1969):

$$\begin{aligned} Ei(x) &= \frac{\exp(x)}{x}(1 + \frac{1}{x} + \frac{2}{x^2} + \frac{6}{x^3} + \ldots) \\ &= \frac{\exp(x)}{x}(1 + O(1/x)) \text{ when } x > 1. \end{aligned}$$

Thus, $Ei(\log x) = \frac{x}{\log x}(1 + O(1/\log x)) \approx \frac{x}{\log x}$ for large x.

As a result, dropping the constants and small order terms yield

$$\begin{aligned} \int_1^{\tilde{T}} \int_2^x \frac{1}{\log(u)} du dx &= \int_1^{\tilde{T}} Ei(\log x) dx \\ &= \int_1^{\tilde{T}} \frac{x}{\log x} dx \\ &= Ei(2 \log \tilde{T}) \end{aligned}$$

Hence, we have

$$\lim_{\tilde{T}\to\infty}\left[\int_1^{\tilde{T}}\int_2^x \frac{1}{\log(u)}dudx\right]/\tilde{T}^2 = \lim_{\tilde{T}\to\infty} Ei(2\log\tilde{T})/\tilde{T}^2$$
$$= O(\frac{1}{2\log\tilde{T}}).$$

Lastly, when $p > 1$, $\frac{1}{\log u} > [\frac{1}{\log u}]^p$ uniformly for all $u > 1$. Hence, we have

$$\lim_{\tilde{T}\to\infty}\left[\int_1^{\tilde{T}}\int_2^x \frac{1}{\log(u)^p}dudx\right]/\tilde{T}^2 < \lim_{\tilde{T}\to\infty}\left[\int_1^{\tilde{T}}\int_2^x \frac{1}{\log(u)}dudx\right]/\tilde{T}^2,$$

where the latter limit decays at order $O(\frac{1}{2\log\tilde{T}})$ as shown above. □

Assumption 2. The weights $w_t(x)$ in (17) satisfies that for all $x \in \mathbb{B}$, $w_t(x) = O(1/\tilde{T})$.

Assumption 2 imposes the condition on the decay order of each weights. Note that by the definition of $w_t(x)$ in (17) and (Meinshausen, 2006, Assumption 2), we know that $w_t(x) = o(1)$. Assumption 2 thus assumes an exact order of decay of the weights.

Assumption 3. The true conditional distribution function is Lipschitz continuous with parameter L. That is, for all x, x' in the support of the random variable X.

$$\sup_z |F(z|X=x) - F(z|X=x')| \leq L\|x - x'\|_1.$$

Assumption 4. For every x in the support of X, the conditional distribution function $F(z|X=x)$ is continuous and strictly monotonically increasing in z.

We remark that Assumption 3 and 4 are identical to (Meinshausen, 2006, Assumption 4 and 5), respectively.

Proof of Proposition 2. The proof is motivated by the analyses in (Meinshausen, 2006), which assumes $(\tilde{Y}_t, \tilde{X}_t), t \geq 1$ are independent and identically distributed. In essence, we analyze the point-wise difference between the estimate $\hat{F}(z|x)$ in (18) and the true value $F(z|x)$. The difference can then be broken into two terms. Both terms can be bounded by Chebyshev inequalities, leading to convergence to zero.

For each observation $t = 1, \ldots, \tilde{T}$, denote $U_t := F(\tilde{Y}_t|X = \tilde{X}_t)$ as the quantile of the t-th empirical residual $\tilde{Y}_t$. Note that $U_t \sim$ Unif$[0,1]$ by the property of the distribution function, which is continuous by Assumption 4.

By the form of the estimator $\hat{F}(z|x)$ in (18), we break it into two parts:

$$\begin{aligned}\hat{F}(z|x) &= \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(\tilde{Y}_t \leq z)\\ &\overset{(i)}{=} \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq F(z|\tilde{X}_t))\\ &= \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq F(z|x)) + \sum_{t=1}^{\tilde{T}} w_t(x)(\mathbb{1}(U_t \leq F(z|\tilde{X}_t)) - \mathbb{1}(U_t \leq F(z|x))).\end{aligned}$$

The equivalence (i) holds because the event $\{\tilde{Y}_t \leq z\}$ is identical to the event $\{U_t \leq F(z|X = \tilde{X}_t)\}$ under Assumption 4.

