Conformal Prediction for Time Series

Chen Xu and Yao Xie, *Member, IEEE*

***Abstract*—We present a general framework for constructing distribution-free prediction intervals for time series. We establish explicit bounds on the conditional and marginal coverage gaps of estimated prediction intervals, which asymptotically converge to zero under additional assumptions. We also provide similar bounds on the size of set differences between oracle and estimated prediction intervals. To implement this framework, we introduce an efficient algorithm called `EnbPI`, which utilizes ensemble predictors and is closely related to conformal prediction (CP) but does not require data exchangeability. Unlike other methods, `EnbPI` avoids data-splitting and is computationally efficient by avoiding retraining, making it scalable for sequentially producing prediction intervals. Extensive simulation and real-data analyses demonstrate the effectiveness of `EnbPI` compared to existing methods.**

***Index Terms*—Time series predictive inference, conformal prediction.**

I. Introduction

MODERN applications, including energy and supply chains [1], [2], require sequential prediction with uncertainty quantification for time-series observations with highly complex dependency. In addition to point prediction, it is typical to construct prediction intervals for uncertainty quantification, a fundamental task in statistics and machine learning.

Constructing accurate prediction intervals for time series is highly challenging yet crucial in many high-stakes applications. In power systems, as outlined in the National Renewable Energy Lab report [2], solar and wind power generation data are non-stationary, exhibit significant stochastic variations, and have spatial-temporal correlations among regions. The inherent randomness of renewable energy sources presents significant challenges for prediction and inference. To overcome these challenges, it is essential to use historical data to accurately predict energy levels from wind farms and solar roof panels and establish prediction intervals. These prediction intervals provide critical information for power network operators, enabling them to understand the uncertainty of the power generation and make necessary arrangements. Incorporating renewable energy into existing power systems requires the prediction of power generation with uncertainty quantification [3], [4]. Although there are various neural-network based quantile prediction models [5], [6], the resulting prediction intervals frequently lack theoretical guarantees, causing concern about their reliability in high-stakes situations. Currently, there is a need for a distribution-free framework that produces prediction intervals for time-series data, along with provable guarantees for interval coverage, which remains an open question in the field.

In addition to the difficulties posed by the inherent stochasticity of time-series, constructing prediction intervals for user-specified predictive models also presents further challenges. For example, complex prediction models such as random forest [7] and deep neural networks [8] are often employed for accurate predictions. Unlike classical linear regression models, these prediction algorithms do not have straightforward methods for calculating prediction intervals. To construct prediction intervals for such models, practitioners often resort to heuristics like bootstrapping, which lack guarantees. In practice, ensemble methods [9] are also frequently used to enhance prediction performance by combining multiple prediction algorithms, further complicating the model. Despite this, constructing efficient prediction intervals for time-series data using general prediction methods, which can be arbitrarily complex, remains an under-explored area.

A. Contributions

In this paper, we develop distribution-free prediction intervals for time series data with a coverage guarantee, inspired by recent works on conformal prediction. Our proposed method, `EnbPI`, can provide prediction intervals for ensemble algorithms. The main contributions of this paper are summarized as follows.

- We present a general framework for constructing prediction intervals for time series, which can be asymmetrical. We theoretically upper-bound the conditional and marginal coverage gaps, which converge to zero under mild assumptions on the dependency of stochastic errors and the quality of estimation. We also obtain similar bounds on the size of the set difference between the *oracle* and estimated prediction intervals.
- We develop `EnbPI`, a robust and computationally efficient algorithm for constructing prediction intervals around ensemble estimators. The algorithm is designed to avoid expensive model retraining during prediction and requires no data splitting, thanks to a carefully constructed bootstrap procedure. `EnbPI` is particularly suitable for small-sample problems, and its versatility makes it applicable in various practical settings, such as network prediction and anomaly detection.

Manuscript received 14 July 2022; revised 13 February 2023; accepted 26 April 2023. Date of publication 8 May 2023; date of current version 5 September 2023. This work was supported in part by the NSF CAREER under Grant CCF-1650913, and in part by the NSF under Grants DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210. Recommended for acceptance by S. C. H. Hoi. *(Corresponding author: Yao Xie.)*

The authors are with the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: cx9711@gatech.edu; yao.xie@isye.gatech.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3272339, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3272339

- We present extensive numerical experiments to study the performance of EnbPI on simulated and real-time series data. The results show that EnbPI can maintain a target coverage when other competing methods fail to do so, and it can yield shorter intervals. Additionally, the experiments demonstrate that EnbPI is robust to missing data.

The rest of this paper is organized as follows. Section II describes the problem setup and introduces the oracle prediction interval, which motivates our proposed method. Section III presents asymptotic guarantees for the interval coverage and width and highlights the generality of such guarantees. Section IV presents EnbPI. Section V contains numerical examples with simulated and real data that compare EnbPI with competing methods to demonstrate its good performance in various scenarios. Section VI extends the use of EnbPI when a change point exist. Section VII concludes the paper with discussions. Appendix A, (available online), contains proofs and B contains experiments. Code for this paper can be found at https://github.com/hamrel-cxu/EnbPI.

B. Literature Review

Conformal Prediction (CP) is a popular method for constructing distribution-free prediction intervals. It was formally introduced in [10], and it assigns "conformity scores" to both training and test data. By inverting hypothesis tests using these scores, prediction intervals can be obtained for the test data. It has been shown that under the assumption of exchangeability in data, this procedure generates valid marginal coverage for the test point. Many CP methods have been developed to quantify uncertainty in predictive models. To efficiently compute the conformity scores, a data-splitting method is developed in [11], which computes the scores on a hold-out set of the training data. [12] builds on this data-splitting idea for quantile regression models. To avoid data splitting which affects the accuracy of trained predictive model, "leave-one-out" (LOO) CP methods are developed to use the entire training samples for computing prediction residuals, a particular choice of conformity scores [13]. Subsequent works develop more computationally efficient way of training LOO models [14] and generalize the approach to other conformity scores [15]. Comprehensive surveys and tutorials can be found in [10], [16]. Although no assumption other than data exchangeability is required for marginally exact coverage, the exchangeability assumption is hardly reasonable for time series, making works above not directly applicable to our setting.

Adapting CP methods beyond exchangeable data has also been gaining significant interest. A widely popular type of approach assumes unknown distribution shifts in the test data and weighs the past conformity scores to restore valid coverage. For instance, the work by [17] uses weighted conformal prediction intervals when the test data distribution is proportional to the training distribution. The work by [18] builds on this idea when the shifted test distribution lies in an f-divergence ball around the training distribution. However, both works still assume *i.i.d.* or exchangeable training data, making them not directly applicable for time series. A concurrent work [19] considers a general set-up for bounding coverage gap using total variation distances. It then proposes to use fixed weights to correct for the coverage gap. In retrospect, we consider a more specific setting involving time series, and the upper bounds are captured differently and explicitly using the quality of the estimator and the noise characteristics. Meanwhile, a recent work for non-exchangeable data sequentially adjusts the significance level α during prediction. For instance, [20] provides approximately valid coverage on sequential data by re-weighting the value α based on online coverage values on test data. The subsequent work [21] proposes more sophisticated re-weighting techniques of α. However, whether such adjustments are applicable to data with general dependency remain unclear, and we compare with [20] in experiments to show the improved performance of EnbPI.

Meanwhile, there are many non-CP prediction interval methods. In the traditional time series literature [22], there have been abundant work for prediction interval construction, such as ARIMA(p, d, q) [23], exponential smoothing [24], dynamic factor models [25] and so on. However, they rely on strong parametric distributional assumptions that limit their applicability. On the other hand, recent works have notably leveraged the predictive power of deep neural networks for neural quantile regression. Two of the most popular approaches are MQ-CNN [6] and DeepAR [5]; additional approaches can be found in [26]. More precisely, MQ-CNN [6] leverages the power of sequence-to-sequence neural networks to predict the multi-horizon quantile value of future response variables directly. The framework can also incorporate various temporal and static features and remains scalable to large-scale forecasting. Meanwhile, DeepAR [5] models the conditional distribution of future response using an autoregressive recurrent network. The network is trained by maximizing the log-likelihood of data, assuming Gaussian likelihood for real-valued data and negative-binomial for positive count data. Extensive experiments show its improvement over state-of-the-art methods. Although both MQ-CNN and DeepAR have promising performances for a variety of time-series data, they have limitations in requiring special network architecture (not model-free) and providing no theoretical guarantees on coverage. In addition, [5] imposes distributional assumptions on data through the parametric likelihood models (not distribution-free). In contrast, EnbPI leverages the benefits of conformal prediction to present a general framework for an arbitrary point-prediction model (model-free), with provable guarantees on coverage and without distributional assumption on data (distribution-free).

Finally, we remark that our assumptions and proof techniques avoid data exchangeability and differ significantly from existing CP works. Most CP methods ensure the finite-sample marginal coverage and distribution-free conditional coverage is impossible at a finite sample size [27]. In contrast, we achieve an asymptotic conditional coverage guarantee. Such theoretical analyses are inspired by [28], [29], yet we refine the proof techniques to improve the convergence rates and extend results under different assumptions. We further analyze the convergence of prediction interval widths. We would also like to remark that

our work is titled "conformal prediction" because `EnbPI` builds on the conformal prediction framework in this more general context—in terms of construction, `EnbPI` intervals closely resemble intervals by existing CP methods (especially J+aB [14]). Meanwhile, the theoretical results in this work can hold for prediction intervals produced by other conformal prediction methods, such as split conformal [11], J+aB [14], and so on (see Remark 1). Thus, the theoretical tools presented in this work are general for analyzing CP methods for time series.