Thus, we have that

$$|\hat{F}(z|x) - F(z|x)| \leq \underbrace{\left|\sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq F(z|x)) - F(z|x)\right|}_{(a)} + \\ \underbrace{\left|\sum_{t=1}^{\tilde{T}} w_t(x)(\mathbb{1}(U_t \leq F(z|\tilde{X}_t)) - \mathbb{1}(U_t \leq F(z|x)))\right|}_{(b)}.$$

1) Bound of term (a). The first term can be bounded using Chebyshev inequality. Let $z' := F(z|x)$. Define $U' := \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq z')$. By the linearity of expectation taken over U_t, we have

$$\begin{aligned}\mathbb{E}[U'] &= \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{E}[\mathbb{1}(U_t \leq z')] \\ &= \left[\sum_{t=1}^{\tilde{T}} w_t(x)\right] z' \overset{(i)}{=} z',\end{aligned}$$

where (i) holds under the definition of $w_t(x)$ in (17), which satisfies $\sum_{t=1}^{\tilde{T}} w_t(x) = 1$ as remarked earlier. Now, for any $\epsilon > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq F(z|x)) - F(z|x)\right| \geq \epsilon\right) \\ =&\mathbb{P}(|U' - z'| \geq \epsilon) \leq \text{Var}(U')/\epsilon^2.\end{aligned}$$

Note that

$$\begin{aligned}\text{Var}(U') =&\text{Var}(\sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq z')) \\ =&\underbrace{\sum_{t=1}^{\tilde{T}} w_t(x)^2\text{Var}(\mathbb{1}(U_t \leq z'))}_{(i)} + \underbrace{\sum_{i\neq j} w_i(x)w_j(x)\text{Cov}(\mathbb{1}(U_i \leq z'), \mathbb{1}(U_j \leq z'))}_{(ii)}. \qquad (25)\end{aligned}$$

We need to show that (i) and (ii) in (25) both converge to zero. To show the convergence of (i), we have $w_t(x) = O(1/\tilde{T})$ by Assumption 2 and note that $\text{Var}(\mathbb{1}(U_t \leq z')) = \mathbb{E}(\mathbb{1}(U_t \leq z')^2) - E(\mathbb{1}(U_t \leq z'))^2 = z' - z'^2$. Hence, $\text{Var}(\mathbb{1}(U_t \leq z')) < 1$ and we have $\sum_{t=1}^{\tilde{T}} w_t(x)^2\text{Var}(\mathbb{1}(U_t \leq z')) < \sum_{t=1}^{\tilde{T}} w_t(x)^2 = O(1/\tilde{T})$.

To show the convergence of (ii), we have by Assumption 1 that

$$\begin{aligned}\sum_{i\neq j} w_i(x)w_j(x)\text{Cov}(\mathbb{1}(U_i \leq z'), \mathbb{1}(U_j \leq z')) &\leq \sum_{k=1}^{\tilde{T}-1} O\left(\frac{\tilde{T}-k}{\tilde{T}^2}\right)\tilde{g}(k) \\ &\leq \int_1^{\tilde{T}} O\left(\frac{\tilde{T}-k}{\tilde{T}^2}\right)\tilde{g}(k)dk \\ &=O\left(\tilde{T}^{-1}\right)\int_1^{\tilde{T}} \tilde{g}(k)dk - O\left(\tilde{T}^{-2}\right)\int_1^{\tilde{T}} k\tilde{g}(k)dk \\ &=O\left(T^{-1}\right)[G(\tilde{T}) - G(1)] - O\left(\tilde{T}^{-2}\right)\int_1^{\tilde{T}} k\tilde{g}(k)dk,\end{aligned}$$

where $G(x) := \int_1^x \tilde{g}(k)dk$ is the anti-derivative. Using integration by part with $u = k, dv = \tilde{g}(k)dk$, we have

$$\int_1^{\tilde{T}} k\tilde{g}(k)dk = \tilde{T}G(\tilde{T}) - G(1) - \int_1^{\tilde{T}} G(x)dx.$$