II. Problem Setup

Given an unknown model $f : \mathbb{R}^d \to \mathbb{R}$, where d is the dimension of the feature vector, we observe data (x_t, y_t) generated according to the following model

$$\widehat{Y_t} = f(X_t) + \epsilon_t, \quad t = 1, 2, \dots \tag{1}$$

where ϵ_t is distributed following a continuous cumulative distribution function (CDF) F_t. Note that we do not need ϵ_t to be independent and F_t needs not be the same across all t. Features X_t can contain exogenous time series sequences that predict Y_t and/or the history of Y_t. We assume that the first T samples $\{(x_t, y_t)\}_{t=1}^T$ are training data or initial state of the random process that are observable. Above, upper case X_t, Y_t denote random variables and lower case x_t, y_t denote data.

Our goal is to construct a sequence of prediction intervals as narrow as possible with a certain coverage guarantee. Given a user-specified prediction algorithm, using T training samples, we obtain a trained model represented by $\hat{f}$. Then we construct $s \geq 1$ prediction intervals $\{\widehat{C}_{T+i}^{\alpha}\}_{i=1}^s$ for $\{Y_{T+i}\}_{i=1}^s$, where α is the *significance level*, and the *batch size* s is a pre-specified parameter for how many steps we want to look ahead. Once new samples $\{(x_{T+i}, y_{T+i})\}_{i=1}^s$ become available, we deploy the pre-trained $\hat{f}$ on new samples and use the most recent T samples to produce prediction intervals for $Y_j, j = T + s + 1$ onward without re-training the model on new data.

The meaning of significance level α is as follows. We consider two types of coverage guarantees. The *conditional* coverage guarantee ensures that each prediction interval $\widehat{C}_t^{\alpha}, t > T$ satisfies:

$$P(Y_t \in \widehat{C}_t^{\alpha} | X_t = x_t) \geq 1 - \alpha. \tag{2}$$

The second type is the *marginal* coverage guarantee:

$$P(Y_t \in \widehat{C}_t^{\alpha}) \geq 1 - \alpha. \tag{3}$$

Note that (2) is much stronger than (3), which is satisfied whenever data are exchangeable using split conformal prediction [11]. For instance, suppose a doctor reports a prediction interval for one patient's blood pressure. An interval satisfying (3) averages over all patients in different age groups, but may not satisfy (2) for the current patient precisely. In fact, satisfying (2), even for exchangeable data, is impossible without further assumptions [27]. In general, it is challenging to ensure either (2) or (3) under complex data dependency without distributional assumptions. Despite such difficulty, our theory provides a way to bound the worst-case gap in conditional coverage (2) and marginal coverage (3), under certain assumptions on the error process $\{\epsilon_t\}_{t\geq 1}$ and $\hat{f}$. From now on, we call a prediction interval conditionally or marginally *valid* if it achieves (2) or (3), respectively.

A. Oracle Prediction Interval

To motivate the construction of $\widehat{C}_t^{\alpha}$, we first consider the *oracle* prediction interval C_t^{α}, which contains Y_t with an exact conditional coverage at $1 - \alpha$ and is the shortest among all possible conditionally valid prediction intervals. The oracle prediction assumes perfect knowledge of f and F_t in (1). Denote $F_{t,Y}$ as the CDF of Y_t conditioning on $X_t = x_t$, then we have

$$\begin{aligned} F_{t,Y}(y) &= \mathbb{P}(Y_t \leq y | X_t = x_t) \\ &= \mathbb{P}(\epsilon_t \leq y - f(x_t)) = F_t(y - f(x_t)). \end{aligned}$$

For any $\beta \in [0, \alpha]$, we also know that

$$\mathbb{P}(Y_t \in [F_{t,Y}^{-1}(\beta), F_{t,Y}^{-1}(1 - \alpha + \beta)] | X_t = x_t) = 1 - \alpha,$$

where $F_{t,Y}^{-1}(\beta) := \inf\{y : F_{t,Y}(y) \geq \beta\}$. Assume $F_{t,Y}^{-1}(\alpha)$ is attained for each $\alpha \in [0, 1]$, and let $y_\beta = F_{t,Y}^{-1}(\beta)$. Clearly,

$$y_\beta = f(x_t) + F_t^{-1}(\beta),$$

which allows us to find C_t^{α} – the oracle prediction interval with the narrowest width:

$$\begin{aligned} C_t^{\alpha} &= [f(x_t) + F_t^{-1}(\beta^*), f(x_t) + F_t^{-1}(1 - \alpha + \beta^*)], \\ \beta^* &:= \underset{\beta \in [0,\alpha]}{\arg\min} \left(F_t^{-1}(1 - \alpha + \beta) - F_t^{-1}(\beta) \right). \end{aligned} \tag{4}$$

A similar oracle construction to (4) appeared in [30]. Thus, if we can approximate unknown $f(x_t), F_t^{-1}(x), x \in [0, 1]$, and β^* reasonably well, the prediction intervals $\widehat{C}_t^{\alpha}$ should be close to the oracle C_t^{α}.

B. Proposed Prediction Interval

We now construct $\widehat{C}_t^{\alpha}$ based on ideas above. Recall that the first T data $\{(x_t, y_t)\}_{t=1}^T$ are observable. Denote $\hat{f}_{-i}$ as the i-th "leave-one-out" (LOO) estimator of f, which is not trained on the i-th datum (x_i, y_i) and may include the remaining $T - 1$ points. Then,

$$\begin{aligned} \widehat{C}_t^{\alpha} &:= [\hat{f}_{-t}(x_t) + \hat{\beta} \text{ quantile of } \{\hat{\epsilon}_i\}_{i=t-1}^{t-T}, \\ &\quad \hat{f}_{-t}(x_t) + (1 - \alpha + \hat{\beta}) \text{ quantile of } \{\hat{\epsilon}_i\}_{i=t-1}^{t-T}], \end{aligned} \tag{5}$$

where the LOO prediction residual $\hat{\epsilon}_i$ and the corresponding $\hat{\beta}$ are defined as

$$\begin{aligned} \hat{\epsilon}_i &:= y_i - \hat{f}_{-i}(x_i) \\ \hat{\beta} &:= \underset{\beta \in [0,\alpha]}{\arg\min} \left((1 - \alpha + \beta) \text{ quantile of } \{\hat{\epsilon}_i\}_{i=t-1}^{t-T} \right. \\ &\qquad \left. - \beta \text{ quantile of } \{\hat{\epsilon}_i\}_{i=t-1}^{t-T} \right). \end{aligned}$$

Thus, the interval centers at the point prediction $\hat{f}_{-t}(x_t)$ and the width is the difference between the $(1 - \alpha + \hat{\beta})$ and $\hat{\beta}$ quantiles over the past T residuals.

Note that we have to split the training data into two parts: one part is used to estimate f, and the second part is used to obtain prediction residuals for the prediction interval. There is a trade-off. On the one hand, we desire the estimator $\hat{f}$ to be trained on as much data as possible. On the other hand, the quantile of prediction residuals should well approximate the tails of F_t^{-1}. These two objectives contradict each other. If we train $\hat{f}$ on all training data, then we overfit; if we train on a subset of training data and obtain prediction residuals on the rest [11], the approximation $\hat{f}$ to f is poorer. The LOO estimator is known to achieve a good trade-off in this regard. When obtaining the i-th residual, the i-th LOO estimator trains on all except the i-th training datum so that the LOO estimator is not overfitted on that datum. Then repeating over T training data yields T LOO estimators with good predictive power and T residuals to calibrate the prediction intervals well. The LOO idea is related to the Jackknife+ procedure [13], but it is known to be costly due to the retraining of the model. To address this issue, we will develop a computationally efficient method called `EnbPI` in Section IV, which constructs the LOO estimators as *ensemble* estimators of pre-trained models.

III. Theoretical Analysis

We first present theoretical results for bounding the worst-case coverage gap in conditional and marginal coverage. We then establish similar bounds on the difference between estimated and oracle intervals. The results are general for methods beyond `EnbPI` (for example, the split conformal method [11]). Without loss of generality and for notation simplicity, we only show guarantees when $t = T + 1$, i.e., the one-step-ahead prediction. We will explain how guarantees naturally extend to all prediction intervals from $t = T + 2$ onward in Remark 1. In particular, our proof removes the assumptions on data exchangeability by replacing them with general and verifiable assumptions on the error process and estimation quality. All proofs can be found in Appendix A, available in the online supplemental material.

A. Coverage Guarantees

Following notations in Section II-A, we first define the empirical p-value at $T + 1$:

$$\hat{p}_{T+1} := \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{\hat{\epsilon}_i \le \hat{\epsilon}_{T+1}\}.$$

As a result, we see the following equivalence between events:

$$\begin{aligned} & Y_{T+1} \in \widehat{C}^{\alpha}_{T+1} \Big| X_{T+1} = x_{T+1} \\ \iff & \hat{\epsilon}_{T+1} \in [\hat{\beta} \text{ quantile of } \{\hat{\epsilon}_i\}_{i=1}^{T}, \\ & (1-\alpha+\hat{\beta}) \text{ quantile of } \{\hat{\epsilon}_i\}_{i=1}^{T}] \Big| X_{T+1} = x_{T+1} \\ \iff & \hat{\beta} \le \hat{p}_{T+1} \le 1 - \alpha + \hat{\beta}, \end{aligned}$$

where $A|B$ means that the event A conditions on event B. Therefore, our method covers Y_{T+1} given $X_{T+1} = x_{T+1}$ with probability $1 - \alpha$, hence, being conditionally valid if the distribution of $\hat{p}_{T+1}$ is uniform. More precisely, we aim to ensure that $|\mathbb{P}(\beta \le \hat{p}_{T+1} \le 1 - \alpha + \beta) - (1 - \alpha)|$ is small for any $\beta \in [0, \alpha]$.

Due to the fact that $F_{T+1}(\epsilon_{T+1}) \sim \text{Unif}[0, 1]$ [31], $\mathbb{P}(\beta \le F_{T+1}(\epsilon_{T+1}) \le 1 - \alpha + \beta) = 1 - \alpha$. Define

$$\hat{F}_{T+1}(x) := \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{\hat{\epsilon}_i \le x\},$$

whereby we have $\hat{p}_{T+1} = \hat{F}_{T+1}(\hat{\epsilon}_{T+1})$. As a consequence:

$$\begin{aligned} & |\mathbb{P}(\beta \le \hat{p}_{T+1} \le 1 - \alpha + \beta) - (1 - \alpha)| \\ & \quad = |\mathbb{P}(\beta \le \hat{F}_{T+1}(\hat{\epsilon}_{T+1}) \le 1 - \alpha + \beta) \\ & \qquad - \mathbb{P}(\beta \le F_{T+1}(\epsilon_{T+1}) \le 1 - \alpha + \beta)|. \end{aligned}$$

Thus, intuitively, we can bound gap in conditional coverage using the worst-case difference between $\hat{F}_{T+1}(\hat{\epsilon}_{T+1})$ and $F_{T+1}(\epsilon_{T+1})$. Notice the following coupling between $\hat{\epsilon}_{T+1}$ and ϵ_{T+1} under model (1) when $X_{T+1} = x_{T+1}$:

$$\hat{\epsilon}_{T+1} = \epsilon_{T+1} + (f(x_{T+1}) - \hat{f}_{-(T+1)}(x_{T+1})). \tag{6}$$

Therefore, the pointwise function estimation error $f(x_{T+1}) - \hat{f}_{-(T+1)}(x_{T+1})$ should be small for $\hat{\epsilon}_{T+1}$ to be a good estimate for ϵ_{T+1}. We will impose this condition when analyzing difference in interval width.

For the analyses, we now introduce another empirical CDF using unknown "true" errors $\epsilon_i, i \ge 1$, denoted as $\tilde{F}_{T+1}$:

$$\tilde{F}_{T+1}(x) := \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{\epsilon_i \le x\}.$$

Note that $\hat{F}_{T+1}(\hat{\epsilon}_{T+1})$ is close in distribution to $\tilde{F}_{T+1}(\epsilon_{T+1})$ under the same pointwise estimation assumption of f by $\hat{f}$, due to (6). Meanwhile, the convergence of $\tilde{F}_{T+1}(x)$ to $F_{T+1}(x)$ is well-studied in the literature, which addresses the rate of convergence of an empirical distribution to the actual CDF [32], [33], [34]. Building on notations and ideas above, we now state the precise assumptions with discussions and present the following results: we first bound the worst deviation between $\tilde{F}_{T+1}(x)$ and $F_{T+1}(x)$ in Lemma 1. We then bound that between $\hat{F}_{T+1}(x)$ and $\tilde{F}_{T+1}(x)$ in Lemma 2. These lemmas are essential to proving our main theoretical results in Theorem 1, which has several useful corollaries under slightly modified assumptions on error dependencies.

Assumption 1 (Errors are short-term i.i.d.): Assume $\{\epsilon_t\}_{t=1}^{T+1}$ are independent and identically distributed (i.i.d.) according to a common CDF F_{T+1}, which is Lipschitz continuous with constant $L_{T+1} > 0$.

Lemma 1: Under Assumption 1, for any training size T, there is an event A_T which occurs with probability at least $1 - \sqrt{\log(16\,T)/T}$, such that conditioning on A_T,

$$\sup_x |\tilde{F}_{T+1}(x) - F_{T+1}(x)| \le \sqrt{\log(16\,T)/T}.$$

Discussion on Assumption 1: We call it the short-term i.i.d. assumption, since it only requires the past $T + 1$ errors to be

independent. It is a reasonably mild assumption on the original process $\{(X_t, Y_t)\}_{t\geq 1}$, because the process can exhibit arbitrary dependence and be highly non-stationary but still have i.i.d. errors. Later on we can relax this assumption for more general cases, for instance, when errors follow linear processes (see Corollary 1) or are strongly mixing (see Corollary 2). We can empirically examine whether or not the assumptions on residuals hold by using the LOO residuals as surrogates. The procedure is similar to examining the autocorrelation function after fitting a time series model.

Assumption 2 (Estimation quality): There exists a real sequence $\{\delta_T\}_{T\geq 1}$ such that

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{f}_{-t}(x_t) - f(x_t))^2 \leq \delta_T^2 \text{ and}$$

$$|\hat{f}_{-(T+1)}(x_{T+1}) - f(x_{T+1})| \leq \delta_T.$$

Lemma 2: Under Assumptions 1 and 2, we have

$$\sup_x |\hat{F}_{T+1}(x) - \tilde{F}_{T+1}(x)|$$

$$\leq (L_{T+1}+1)\delta_T^{2/3} + 2\sup_x |\tilde{F}_{T+1}(x) - F_{T+1}(x)|.$$

Discussion on Assumption 2: There are two situations affecting asymptotic guarantees: δ_T never decays as T grows or converges to zero as $T \to \infty$. The first situation can happen due to data overfitting, which leads to $\hat{f}_{-t}(x_t) \approx y_t$ and therefore, $\sum_{t=1}^{T}(\hat{f}_{-t}(x_t) - f(x_t))^2 \approx \sum_{t=1}^{T}\epsilon_t^2$. If $\sum_{t=1}^{T}\epsilon_t^2 \in \Omega(T)$, the same order holds for the sequence $\{\delta_T\}_{T\geq 1}$, so that the worst-case coverage gap always exists (see Theorem 1). On the other hand, there are examples in the second situation where $\{\delta_T\}_{T\geq 1}$ converges to zero. Note that assumptions for estimating unknown f are necessary due to the well-known *No Free Lunch Theorem* [35]. The decay rate of δ_T is explicit for two classes of f and the following $\mathcal{A}$:

(Example 1) If f is sufficiently smooth, $\delta_T = o_P(T^{-1/4})$ for general neural networks sieve estimators [36, see Corollary 3.2].

(Example 2) If f is a sparse high-dimensional linear model, $\delta_T = o_P(T^{-1/2})$ for the Lasso estimator and Dantzig selector. [37, see 7.7].

In general, one needs to analyze the convergence rate of estimators $\hat{f}$ to the unknown true f. This task is different from analyzing the Mean Squared Error (MSE) of ensemble estimators [9] and likely requires case-by-case analyses, which we leave for future work.

Our main theoretical result is the following Theorem 1, which establishes the asymptotic conditional coverage as a consequence of Lemmas 1 and 2.

Theorem 1 (Conditional coverage gap; errors are short-term i.i.d.): Under Assumption 1 and 2, for any training size T, $\alpha \in (0,1)$, and $\beta \in [0, \alpha]$, we have:

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} | X_{T+1} = x_{T+1}) - (1-\alpha)|$$

$$\leq 12\sqrt{\log(16\,T)/T} + 4(L_{T+1}+1)(\delta_T^{2/3} + \delta_T). \quad (7)$$

Furthermore, if $\{\delta_T\}_{T\geq 1}$ converges to zero, the upper bound in (7) converges to 0 when $T \to \infty$, and thus the conditional coverage is asymptotically valid.

We briefly comment on the proof techniques and the role of Assumption 1. The term $\sqrt{\log(16\,T)/T}$ on the right-hand side directly relates to how quickly the empirical CDF $\tilde{F}_{T+1}$ converges to the actual CDF F_{T+1}. In general, we find sequences $\{s_T\}_{T\geq 1}$ and $\{g(s_T)\}_{T\geq 1}$, both of which converge to zero, such that

$$\mathbb{P}(\sup_x |\tilde{F}_{T+1}(x) - F_{T+1}(x)| > s_T) \leq g(s_T).$$

The optimal rate of decay reduces to finding s_T such that $s_T = g(s_T)$. Then, the event A_T is chosen to happen with probability at least $1 - s_T$, where conditioning on this event, $\sup_x |\tilde{F}_{T+1}(x) - F_{T+1}(x)| \leq s_T$. As a result, there are decay rates different from $\sqrt{\log(16\,T)/T}$ under more relaxed assumptions on $\{\epsilon_t\}_{t=1}^{T+1}$. We summarize two possible results in Corollaries 1 and 2; certain technical assumptions, precise statements, and definitions are presented in the appendix, available in the online supplemental material.

Corollary 1 (Conditional coverage gap; errors follow linear processes): Under Assumption 2, suppose that $\{\epsilon_t\}_{t=1}^{T+1}$ satisfy $\epsilon_t = \sum_{j=1}^{\infty}\delta_j z_{t-j}$, with regularity conditions on δ_j and z_{t-j}. There exists a constant K so that for any training size T, $\alpha \in (0,1)$, and $\beta \in [0, \alpha]$, we have:

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} | X_{T+1} = x_{T+1}) - (1-\alpha)|$$

$$\leq 12\,K\log T/\sqrt{T} + 4(L_{T+1}+1)(\delta_T^{2/3} + \delta_T). \quad (8)$$

To introduce the last corollary, we first define the strong mixing coefficient between two $\sigma-$fields $\mathcal{A}$ and $\mathcal{B}$, which measures the dependence between them:

$$\alpha(\mathcal{A}, \mathcal{B}) = 2\sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| : (A, B) \in \mathcal{A}\times\mathcal{B}\}.$$

This definition is equivalent to that in [38] up to a multiplicative factor of 2. For the sequence $\{\epsilon_t\}_{t\geq 1}$, let $\mathcal{A}_k := \sigma(\epsilon_t : t \leq k)$ and $\mathcal{B}_c := \sigma(\epsilon_t : t \geq l)$. The coefficients $\{\alpha_n\}_{n\geq 1}$ are defined as

$$\alpha_0 = 1/2 \text{ and } \alpha_n = \sup_{k\in\mathbb{N}} \alpha(\mathcal{A}_k, \mathcal{B}_{k+n}) \text{ for any } n > 0.$$

The sequence is said to be *strongly mixing* if $\lim_{n\to\infty}\alpha_n = 0$.