Thus, dropping constants and small order terms yield

$$\sum_{i\neq j} w_i(x)w_j(x)\text{Cov}(\mathbb{1}(U_i \leq z'), \mathbb{1}(U_j \leq z')) \leq \left[\int_1^{\tilde{T}} \left[\int_1^x \tilde{g}(k)dk\right] dx\right] /\tilde{T}^2.$$

By (24) in Assumption 1, we thus have the desired convergence result.

2) Bound of term (b). Define $W := \sum_{t=1}^{\tilde{T}} w_t(x)\mathbb{1}(U_t \leq F(z|\tilde{X}_t))$. Note that $\mathbb{E}(W) = \sum_{t=1}^{\tilde{T}} w_t(x)F(z|\tilde{X}_t)$. We have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(|W - \mathbb{E}(W)| > \epsilon)\\
&\leq \text{Var}(W)/\epsilon^2\\
&= (\epsilon)^{-2}\left[\sum_{t=1}^{\tilde{T}} w_t(x)^2\text{Var}(\mathbb{1}(U_t \leq F(z|\tilde{X}_t))) + \sum_{i\neq j} w_i(x)w_j(x)\text{Cov}(\mathbb{1}(U_i \leq F(z|\tilde{X}_i)), \mathbb{1}(U_j \leq F(z|\tilde{X}_j)))\right].
\end{aligned}$$

By the same argument for bounding term (a) above, we have that $W \xrightarrow{p} \mathbb{E}[W]$ as sample size $\tilde{T} \to \infty$.

As a result, we have

$$\left|\sum_{t=1}^{\tilde{T}} w_t(x)(\mathbb{1}(U_t \leq F(z|\tilde{X}_t)) - \mathbb{1}(U_t \leq F(z|x)))\right| \xrightarrow{p} \left|\sum_{t=1}^{\tilde{T}} w_t(x)(F(z|\tilde{X}_t) - F(z|x))\right|.$$

By Assumption 3, we have

$$\left|\sum_{t=1}^{\tilde{T}} w_t(x)(F(z|\tilde{X}_t) - F(z|x))\right| \leq \sum_{t=1}^{\tilde{T}} w_t(x)L\|\tilde{X}_t - x\|_1.$$

The rest of proof follows due to (Meinshausen, 2006, Lemma 2), which shows that

$$\sum_{t=1}^{\tilde{T}} w_t(x)\|\tilde{X}_t - x\|_1 = o_p(1).$$

□

Proof of Theorem 2. Under SPCI interval construction in (10), the equivalence in (6) implies that

$$\mathbb{P}(Y_t \in \widehat{C}_{t-1}(X_t)|X_t) = F(\widehat{Q}_t(1-\alpha+\hat{\beta})|\mathcal{E}_t^w) - F(\widehat{Q}_t(\hat{\beta})|\mathcal{E}_t^w),$$

where $\widehat{Q}_t(p), p \in [0,1]$ is the estimated p-th quantile of $\hat{\epsilon}_t$, $F(z|\mathcal{E}_t^w)$ is the unknown distribution function of $\hat{\epsilon}_t$, and $\hat{\beta}$ minimizes interval width per the procedure in Algorithm 1.

To finish the proof, by Proposition 2, we know that the conditional distribution estimator $\hat{F}(z|\mathcal{E}_t^w)$ using QRF converges point-wise to the true $F(z|\mathcal{E}_t^w)$ as the sample size (hence the number of residuals) approaches infinity. By Lemma 1, we thus know that $\widehat{Q}_t(p) \to Q_t(p)$ in probability for all $p \in [0,1]$.