Corollary 2 (Conditional coverage gap; errors are strongly mixing): Under Assumption 2, suppose $\{\epsilon_t\}_{t=1}^{T+1}$ are stationary and strongly mixing, where mixing coefficients are summable with $0 < \sum_{k\geq 0}\alpha_k < M$. For any training size T, $\alpha \in (0,1)$, and $\beta \in [0, \alpha]$, we have:

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}_{T+1}^{\alpha} | X_{T+1} = x_{T+1}) - (1-\alpha)|$$

$$\leq 12(M/2)^{1/3}(\log T)^{2/3}/T^{1/3} + 4(L_{T+1}+1)(\delta_T^{2/3} + \delta_T). \quad (9)$$

Lastly, the following asymptotic marginal validity guarantee holds as a consequence of earlier results by the tower law property (proof omitted):

Theorem 2 (Marginal coverage gap): Under Assumption 1 and 2, for any training size T, $\alpha \in (0,1)$, and $\beta \in [0, \alpha]$, we

have:

$$|\mathbb{P}(Y_{T+1} \in \widehat{C}^{\alpha}_{T+1}) - (1-\alpha)| \leq 12\sqrt{\log(16\,T)/T} + 4(L_{T+1}+1)(\delta_T^{2/3} + \delta_T). \quad (10)$$

Moreover, the right-hand side decay rate in (10) is $\mathcal{O}(\log T/\sqrt{T} + \delta_T^{2/3})$ if $\{\epsilon_t\}$ follow a linear process as in Corollary 1, and $\mathcal{O}((\log T)^{2/3}/T^{1/3} + \delta_T^{2/3})$ if $\{\epsilon_t\}$ are strongly mixing with summable mixing coefficients as in Corollary 2.

We make two final comments for the above theorems and corollaries. First, to build prediction intervals that have at least $1-\alpha$ coverage, one needs to incorporate the upper bounds on the right-hand side of (7)—(10) into the prediction interval construction. However, we will not do so in EnbPI (our proposed algorithm), which is a *general wrapper* that can be applied to most regression models $\mathcal{A}$. Second, The rate $\mathcal{O}(\sqrt{\log(16\,T)/T} + \delta_T^{2/3})$ is a worst-case analysis for both marginal and conditional coverage; empirical results show that even at a small training data size T, EnbPI can achieve both marginal and conditional validity.

B. Width Guarantees

Our next goal is to bound the gap between the estimated prediction interval $\widehat{C}^{\alpha}_{T+1}$ and the oracle C^{α}_{T+1} in (4). Define set difference $\Delta : \mathbb{N} \to \mathbb{R}$ such that $\Delta(T) = \widehat{C}^{\alpha}_{T+1} \triangle C^{\alpha}_{T+1}$, where for any two subsets $A, B \subset \mathbb{R}$ under the Lebesgue measure μ, $A \triangle B := \mu(\{x \in \mathbb{R} : x \in A, x \notin B\}) + \mu(\{x \in \mathbb{R} : x \in B, x \notin A\})$. Theorem 3 below bounds $\Delta(T)$ under Assumptions 1, 2, and other regularity conditions; the bound is similar to that in Theorem 1.

Theorem 3 (Width gap bound; errors are i.i.d.): Under Assumption 1 and 2, further assume F^{-1}_{T+1} is Lipschitz continuous with constant K_{T+1}. With probability at least $1 - \sqrt{\log(16\,T)/T}$,

$$\Delta(T) \leq \delta_T + \alpha K'_{T+1}/m + 2(K_{T+1} + M_{T+1}) \cdot \times \left(3\sqrt{\log(16\,T)/T} + (L_{T+1}+1)(\delta_T^{2/3} + \delta_T)\right),$$

where m is the number of grids for line-search of $\hat{\beta}$ based on the past T LOO residuals, $K'_{T+1} := \max_{j=1,\ldots,T-1} \hat{\epsilon}_{j+1} - \hat{\epsilon}_j$ using sorted LOO residuals indexed from the smallest to the largest, and M_{T+1} is a constant that depends only on L_{T+1}, K_{T+1}, and K'_{T+1}.

When $\{\epsilon_t\}_{t=1}^{T}$ are not i.i.d., results similar to Corollaries 1 and 2 can be established for Theorem 3 using similar proof techniques. More precisely, the rate $\sqrt{\log(16\,T)/T}$ will be replaced by $\log T/\sqrt{T}$ when errors follow linear processes, and by $(\log T)^{2/3}/T^{1/3}$ when errors are strongly mixing with summable mixing coefficients.

Remark 1 (Theorem applicability and caveats): All theoretical results hold for $t > T+1$, as long as Assumptions 1 and 2 hold at indices $t-T, \ldots, t$. The same proof techniques apply. Meanwhile, as long as the same assumptions hold, all previous results apply to other conformal prediction methods, such as split conformal [11]. However, unlike our EnbPI that requires no data-splitting, split conformal and its variants require data splitting by treating a subset of training data as the "calibration data." As a result, the value T on the right-hand side of Theorem 1 and all subsequent corollaries become the size of the calibration data, not that of the full training data. This is because prediction residuals $\hat{\epsilon}$ are only computed on calibration data, whose empirical distribution is used to approximate that of the true distribution of errors ϵ. In such cases, the worst-case coverage gap becomes larger.

IV. EnbPI Algorithm

We now present a general conformal prediction algorithm for time series in Algorithm 1, which is named EnbPI. On a high-level, EnbPI has a training phase and the prediction phase. In the training phase, EnbPI first fits a fixed number of bootstrap estimators from subsets of the training data. Then, it aggregates predictions from these bootstrap estimators on the training data in an efficient leave-one-out (LOO) fashion, resulting in *both* LOO predictors and LOO residuals for prediction. In the prediction phase, EnbPI aggregates predictions from LOO predictors on each test datum to compute the center of the prediction interval. Then, it builds the prediction interval using the past LOO residuals, where the interval width is also optimized through a simple one-dimensional line search. Lastly, residuals are slid forward as soon as actual response variables in test data are observed to ensure adaptiveness in the prediction intervals.

In the algorithm description, $\hat{f}^b$ is the b-th bootstrap estimator, the superscript ϕ denotes variables with dependence on the aggregation function ϕ. The *block bootstrap* with T non-overlapping blocks is used in line 2, which is a popular method for bootstrapping dependent data [39]. The basic idea is to split the T training samples into l (non-)overlapping blocks, each with a size $\lfloor T/l \rfloor$. Then, sample from l blocks randomly with replacement.

We comment on the choice of hyperparameters as follows. (1) In general, $\mathcal{A}$ can be a family of (parametric and non-parametric) prediction algorithms. (2) Different choices of aggregation functions ϕ bring different benefits, such as reducing the MSE under mean, avoiding sensitivity to outliers under median, or achieving both under trimmed mean. (3) As the number of pre-trained bootstrap models B increases, interval widths may be narrower. Empirically, we find that choosing B between 20 and 50 is sufficient, especially for computationally intensive methods such as neural networks. (4) Larger s requires prediction further in the future without feedback; however, as s increases, the prediction becomes harder, which is reflected in that intervals become wider and the coverage deteriorates; how large s can be is determined by the dynamics of the data.

A. Properties of EnbPI

Computational Efficiency: Note that in EnbPI, the prediction models in the ensemble are pre-trained once and stored; when deploying EnbPI for prediction, residuals are computed from T pre-trained models on the fly, and the interval is constructed

Algorithm 1: Ensemble batch prediction intervals (EnbPI).

Require: Training data $\{(x_i, y_i)\}_{i=1}^{T}$, prediction algorithm $\mathcal{A}$, significance level α, aggregation function ϕ, number of bootstrap models B, batch size s, and test data $\{(x_t, y_t)\}_{t=T+1}^{T+T_1}$; y_t is revealed as feedback only after prediction at t is done.

Ensure: Ensemble prediction intervals $\{C_t^{\phi,\alpha}(x_t)\}_{t=T+1}^{T+T_1}$

1: **for** $b = 1, \ldots, B$ **do**
2: Sample with replacement an index set $S_b = (i_1, \ldots, i_T)$ from indices $(1, \ldots, T)$.
3: Compute $\hat{f}^b = \mathcal{A}((x_i, y_i), i \in S_b)$.
4: **end for**
5: Initialize $\widehat{\boldsymbol{\epsilon}} = \{\}$ as an ordered set.
6: **for** $i = 1, \ldots, T$ **do**
7: $\hat{f}^{\phi}_{-i}(x_i) = \phi(\hat{f}^b(x_i), i \notin S_b)$
8: Compute $\hat{\epsilon}_i^{\phi} = y_i - \hat{f}^{\phi}_{-i}(x_i)$
9: $\widehat{\boldsymbol{\epsilon}} = \widehat{\boldsymbol{\epsilon}} \cup \{\hat{\epsilon}_i^{\phi}\}$
10: **end for**
11: **for** $t = T+1, \ldots, T+T_1$ **do**
12: $\hat{f}^{\phi}_{-t}(x_t) = \phi(\hat{f}^{\phi}_{-i}(x_t), i = 1, \ldots, T)$
13: Compute $\hat{\beta}$ as

$$\arg\min_{\beta \in [0,\alpha]} (1-\alpha+\beta) \text{ quantile of } \widehat{\boldsymbol{\epsilon}} - \beta \text{ quantile of } \widehat{\boldsymbol{\epsilon}})$$

14: $w^{\phi,\alpha}_{t,\text{lower}} = \hat{\beta}$ quantile of $\widehat{\boldsymbol{\epsilon}}$
15: $w^{\phi,\alpha}_{t,\text{upper}} = (1 - \alpha + \hat{\beta})$ quantile of $\widehat{\boldsymbol{\epsilon}}$.
16: Return $C_t^{\phi,\alpha}(x_t) = [\hat{f}^{\phi}_{-t}(x_t) + w^{\phi,\alpha}_{t,\text{lower}}, \hat{f}^{\phi}_{-t}(x_t) + w^{\phi,\alpha}_{t,\text{upper}}]$
17: **if** $t - T \equiv 0 \bmod s$ **then**
18: **for** $j = t-s, \ldots, t-1$ **do**
19: Compute $\hat{\epsilon}_j^{\phi} = y_j - \hat{f}^{\phi}_{-j}(x_t)$
20: $\widehat{\boldsymbol{\epsilon}} = (\widehat{\boldsymbol{\epsilon}} - \{\hat{\epsilon}_1^{\phi}\}) \cup \{\hat{\epsilon}_j^{\phi}\}$ and reset index of $\widehat{\boldsymbol{\epsilon}}$.
21: **end for**
22: **end if**
23: **end for**

based on quantile values of T residuals. Thus, the main computation of EnbPI for obtaining the prediction interval is tolerable in calling the prediction algorithm $\mathcal{A}$ B times. In comparison, the Jackknife+ approach [13] requires requires B times training of $\mathcal{A}$ on *each* leave-i-out sample $\{(x_j, y_j)\}_{j=1, j\neq i}^{T}$. This requires BT training of $\mathcal{A}$, which can be computationally intensive for complex prediction algorithms such as deep neural networks.

No Overfitting or Data Splitting: Traditional CP methods such as split conformal [11] use data-splitting to avoid overfitting. In contrast, inspired by the J+aB procedure in [14], EnbPI trains LOO ensemble models on full data and avoids overfitting through thoughtful aggregations in lines 6-10. In particular, to construct the i-th LOO ensemble predictor, EnbPI aggregates all B bootstrap models that are *not* trained on the training datum (x_i, y_i). Thus, the actual number of aggregated models is a Binomial random variable with parameters B and $(1 - 1/T)^T$; the Chernoff bound ensures that each ensemble predictor aggregates a balanced number of pre-trained models.

Leverage New Data Without Model Retraining: EnbPI constructs sequential prediction intervals without retraining $\mathcal{A}$. Instead, it leverages feedback by updating past residuals through a sliding window of size T, which adapts the interval widths to data and can better adapt to data non-stationarity. In practice, we acknowledge the benefits of retraining, especially in reducing the widths of the prediction intervals. However, retraining can be costly for certain models, and one should consider the trade-off between interval widths and computation involved in retraining.

B. EnbPI *on Challenging Tasks*

We comment that EnbPI is flexible and can handle various challenging tasks. In Appendix B.4, available in the online supplemental material, we also discuss how EnbPI can construct prediction intervals for outputs from each node of a network.

Handle missing data: We suggest a heuristic approach to handle missing data by EnbPI, which is verified in Section V-C. When training and/or test data have missing entries, we can properly increase the size of bootstrap samples being drawn from the rest available training data— this is appropriate since a common data model f is assumed. On test data, when EnbPI encounters a missing index t', we impute the feature $x_{t'}$ if it is missing to compute $\hat{f}_t(x_{t'})$, the interval center, and use the most recent T residuals to compute the interval width. The sliding window would skip over the residual $\epsilon_{t'}^{\phi}$ when $y_{t'}$ is unobserved. Section V-C considers the solar dataset with missing data.

Unsupervised Anomaly Detection: Suppose there is an anomalous point y_{t^*} at time t^*, due to either a change in model f at t^* or an unusually large stochastic error ϵ_{t^*}. As a result, y_{t^*} tends to lie far outside the interval (equivalently, $\epsilon_{t^*}^{\phi}$ is well below or above the $\hat{\beta}$ or $(1 - \alpha + \hat{\beta})$ quantile of past T residuals) and thus can be detected using the prediction interval. An example applying EnbPI to detect anomalous traffic flows appears in Section V-D.

V. Experimental Results

The experiments are organized as follows. In Section V-A, we provide extensive simulations to examine the coverage and width of EnbPI intervals. In Section V-B, we show that EnbPI attains valid marginal coverage on real data, whereas competing methods may fail. In Section V-C, we present real-data experiments to examine the conditional coverage of EnbPI against other methods when missing data are present. In Section V-D, we present an example for anomaly detection in traffic flow using EnbPI. In Appendix B.4 and B.5, available in the online supplemental material, we present more time-series data examples to demonstrate that EnbPI has valid coverage and shorter intervals than the competing methods.

A. Simulation Results

We first conduct three simulated examples based on the assumption $Y_t = f(X_t) + \epsilon_t$ to examine the performance of EnbPI. We then consider a more complex example based on a noisy helix trajectory.

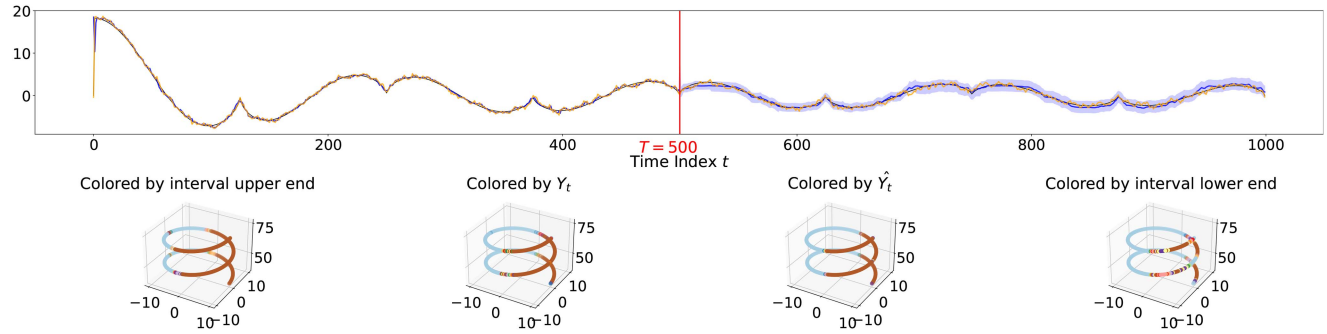

Fig. 1. Helix colored by Y_t. We observe that the predicted colors closely match the actual color on the bottom row because values of Y_t colored in orange are contained in prediction intervals colored in shaded blue with high probability, and intervals are very narrow, as shown on the top row.

Three simulated examples: We construct these examples with increasing levels of model sophistication in the design of $f(X_t)$ and under more complex error dependency in ϵ_t. The detailed data-generating procedures and additional details are described in Appendix B.1, available in the online supplemental material. The results shown in Fig. 5 of Appendix B.1, available in the online supplemental material indicate the satisfactory performance of `EnbPI` to maintain valid coverage. The interval widths also converge to the oracle width as the training sample size grows, validating Theorem 3.

Simulation With a Noisy Helix Trajectory: Consider Y_t given by a nonlinear mapping of components of a helix in three-dimensional space contaminated by noise: $X_t = [r\cos(\theta_t), r\sin(\theta_t), H\theta_t]$, $f(X_t) = r\cos(\theta_t) \cdot (|r\sin(\theta_t)|)^{1/2} \cdot (H\theta_t + \varepsilon)^{-1/2}$, $\varepsilon = 10^{-3}$, and $\epsilon_t = \rho\epsilon_{t-1} + e_t$ where $\rho = 0.6$ and e_t are *i.i.d.* normal random variables with zero mean and unit variance. The color map of the helix is proportional to Y_t. We fix $H = 3$, $r = 10$ and generate 1000 samples parametrized by θ_t, which are uniformly spaced between 0 and 8π. The first 500 data points are used for training `EnbPI` with random forest regression (RF) and the rest 500 are used for testing. The RF setup is described in Appendix B.2, available in the online supplemental material. In Fig. 1, we see that in the test phase intervals by `EnbPI` tightly cover the unknown response Y_t. Moreover, the blue and orange curves corresponding to $\hat{Y}_t$ and Y_t are very close, which indicates that LOO ensemble predictors approximate the unknown model f very well.

B. Real-Data: Marginal Validity and the Interval Width

In this section, we consider predictions for renewable energy generation. In this setting, the prediction and uncertainty quantification is critical due to their high stochasticity and non-stationarity.

Data Description: The renewable energy data are from the National Solar Radiation Database and the Hackberry wind farm in Austin.[1] We use 2018 hourly solar radiation data from Atlanta and nine cities in California and 2019 hourly wind energy data. We remove recordings before 6 a.m. and after 8 p.m. for the solar radiation data due to zero radiation levels during the period. In total, there are 11 time series from 11 sensors (one from each sensor), and each time series contain other features such as temperature, humidity, wind speed, etc. In particular, California solar data constitute a network, where each node is a sensor. From now on, we call X_t *univariate* if it is the history of Y_t and *multivariate* if it contains other features that predict Y_t.

Comparison Methods: We compare `EnbPI`with traditional time series and other conformal prediction methods. The time series methods are ARIMA(10,1,10), Exponential Smoothing (ExpSmoothing), and Dynamic Factor model (DynamicFactor). The CP methods are split/inductive conformal predictor (ICP) [11] and, weighted ICP (WeightedICP) [17], quantile out-of-bag method (QOOB) [15], adaptive conformal inference (AdaptCI) [20], and jackknife+-after-bootstrap (J+aB) [14]. For the former two CP methods (resp. AdaptCI), we split the training data into 50% (resp. 75%) proper training set for training a predictor and 50% (resp. 25%) calibration set for computing non-conformity scores. Appendix B.2, available in the online supplemental material describes more detailed setup.

Prediction Algorithm $\mathcal{A}$: We choose four prediction algorithms: ridge regression, random forest (RF), neural networks (NN), and recurrent neural networks (RNN) with LSTM layers. The first two are implemented in the Python `sklearn` library, and the last two are built using the `keras` library. See Appendix B.2, available in the online supplemental material for their specifications.

Other Hyperparameters: Since the three CP methods are trained on random subsets of training data, we repeat all experiments below for ten trials with an independent random split in each trial. The time series methods are only applied once on training data because they do not use random subsets. Throughout this subsection, we fix $s = 1$. Let $\alpha = 0.1$ and use the first 20% of the total hourly data for training unless otherwise specified. This creates small training samples for a challenging long-term predictive inference task. We use `EnbPI` under $B = 25$ and ϕ as taking the sample mean.