We can thus use the continuous mapping theorem (Van der Vaart, 2000, Theorem 2.3) to finish the proof: by Assumption 3, the true conditional distribution function F is absolutely continuous and therefore differentiable almost everywhere. Thus, the set of discontinuity points of F has measure zero. As the number of data $\tilde{T} \to \infty$ when training QRF, we finally have that in probability,

$$\begin{aligned}
&F(\widehat{Q}_t(1-\alpha+\hat{\beta})|\mathcal{E}_t^w) - F(\widehat{Q}_t(\hat{\beta})|\mathcal{E}_t^w)\\
\to &F(Q_t(1-\alpha+\hat{\beta})|\mathcal{E}_t^w) - F(Q_t(\hat{\beta})|\mathcal{E}_t^w) = 1-\alpha.
\end{aligned}$$

□

B. Experimental details

(1) Baseline methods. We compare `SPCI` with three recent CP methods for non-exchangeable data or time series, which have also been carefully described in the literature review. In particular, they all leverage the feedback Y_t after it is sequentially revealed.

- `EnbPI` (Xu & Xie, 2021b) proposes a general framework for constructing time-series prediction intervals. In particular, it fits LOO regression models and uses residuals as non-conformity scores. Comparing our use of `SPCI` in experiments, the only difference appears in using conditional rather than empirical quantiles for the calibration of interval width.
- AdaptiveCI (Gibbs & Candes, 2021) is an adaptive procedure that adjusts the significance level α based on historical information of interval coverage. It leverages CQR (Romano et al., 2019) to produce intervals that maintain coverage validity in theory. We use the quantile random forest as the predictor and update α according to the simple online update (ibid., Eq (2)).
- NEX-CP (Barber et al., 2022) uses weighted quantiles to tackle arbitrary distribution drift in test data. In particular, the implementation is based on full conformal with weighted least squares regression models, which empirically yields more stable coverage than the naive split conformal method.

(2) Real-data description. We describe the three real time-series for results in Section 5.2. The first dataset is the wind speed data (m/s) at wind farms operated by the Midcontinent Independent System Operator (MISO) in the US (Zhu et al., 2021). The wind speed record was updated every 15 minutes over a one-week period in September 2020. The second dataset contains solar radiation information[1] in Atlanta downtown, which is measured in Diffuse Horizontal Irradiance (DHI). The full dataset contains a yearly record in 2018 and is updated every 30 minutes. We remark that uncertainty quantification for both wind and solar is important for accurate and reliable energy dispatch. The last dataset tracks electricity usage and pricing (Harries et al., 1999) in the states of New South Wales and Victoria in Australia, with an update frequency of 30 minutes over a 2.5-year period in 1996–1999. We are interested in tracking the quantity of electricity transferred between the two states.

B.1. Simulation

We first describe details regarding data simulation procedures. We then show additional rolling coverage and width results when comparing with NEX-CP.

B.1.1. Data simulation

For the results in Table 1, we simulate the non-stationary and heteroskedastic time-series as follows:

1. Non-stationary (Nstat) time-series: We let

$$\begin{aligned}
f(X_t) &= g(t)h(X_t). \\
g(t) &= \log(t')\sin(2\pi t'/12), t' = \text{mod}(t,12). \\
h(X_t) &= (|\beta^T X_t| + (\beta^T X_t)^2 + |\beta^T X_t|^3)^{1/4}.
\end{aligned} \tag{26}$$

Note that the model in (26) can represent non-stationary time-series due to additional time-related effects (e.g., time drift, seasonality, periodicity, etc.). For a fixed window size $w \geq 1$, each feature observation $X_t = [Y_{t-w}, \ldots, Y_{t-1}]$ contains

[1] Collected from National Solar Radiation Database (NSRDB): https://nsrdb.nrel.gov/.

Table A.1: Simulation on non-stationary time-series: the setup is identical to Table 1. We compare SPCI against baseline CP methods when no time information is assumed known (i.e., $\tilde{X}_t = X_t$).