Results: All results in Section V-B and V-C come from using the Atlanta solar data. Similar results using California solar data and Hackberry wind data are in Appendix B.4, available in the online supplemental material. We first compare `EnbPI` with the conformal prediction methods at a fixed $\alpha = 0.1$. `EnbPI` results

[1]NSRDB: https://nsrdb.nrel.gov/. Wind farm: https://github.com/Duvey314/austin-green-energy-predictor

TABLE I
SOLAR POWER PREDICTION IN ATLANTA, COMPARISON OF EnbPI WITH ADAPTCI, J+AB, QOOB, ICP, AND WEIGHTED ICP

Train ratio	0.10						0.19						0.28					
CP method	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP
Coverage	0.893 (1.8e-03)	0.828 (2.4e-02)	0.747 (2.7e-03)	0.684 (1.1e-02)	0.646 (1.2e-01)	0.608 (1.4e-01)	0.897 (5.9e-04)	0.891 (6.1e-03)	0.777 (3.0e-03)	0.783 (2.8e-03)	0.703 (1.2e-01)	0.698 (1.2e-01)	0.905 (7.0e-04)	0.909 (1.5e-03)	0.819 (1.9e-03)	0.850 (3.5e-03)	0.760 (1.1e-01)	0.746 (1.3e-01)
Width	204.597 (1.8e+00)	178.870 (1.7e+01)	116.129 (1.3e+00)	106.199 (2.1e+00)	104.745 (4.0e+01)	96.728 (4.2e+01)	215.442 (4.4e-01)	222.328 (2.2e+00)	148.174 (1.6e+00)	140.723 (1.9e+00)	132.888 (4.8e+01)	131.247 (4.9e+01)	227.286 (6.8e-01)	211.686 (2.6e+00)	180.081 (7.0e-01)	160.231 (1.4e+00)	165.545 (6.5e+01)	163.855 (6.8e+01)

We vary the percentage of total data as training data at $\alpha = 0.1$. Cells in brackets for CP methods indicate standard deviation over ten trials.

TABLE II
SOLAR POWER PREDICTION IN ATLANTA

Train ratio	0.10						0.19						0.28					
CP method	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP	EnbPI	AdaptCI	J+aB	QOOB	ICP	Weighted ICP
α value	0.1	0.05	0.0115	0.0075	0.018	0.0125	0.1	0.13	0.04	0.025	0.04	0.04	0.1	0.125	0.055	0.03	0.07	0.06
Coverage	0.893 (1.8e-3)	0.855 (4.5e-3)	0.844 (2.5e-3)	0.840 (9.8e-3)	0.848 (1.4e-2)	0.828 (3.4e-2)	0.897 (5.9e-4)	0.869 (1.0e-3)	0.850 (2.4e-3)	0.876 (2.4e-3)	0.843 (1.8e-2)	0.844 (1.2e-2)	0.905 (7.0e-4)	0.883 (6.1e-4)	0.879 (2.3e-3)	0.931 (1.7e-3)	0.859 (7.4e-3)	0.869 (7.1e-3)
Width	204.597 (1.8e+0)	210.124 (4.7e+0)	210.747 (2.7e+0)	203.400 (1.0e+1)	214.308 (1.3e+1)	203.511 (3.2e+1)	215.442 (4.4e-1)	215.896 (1.1e+0)	214.418 (1.9e+0)	212.158 (2.0e+0)	218.597 (1.2e+1)	217.359 (8.7e+0)	227.286 (6.8e-1)	224.981 (9.7e-1)	224.418 (1.7e+0)	223.890 (1.6e+0)	220.091 (6.1e+0)	225.935 (5.2e+0)

We adjust the hyper-parameter α for baseline methods to ensure they yield intervals with nearly the same width as EnbPI, under identical setup to table I.

TABLE III
SOLAR POWER PREDICTION IN ATLANTA, COMPARISON OF EnbPI WITH ADAPTCI, ARIMA, EXPONENTIAL SMOOTHING, AND DYNAMIC FACTOR MODELS

α	0.05					0.10					0.15					0.20				
Method	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor	EnbPI	AdaptCI	ARIMA	Exp Smoothing	Dynamic Factor
Coverage	0.950	0.863	0.839	0.900	0.917	0.896	0.831	0.784	0.868	0.887	0.846	0.806	0.743	0.852	0.855	0.798	0.776	0.711	0.840	0.832
Width	288.581	215.258	158.581	351.181	262.006	216.989	187.504	135.404	313.185	229.151	178.140	173.079	119.870	288.428	206.448	147.297	154.322	107.652	269.379	187.840

We vary $\alpha \in [0.05, 0.10, 0.15, 0.20]$ and use the first 20% data as training data.

are based on one of the four prediction algorithms that yield the narrowest interval when reaching valid $1-\alpha$ coverage. Table I shows that out of all the CP methods, EnbPI is the only choice that consistently yields valid coverage at 0.9 regardless of the amount of training data. In contrast, the baseline CP methods may yield narrower intervals than EnbPI, yet their intervals often have a high coverage gap with respect to the 0.9 target level. Hence, this indicates that EnbPI is the most suitable method for this dataset. To better compare EnbPI with the baselines, we adjust the α parameter for each baseline method so that they yield approximately the same interval widths as EnbPI. Table II compares the performance of all methods under adjusted α, where we see that baseline methods often fail to reach valid $1-\alpha$ coverage as EnbPI. In addition, we often need to use extremely conservative values of α to reach the same interval widths as EnbPI (e.g., reduce to 0.03 for QOOB under 0.28 train ratio). Furthermore, EnbPI intervals also have the smallest standard deviation in width, indicating more stable interval construction by our proposed method.

In addition, Table III compares EnbPI with commonly used time-series methods, where we also include AdaptCI as the best-performing CP baseline method. Compared to EnbPI, the time-series baseline methods either yield conservative intervals under valid coverage or narrower intervals which nevertheless fail to cover at target $1-\alpha$ levels.

Remark 2 (Computational challenges of quantile-based conformal inference methods): Quantile regression models aim to predict quantiles of the response distribution accurately and capture the unknown distribution during inference. Such benefit can be reflected in the narrow prediction intervals by quantile-based conformal inference methods [12], [15], [20]. However, one should be cautious with the following subtle computational concern.

To fit a quantile regression model, one uses the empirical risk minimization under the following loss, which depends on the quantile α and the sign of the residual $\hat{\epsilon}_i := y_i - \hat{f}(x_i)$:

$$\mathcal{L}(\hat{\epsilon}_i, \alpha) = \begin{cases} \alpha\hat{\epsilon}_i & \text{if } \hat{\epsilon}_i \geq 0, \\ (\alpha-1)\hat{\epsilon}_i & \text{if } \hat{\epsilon}_i < 0. \end{cases} \tag{11}$$

Therefore, producing intervals at different desired $1-\alpha$ coverage levels requires fitting the baseline algorithm $\mathcal{A}$ inside a quantile-based conformal method multiple times.

In comparison, EnbPI trains the LOO estimators only once to compute all LOO residuals, during which one needs not to specify the desired α value (see Algorithm 1, line 1-10). Then, constructing intervals at a particular $1-\alpha$ only requires making a point prediction using fitted LOO estimators and evaluating the empirical quantiles of LOO residuals. The whole procedure is computationally efficient when different target coverage levels are specified.

C. Real-Data: Missing Data, Conditional Coverage

In this section, we move beyond marginal coverage with two particular goals. First, we aim to show conditional validity of EnbPI as it looks ahead beyond one step to construct multiple prediction intervals before receiving feedback (that is, $s > 1$). Second, we show that EnbPI can handle time series with missing data, which commonly exist in reality. We compare EnbPI against QOOB and AdaptCI in this setting.

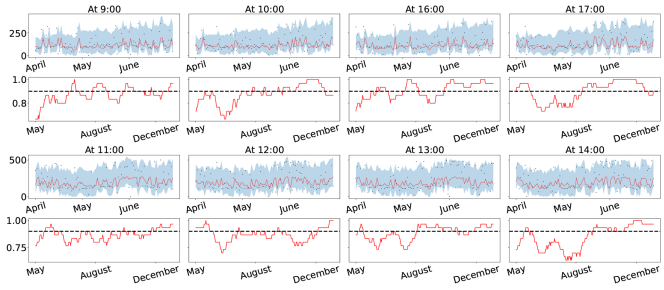

Fig. 2. Solar power prediction in Atlanta, when EnbPI looks ahead beyond one step. At each hour (i.e., a two-row subfigure), the top figure visualizes observations in black, estimates in red, and prediction intervals in blue for three months (April-June). The bottom subfigures compute coverage using a sliding window of 30 days. The sliding coverage is much poorer near summertime (for example, August), when the data distribution may differ. Conditional coverage at each hour is always near 0.9 (cf. Table 5).

Setup: The same setup applies to all three conformal inference methods, so we only describe the general setup. All hyperparameters except choices of s are kept the same unless otherwise specified. We fit each CP method separately on subsets of hourly data, given that radiation data exhibit significant periodic variations (for example, recordings near noon have much larger magnitudes than the rest). More precisely, we fit each CP method once on data between 10 AM —2 PM and once on data from the rest 5 hours. Then, we let $s = 5$ hours, so EnbPI constructs five-hour ahead prediction intervals every day, after which the conditional coverage is computed separately at each hour. To create a more challenging missing data situation, we randomly drop 25% of both training and test data. As X_t may contain the history of Y_t for prediction, we impute missing entries as independent random samples from a normal distribution, whose mean and variance parameters are empirical mean and standard error of the most recent s observations. We assume exogenous features (temperature, humidity, wind speed, etc.) are readily available and perform no imputation on them. The training data come from the first 92 days of observation (January-March), and intervals always lie within $[0, \infty)$, as solar radiation value cannot be negative. For clarity, we only show results under one typical trial.