SPCI		EnbPI		AdaptiveCI		NEX-CP	
Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
0.92 (2.75e-3)	12.96 (2.56e-2)	0.90 (2.21e-3)	25.41 (4.79e-2)	0.90 (4.12e-3)	28.00 (5.81e-2)	0.93 (3.10e-3)	46.50 (6.29e-2)

the past w observations of the response Y. We sample the errors ϵ_t from an AR(1) process, where $\epsilon_t = \rho\epsilon_{t-1} + e_t$ and e_t are *i.i.d.* normal random variables with zero mean and unit variance with $\rho = 0.6$.

We want to compare the performance of EnbPI and SPCI assuming no feature mis-specification, so that the only difference in interval coverage/width lies in how the residuals are used to construct the intervals. Therefore, because f in (26) explicitly depends on t and X_t, we use the new feature $\tilde{X}_t := [\text{mod}(t, 12), X_t]$ to predict Y_t. We acknowledge that in practice, the true periodicity constant 12 in (26) is unknown, and one must estimate it before constructing the new feature $\tilde{X}_t$. Meanwhile, Table A.1 compares all four CP methods when the time information is unknown (i.e., $\tilde{X}_t = X_t$), and we still observe much narrower intervals by SPCI than the baselines.

2. Heteroskedastic (Hetero) time-series: We let

$$f(X_t) = (|\beta^T X_t| + (\beta^T X_t)^2 + |\beta^T X_t|^3)^{1/4}. \tag{27}$$

$$\text{Var}(\epsilon_t) = \sigma(X_t)^2, \sigma(X_t) = \mathbf{1}^T X_t. \tag{28}$$

Note that the model above represents the generalized autoregressive conditional heteroskedasticity (GARCH) model (Engle, 1982), where variances of response Y_t depend on its feature X_t. We let features $X_t \in \mathbb{R}^{20}$, with i.i.d. entries from Uniform$[0, e^{0.01\text{mod}(t,100)})$. Due to heteroskedastic errors, we estimate conditional quantile of normalized residuals $\hat{\epsilon}_t := (Y_t - \hat{f}_t(X_t))/\hat{\sigma}(X_t)$ and multiply the quantile values by estimates $\hat{\sigma}(X_t)$ to construct the prediction intervals.

For the simulated results in Table 2, the data with distribution-shift and change-points are simulated as follows. For $N = 2000$ and $X_i \sim \mathcal{N}(0, \boldsymbol{I}_4), i = 1, \ldots, N$:

1. Distribution-drift (Drift): $Y_i \sim X_i^T\beta_i + \mathcal{N}(0, 1)$, where $\beta_1 = (2, 1, 0, 0), \beta_N = (0, 0, 2, 1)$, and $\beta_i, i = 2, \ldots, N-1$ is a linear interpolation of β_1 and β_N.
2. Changepoints (Change): $Y_i \sim X_i^T\beta_i + \mathcal{N}(0, 1)$,

$$\beta_1 = \ldots = \beta_{500} = (2, 1, 0, 0)$$
$$\beta_{501} = \ldots = \beta_{1500} = (0, -2, -1, 0)$$
$$\beta_{1501} = \ldots = \beta_N = (0, 0, 2, 1).$$

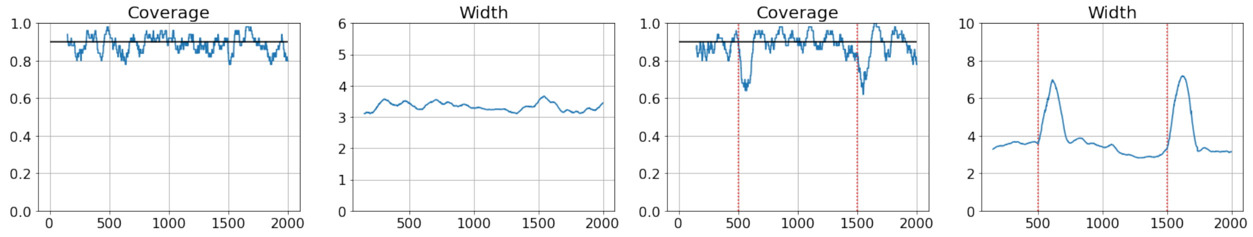
Similar to NEX-CP, we apply SPCI after a burn-in period of the first 100 sample points, and in addition, adaptively refit the point estimator $\hat{f}$ using a rolling window of $\min(T, T_0)$ points during testing for $T = 101, \ldots, 2000$. We choose $T_0 = 300$ under distribution shifts and $T_0 = 200$ under changepoints. Similar to NEX-CP, we use weighted linear regression with exponentially decaying weights to train the point estimator $\hat{f}$ in SPCI.