Results: Fig. 2 shows conditional coverage of EnbPI under RF. We title each subfigure by the hour, in which the bottom row visualizes the coverage over a sliding window to illustrate how EnbPI performance evolves. Several things are noticeable. First, despite not being shown, empirical distributions of LOO residuals in the rightmost figures are asymmetric around 0, justifying the need to build asymmetric intervals in EnbPI. Second, EnbPI can nearly obtain conditional coverage at all these hours (see the first row of Table 5) even with missing data. We note that the sliding coverage can be much poorer near the summer (for example, in August), likely because radiation data near the summer experience unknown shifts in the model f and violate our assumption for the data-generating process. Lastly, applying EnbPI separately onto group training data that are more "similar" (for example, by morning and afternoon) can be essential, especially when the data-generating processes are heterogeneous over subgroups. In general, we believe EnbPI

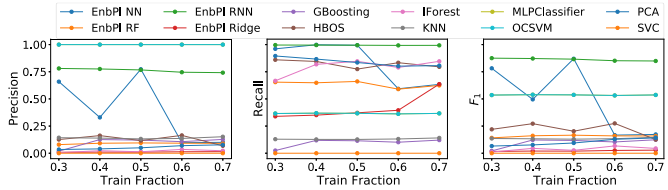

Fig. 3. Traffic flow anomaly detection. Precision, Recall, and F_1 scores versus different amounts of training data (as percentages of total data) for different detectors. EnbPI under RNN and NN outperforms the other methods.

can obtain conditionally valid coverage on real data even in missing data. In Appendix B.3, available in the online supplemental material, we show more results when no feedback is available to EnbPI (that is, $s = \infty$), illustrating the necessity to slide past residuals for a dynamic interval calibration. Table 5 in Appendix B.3, available in the online supplemental material reports the conditional coverage and width for EnbPI, QOOB, and AdaptCI. We see that QOOB can lose coverage at all hours, but AdaptCI can maintain conditional validity. In particular, AdaptCI prediction intervals for radiation levels in the morning are almost identical in width to those by EnbPI. However, those for radiation levels in the afternoon are wider than those by EnbPI. In Appendix B.3, available in the online supplemental material, we also visualize the sliding coverage and prediction intervals by QOOB and AdaptCI as in Fig. 2 for EnbPI.

D. Real Data: Unsupervised Anomaly Detection

In this section, we use EnbPI to detect anomalies in traffic flow observations with missing data. In this setting, it is important to dynamically update decision thresholds (for example, upper and lower ends of prediction intervals) based on spatial and temporal information in the traffic sensor network because traffic data are highly correlated and non-stationary. Data description, setup, and comparison methods are described in Appendix B.6, available in the online supplemental material

Results: Fig. 3 compares all methods on a particular traffic sensor as we vary the size of training data. It is clear that EnbPI consistently obtains the highest F_1 scores when RNN is used as the prediction model; F_1 scores by EnbPI also are consistent across over training sample sizes. In addition, Table IV shows the results with more sensors, from which EnbPI under NN or RNN still outperforms the other competitors by a large margin. In the future, we will consider multiple testing corrections to improve the performance [40], [41], [42], where the critical step is to examine the dependency of p-value as a correction step.

VI. EnbPI Under Change Points

In real applications, there can exist abrupt changes in the underlying data distribution, which are called *change points* [43], [44]. In this section, we present numerical experiments to demonstrate the performance of EnbPI in the presence of change points. We also discuss the potential adaption of EnbPI for change point detection.

TABLE IV
TRAFFIC FLOW ANOMALY DETECTION. F_1 SCORES, PRECISION, AND RECALL BY 12 METHODS ON SELECTED SENSORS. BOLD CELLS INDICATE THE HIGHEST SCORES. EnbPI RNN OR NN ARE BETTER ON THIS TASK IN TERMS OF F_1 SCORES

F_1 score

Sensor ID	EnbPI Ridge	EnbPI RF	EnbPI NN	EnbPI RNN	HBOS	IForest	OCSVM	PCA	SVC	GBoosting	KNN	MLPClassifier
282	0.13	0.14	0.88	**0.88**	0.16	0.02	0.51	0.09	0.0	0.04	0.07	0.51
248	0.02	0.17	0.87	**0.87**	0.20	0.03	0.54	0.09	0.0	0.12	0.13	0.54
151	0.02	0.14	**0.81**	0.80	0.11	0.04	0.39	0.08	0.0	0.08	0.12	0.39
235	0.57	0.59	0.77	**0.77**	0.01	0.00	0.45	0.00	0.0	0.23	0.24	0.45

Precision

Sensor ID	EnbPI Ridge	EnbPI RF	EnbPI NN	EnbPI RNN	HBOS	IForest	OCSVM	PCA	SVC	GBoosting	KNN	MLPClassifier
282	0.46	0.59	0.96	**0.96**	0.58	0.71	0.34	0.75	0.0	0.04	0.07	0.34
248	0.37	0.66	0.99	**0.99**	0.77	0.85	0.37	0.84	0.0	0.11	0.13	0.37
151	0.24	0.61	0.96	**0.96**	0.30	0.47	0.24	0.46	0.0	0.08	0.11	0.24
235	0.60	0.60	0.70	**0.70**	0.04	0.03	0.29	0.00	0.0	0.23	0.24	0.29

Recall

Sensor ID	EnbPI Ridge	EnbPI RF	EnbPI NN	EnbPI RNN	HBOS	IForest	OCSVM	PCA	SVC	GBoosting	KNN	MLPClassifier
282	0.07	0.08	0.81	0.81	0.10	0.01	**1.0**	0.05	0.0	0.04	0.07	1.0
248	0.01	0.09	0.77	0.77	0.12	0.01	**1.0**	0.05	0.0	0.12	0.13	1.0
151	0.01	0.08	0.69	0.68	0.07	0.02	**1.0**	0.04	0.0	0.09	0.12	1.0
235	0.55	0.59	0.87	0.87	0.01	0.00	**1.0**	0.00	0.0	0.24	0.24	1.0

F_1 scores, precision, and recall by 12 methods on selected sensors. Bold cells indicate the highest scores. EnbPI RNN or NN are better on this task in terms of F_1 scores.

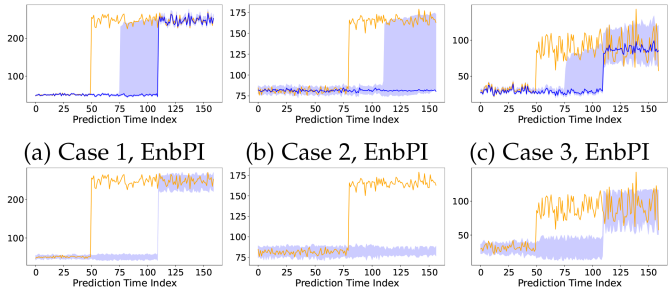

Fig. 4. Simulation with a change point at index 50. We overlay prediction intervals in shaded blue on top of the actual data. In particular, we collect 60 post-change data points to refit $\mathcal{A}$ at index 110. We expect that collecting more post-change data to fit EnbPI will yet better estimation with tighter prediction intervals.

We conider a change point happening during the testing phase and follow the setup in Section V-A. Assume a change point at $T^* = 0.6(T + T_1)$, which alters the underlying model f for the last 40% test data. As a result, the post-change responses Y_t are very different from the pre-change ones. We call the post-change model f_1. For the linear model, let $f_1(X_t) = \beta_1 X_t$ and β_1 be entry-wise *i.i.d.* $U[0, 5]$. Recall the pre-change β is entry-wise *i.i.d.* $U[0, 1]$. For the high-dimensional sparse linear model, β_1 has twice many non-zero components as that of β and the components are drawn from $U[0, 1]$ independently. For the nonlinear model, we keep the same β but square the value $f(X_t)$. Choices of X_t and ϵ_t remain the same in each case.

Recall T is the length of the pre-change training data; let $T = 0.3(T + T_1) = 600$. To adapt to post-change dynamics as quickly as possible, we retrain the prediction algorithm on $0.1\ T$ data after the change point T^*. We assume the T^* is known to us (for instance, we can be detected and estimated using a change point detection algorithm [43]). To quickly detect change points that highly correlate with differences in interval widths, we only take the empirical quantile of the most recent $T', T' < T$ residuals and fix $T' = 100$.

Fig. 4 plots prediction intervals on top of actual data for three cases. First, except for data indexed between T^* and $T^* + 0.1\ T$ (that is, between index 50 and 110 in the figure), most prediction data from both pre-change and post-change models are covered by EnbPI intervals. Second, prediction intervals built with pre-change models on post-change data tend to have much wider widths than others, reflecting a poor estimation of $\hat{f}$ by the pre-change models. Nevertheless, such a dramatic increase in width can enable change point detection, as we elaborate on below. Third, we observe that AdaptCI intervals are non-adaptive in this setting, as they fail to contain the true observations before retraining the predictive model. In Fig. 6, we further compare EnbPI with the ETS model [45], which shows similar performance as AdaptCI.

One can potentially adapt EnbPI to detect change points as follows. From Fig. 4, we observe that the change point leads to unusually wide post-change prediction intervals. As a result, one should monitor *both* the evolution of interval widths *and* coverage performances. On the one hand, when only f changes but the distribution of errors remains the same, the interval tends to be wider, but the coverage is worse. On the other hand, if f remains the same but the distribution changes, intervals may also become wider. However, coverage may not be as greatly affected because estimators by EnbPI can approximate f well. Due to a sliding window over residuals, one can adapt to the post-change distribution. These ideas resonate with several other works: [20] construct prediction sets under distribution shifts sequentially and prove that when shifts are small, the marginal coverage is approximately maintained. As a result, when coverage is significantly less than $1 - \alpha$, it can indicate an abrupt shift in distribution. Such ideas may also be used to test whether the test distribution lies in an f-divergence ball of the training distribution, given i.i.d. training and test data from the corresponding distribution [18]; extensions to time series remain unexplored. On the other hand, a line of works [46], [47], [48] builds martingales to detect change points which however, violates data exchangeability. The lower bound for the average-run-length is established for the Shiryaev–Roberts

procedure using such martingale [46, Proposition 4.1]. How to extend the ideas beyond testing exchangeable data remains an open question.