B.1.2. Comparison with NEX-CP

We explain why SPCI tends to under-cover in these settings before α adjustment. We suspect the primary reasons are that prediction residuals $\hat{\epsilon}_i$ in these settings are *(nearly) independent yet non-identically distributed.* More precisely, regarding independence, suppose we use the split conformal framework in SPCI to train $\hat{f}$ and obtain residuals on the calibration set. We thus have that for each prediction residual $\hat{\epsilon}_i$ in the calibration set,

$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i \sim (X_i^T\beta_i + \mathcal{N}(0, 1)) - \hat{f}(X_i). \tag{29}$$

Note that X_i are all independent by design. Except for the possible dependency in β_i, which is zero in the change-point setting, the (unobserved) test residual $\hat{\epsilon}_{T+1} \perp\!\!\!\perp \hat{\epsilon}_{T+1-k}, k \geq 1$, where $\perp\!\!\!\perp$ denotes independence of random variables.



(a) Distribution shift

(b) Changepoint

Figure A.1: Rolling coverage and width during test time without adjusted α values. Target coverage at 0.9 is marked in the black lines. In (b), the two changepoints are marked in dotted red line at time indices 500 and 1500.

Table A.2: Marginal coverage and width by `SPCI` and AdaptiveCI on three NASDAQ stock market data. The tarrget coverage is 0.9, and entries in the bracket indicate standard deviation over three independent trials.

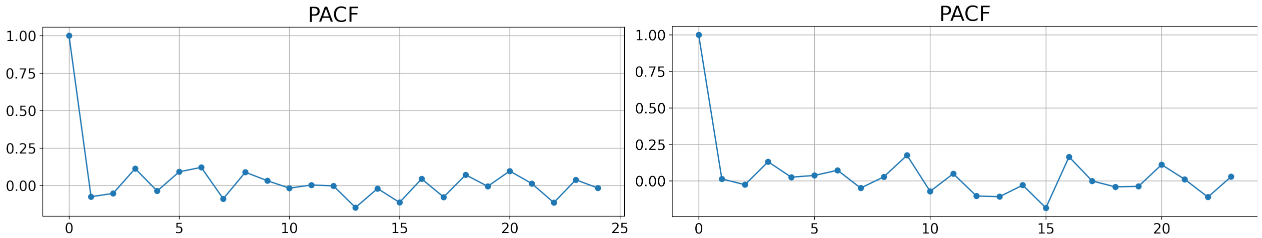
Method	Company AJISF		Company AGTC		Company AAVL	
	Coverage	Width	Coverage	Width	Coverage	Width
SPCI	0.89 (2.34e-3)	17.64 (1.24e-1)	0.95 (2.43e-3)	2.89 (5.23e-2)	0.81 (3.64e-3)	1.03 (2.34e-2)
AdaptiveCI	0.94 (3.43e-3)	30.20 (2.53e-1)	0.71 (1.53e-2)	5.88 (7.43e-2)	0.64 (2.32e-2)	2.18 (3.37e-2)

We empirically verify the independence of residuals through the PACF plot in Figure A.2. On the other hand, regarding non-identical distribution, because of drifts or changepoints through the changes in β_i, the residuals do not follow the same distribution. Thus, the QRF estimated on past residuals may not be a desirable estimator for the conditional quantile of the test residual $\hat{\epsilon}_{T+1}$, hence weakening the performance of `SPCI` in this setting.

B.2. Additional real-data comparisons

We compare `SPCI` with AdaptiveCI on stock market data. Specifically, the dataset is publicly available on Kaggle `https://www.kaggle.com/datasets/paultimothymooney/stock-market-data`, where we are interested in constructing the prediction intervals for the closing price. We randomly select three NASDAQ stock from three companies. Table A.2 shows several findings:

- When AdaptiveCI and SPCI both yield valid coverage (on company AJISF), the width of SPCI is significantly narrower.
- Even when AdaptiveCI loses coverage and SPCI maintains coverage (on company AGTC), the width of SPCI is still significantly narrower.
- When both methods lose coverage (on company AAVL), the loss by SPCI is less and SPCI still yields narrower intervals.



(a) Distribution shift

(b) Changepoint

Figure A.2: PACF using 300 residuals (dist. shift) and 200 residuals (change-point). We see near independence of the residuals, which are non-identically distributed due to the data generation.

Algorithm 2 SPCI for exchangeable data (based on split conformal)

```
Require: Training data $\{(X_t, Y_t)\}_{t=1}^T$, significance level $\alpha$.
Output: Prediction intervals $\widehat{C}_{t-1}(X_t), t > T$
Randomly split $\{1, \ldots, T\}$ into disjoint index sets $\mathcal{I}_1$ and $\mathcal{I}_2$.
Train a point predictor $\hat{f}$ with $\{(X_t, Y_t)\}_{t \in \mathcal{I}_1}$.
Obtain residuals $\hat{\epsilon}_t := Y_t - \hat{f}(X_t)$ for $t \in \mathcal{I}_2$.
for $t > T$ do
    Return the prediction interval $\widehat{C}_{t-1}(X_t)$ as in (3).
end for
```

B.3. Multi-step inference

(1) Motivation and setup. We first motivate the study of multi-step ahead prediction interval. For examples in Section 5.2, all intervals are one step ahead: the response variable Y_t is revealed *before* $\widehat{C}_{t-1}(X_t)$ is constructed, which is the prediction interval for Y_{t+1}. Such immediate feedback is advantageous for all adaptive methods as they thus have access to the most up-to-date information about the data process. Nevertheless, such access can be neither feasible nor desirable for some use cases. In energy systems such as wind or solar prediction, we often need multiple forecasts spanning a long enough future horizon to allow enough time for subsequent dispatch. Meanwhile, lags in data collection can limit the availability of feedback—for $S > 1$, Y_t may not be revealed until all S intervals ahead are constructed.

We consider the following multi-step ahead prediction setting. Fix a value of $S \geq 1$, which denotes the $s-$step ahead prediction setting ($S = 1$ refers to examples in earlier sections). Features $X_t = [Y_{t-1}, \ldots, Y_{t-\tau}]$ are auto-regressive with a pre-specified window $\tau \geq 1$. At prediction time t, we need to construct S prediction intervals at once for time indices $t, \ldots, t+S-1$. In particular, responses $Y_t, \ldots, Y_{t+S-1}$ (and thus features $X_{t+1}, \ldots, X_{t+S}$) are not available until we construct prediction intervals at indices $t+S, \ldots, t+2S-1$.

(2) Multi-step SPCI algorithm. Note that constructing multi-step ahead prediction intervals using SPCI involves estimating the *joint* distribution of $\hat{\epsilon}_{t+1}, \ldots, \hat{\epsilon}_{t+S}$ every S test indices. Doing so can be highly challenging. Instead, we take a simplified "divide-and-conquer" approach based on the LOO fitting in EnbPI. First, we train S sets of LOO predictors for estimating the value of $\widehat{Y}_{t+j}, j = 0, \ldots, S-1$. This is implemented by fitting B bootstrap models on each lagged data $\{(X_t, Y_{t+s})\}_{t=1}^{T-s+1}, s = 1, \ldots, S$. Then, we compute residuals only at $t = 1 + kS : kS \leq T-1$. We do so because on test data, new feature X_t and output Y_t are revealed only in every S step. Lastly, we fit QRF S times using past residuals with lags to obtain s prediction intervals at once.

We briefly compare and contrast Algorithm 1 (SPCI) and 3 (multi-step ahead SPCI) when LOO point predictors are trained. Computationally, we need to refit $S-1$ more sets of LOO predictors in multi-step ahead SPCI for point prediction. On the other hand, both algorithms fit the same number of QRF regressors for constructing prediction intervals. In practice, multi-step SPCI is expected to yield wider intervals as S increases because there is greater uncertainty when fitting the baseline regression or QRF on lagged data. A simple example is the $AR(1)$ process where $x_t = ax_{t-1} + \epsilon_t, \epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Using the present feature x_{t-1}, we have $x_{t+S} = a^{S+1}x_{t-1} + \sum_{i=1}^S a^{i-1}\epsilon_{t+i}$, whereby the error distribution $a^{i-1}\epsilon_{t+i} \sim N(0, \sum_{i=1}^S a^{2(i-1)})$, so width naturally increases.

C. Additional technical details

We first present the SPCI algorithm for exchangeable data in Algorithm 2. We then present the SPCI algorithm for multi-step ahead inference in Algorithm 3.

Algorithm 3 Multi-step `SPCI` (based on LOO prediction in `EnbPI` (Xu & Xie, 2021b))

```
Require: Training data $\{(X_t, Y_t)\}_{t=1}^T$, significance level $\alpha$, number of bootstrap estimators $B$, aggregation function $\phi$,
    conditional quantile regression algorithm $\mathcal{Q}$, multi-step size $S > 1$.
Output: Prediction intervals $\widehat{C}_{t-1}(X_t), t > T$
for $s = 1, \dots, S$ do {$\triangleright$ $s$-step ahead model fitting}
    Sample with replacement $B$ index sets, each of size $T - s + 1$:
    $\{S_b : S_b \subset \{1, \dots, T - s + 1\}\}_{b=1}^B$.
    Train $B$ corresponding bootstrap estimators $\{\hat{f}^b\}_{b=1}^B$ on data $\{(X_t, Y_{t+s-1}) : t \in S_b\}$.
    {$\triangleright$ Leave-one-out aggregation}
    Initialize $\widehat{\boldsymbol{\epsilon}} = [\,]$
    for $t = 1, 1 + S, \dots, 1 + kS$ such that $kS \leq T - 1$ do
        $\hat{f}_t^s(X_t) = \phi(\{\hat{f}^b(X_t), t \notin S_b\}_{b=1}^B)$
        $\widehat{\boldsymbol{\epsilon}}$.append$(Y_{t+s-1} - \hat{f}_t^s(X_t))$
    end for
end for
for $t > T$ do {$\triangleright$ Interval construction}
    Compute $s = \text{mod}(t - T, S + 1)$ and $t' = t - s$
    {$\triangleright$ $t'$ denotes the most recent index where residual $\hat{\epsilon}_{t'}$ and feature $X_{t'+1}$ are available.}
    if $s = 1$ then {$\triangleright$ Fit quantile regressors with updated residuals}
        Re-fit $S$ quantile estimators $\{\widehat{Q}_t(\cdot\,; s')\}_{s'=1}^S$ with $\{(\hat{\epsilon}_j^w, \hat{\epsilon}_{j+s'-1})\}_{j=t-T+w}^{t-1-(S-1)}$.
    end if
    Compute $\hat{\beta} = \arg\min_{\beta \in [0,\alpha]} (\widehat{Q}_t(1 - \alpha + \beta; s) - \widehat{Q}_t(\beta; s))$ using $\hat{\epsilon}_{t'}^w$.
    $\widehat{C}_{t-1}(X_t) = [\hat{Y}_t + w_{\text{left}}(t), \hat{Y}_t + w_{\text{right}}(t)]$, where
    $\hat{Y}_t = \phi(\{\hat{f}_j^s(X_{t'+1})\}_{j=1}^{T/S}), w_{\text{left}}(t) = \widehat{Q}_t(\hat{\beta}; s), w_{\text{right}}(t) = \widehat{Q}_t(1 - \alpha + \hat{\beta}; s)$ .
end for
```