VII. Conclusions and Discussions

In this paper, we present a predictive inference method for time series. Theoretically, we can show that the constructed intervals are asymptotically valid without assuming data exchangeability: relaxing this requirement is crucial for time series data, and the interval width converges to the oracle one. We also present a simple, computationally friendly, and interpretable algorithm called `EnbPI`, which is an efficient ensemble-based wrapper for many prediction algorithms, including deep neural networks. Empirically, it works well on time series from various applications, including network data and data with missing entries, and maintains validity when other predictive inference methods fail. Furthermore, one can use `EnbPI` for unsupervised sequential anomaly detection. While the theoretical guarantee of `EnbPI` requires consistent estimation of the true model, empirical results are valid even under potentially misspecified models, and coverage is almost always valid.

Future work includes several possible directions. We may adapt `EnbPI` for classification problems [49], [50], [51] by defining conformity scores other than residuals. It can also be interesting to further develop `EnbPI` for online change point detection and adaptation for time series, extending the idea of sequential testing of data exchangeability [52] based on the Shiryaev-Roberts procedure.

Acknowledgments

The method presented in this paper has been implemented in open-source packages MAPIE [53] and Fortuna [54].

References

[1] F. Díaz-González, A. Sumper, O. Gomis-Bellmunt, and R. Villafáfila-Robles, "A review of energy storage technologies for wind power applications," *Renewable Sustain. Energy Rev.*, vol. 16, no. 4, pp. 2154–2171, 2012.

[2] J. Cochran, P. Denholm, B. Speer, and M. Miller, "Grid integration and the carrying capacity of the us grid to incorporate variable renewable energy," National Renewable Energy Lab.(NREL), Golden, CO USA, Tech. Rep., 2015.

[3] H. Gangammanavar, S. Sen, and V. M. Zavala, "Stochastic optimization of sub-hourly economic dispatch with wind energy," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 949–959, Mar. 2016.

[4] Y. Gu and L. Xie, "Stochastic look-ahead economic dispatch with variable generation resources," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 17–29, Jan. 2017.

[5] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *Int. J. Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[6] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," 2017, *arXiv: 1711.11053*.

[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[8] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2065–2081, Sep. 2020.

[9] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[10] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, no. Mar, pp. 371–421, 2008.

[11] H. Papadopoulos, V. Vovk, and A. Gammerman, "Conformal prediction with neural networks," in *Proc. IEEE 19th Int. Conf. Tools Artif. Intell.*, 2007, pp. 388–395.

[12] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3543–3553.

[13] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *Ann. Statist.*, vol. 49, no. 1, pp. 486–507, 2021, doi: 10.1214/20-AOS1965.

[14] B. Kim, C. Xu, and R. F. Barber, "Predictive inference is free with the jackknife+-after-bootstrap," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4138–4149. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/2b346a0aa375a07f5a90a344a61416c4-Paper.pdf

[15] C. Gupta, A. K. Kuchibhotla, and A. Ramdas, "Nested conformal prediction and quantile out-of-bag ensemble methods," *Pattern Recognit.*, vol. 127, 2021, Art. no. 108496.

[16] G. Zeni, M. Fontana, and S. Vantini, "Conformal prediction: A unified review of theory and new challenges," 2020, *arXiv: 2005.07972*.

[17] R. J. Tibshirani, R. F. Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2530–2540.

[18] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi, "Robust validation: Confident predictions even when distributions shift," 2020, *arXiv:2008.04267*.

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

[20] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1660–1672.

[21] M. Zaffran, A. Dieuleveut, O. F'eron, Y. Goude, and J. Josse, "Adaptive conformal predictions for time series," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 25834–25866.

[22] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, "Time series: Theory and methods: Theory and methods," *Springer Sci. Bus. Media*, 1991.

[23] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*. Press, London, U.K.: Oxford Univ. Press, 2012.

[24] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting With Exponential Smoothing: The State Space Approach*. Berlin, Germany: Springer, 2008.

[25] M. Bańbura and M. Modugno, "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data," *J. Appl. Econometrics*, vol. 29, no. 1, pp. 133–160, 2014.

[26] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022.

[27] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," 2019, *arXiv: 1903.04684*.

[28] V. Chernozhukov, K. Wüthrich, and Z. Yinchu, "Exact and robust conformal inference methods for predictive machine learning with dependent data," in *Proc. Conf. Learn. Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., PMLR, Jul. 06–09, 2018, pp. 732–749. [Online]. Available: http://proceedings.mlr.press/v75/chernozhukov18a.html

[29] V. Chernozhukov, K. Wuthrich, and Y. Zhu, "An exact and robust conformal inference method for counterfactual and synthetic controls," 2020, *arXiv: 1712.09089*.

[30] M. Sesia and Y. Romano, "Conformal histogram regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 6304–6315.

[31] G. Casella and R. L. Berger, *Statistical inference*. Cengage Learning, 2021.

[32] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statist.*, vol. 27, pp. 642–669, 1956.

[33] C. Hesse, "Rates of convergence for the empirical distribution function and the empirical characteristic function of a broad class of linear processes," *J. Multivariate Anal.*, vol. 35, no. 2, pp. 186–202, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0047259X9090024C

[34] E. Rio, *Asymptotic Theory of Weakly Dependent Random Processes*, vol. 80. Berlin, Germany: Springer, 2017.

[35] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[36] X. Chen and H. White, "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 682–691, Mar. 1999.

[37] P. J. Bickel et al., "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.
[38] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proc. Nat. Acad. Sci. USA*, vol. 42, pp. 43–7, 1956.
[39] J.-P. Kreiss and E. Paparoditis, "Bootstrap methods for dependent data: A review," *J. Korean Stat. Soc.*, vol. 40, pp. 357–378, 2011.
[40] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia, "Testing for outliers with conformal p-values," *Ann. Statist.*, vol. 51, no. 1, pp. 149–178, 2023.
[41] S. Chen and S. Kasiviswanathan, "Contextual online false discovery rate control," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, ser. Proceedings of Machine Learning Research, PMLR, Aug. 26–28 2020, pp. 952–961. [Online]. Available: http://proceedings.mlr.press/v108/chen20b.html
[42] A. Ramdas, F. Yang, M. J. Wainwright, and M. I. Jordan, "Online control of the false discovery rate with decaying memory," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2017, pp. 5650–5659. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/7f018eb7b301a66658931cb8a93fd6e8-Paper.pdf
[43] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 494–514, Jun. 2021.
[44] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 339–367, 2017.
[45] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and practice*. OTexts, 2013.
[46] V. Vovk, "Testing randomness online," *Stat. Sci.*, vol. 36, pp. 595–611, 2021.
[47] V. Vovk, "Conformal testing in a binary model situation," *Conformal Probabilistic Prediction Appl.*, pp. 131–150, 2021.
[48] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, and A. Gammerman, "Retrain or not retrain: Conformal test martingales for change-point detection," *Conformal Probabilistic Prediction Appl.*, pp. 191–210, 2021.
[49] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," 2020, *arXiv: 2009.14193*.
[50] Y. Romano, M. Sesia, and E. Candès, "Classification with valid and adaptive coverage," in *Proc. Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3581–3591.
[51] C. Xu and Y. Xie, "Conformal prediction set for time-series," 2022. [Online]. Available: https://arxiv.org/abs/2206.07851
[52] D. Volkhonskiy, E. Burnaev, I. Nouretdinov, A. Gammerman, and V. Vovk, "Inductive conformal martingales for change-point detection," in *Proc. Conf. Conformal Probabilistic Prediction Appl.*, PMLR, 2017, pp. 132–153.
[53] V. Taquet, V. Blot, T. Morzadec, L. Lacombe, and N. Brunel, "MAPIE: An open-source library for distribution-free uncertainty quantification," 2022, *arXiv:2207.12274*.
[54] G. Detommaso, A. Gasparin, C. Archambeau, M. Donini, M. Seeger, and A. G. Wilson, "Awslabs/fortuna: A library for uncertainty quantification," Dec. 2022. [Online]. Available: https://github.com/awslabs/Fortuna
[55] M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*. Berlin, Germany: Springer, 2007.
[56] K. B. Athreya and S. G. Pantula, "A note on strong mixing of arma processes," *Statist. Probability Lett.*, vol. 4, no. 4, pp. 187–190, 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167715286900647
[57] C. Kath and F. Ziel, "Conformal prediction interval estimation and applications to day-ahead and intraday power markets," *Int. J. Forecasting*, vol. 37, no. 2, pp. 777–799, Apr. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.ijforecast.2020.09.006
[58] D. Lucas et al., "Designing optimal greenhouse gas observing networks that consider performance and cost," *GeoSci. Instrum., Methods Data Syst.*, vol. 4, no. 1, 2015, Art. no. 121.
[59] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy Buildings*, vol. 140, pp. 81–97, 2017.
[60] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Roy. Soc. A Math. Phys. Eng. Sci.*, vol. 473, no. 2205, 2017, Art. no. 20170457.
[61] C. Xu and Y. Xie, "Conformal anomaly detection on spatio-temporal observations with missing data," 2021, *arXiv:2105.11886*.
[62] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1–39, 2012.
[63] C. C. Aggarwal, "Outlier analysis," in *Data mining*. Berlin, Germany: Springer, 2015, pp. 237–263.
[64] M. Goldstein and A. Dengel, "Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm," *KI-2012 Poster Demo Track*, vol. 1, pp. 59–63, 2012.

Chen Xu received the BSc degrees in computational and applied mathematics and economics and the MSc degree in statistics from the University of Chicago in 2020. He is currently working toward the PhD degree with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. His research interests are uncertainty quantification, neural networks, and spatio-temporal data modeling.

Yao Xie (Member, IEEE) received the PhD degree in electrical engineering (minor in mathematics) from Stanford University and was a research scientist with Duke University. She is a professor and Harold R. and Mary Anne Nash Early Career professor with the Georgia Institute of Technology in the H. Milton Stewart School of Industrial and Systems Engineering and associate director of the Machine Learning Center. Her research interests include the intersection of statistics, machine learning, and optimization in providing theoretical guarantees and developing computationally efficient and statistically powerful methods for problems motivated by real-world applications. She received the National Science Foundation (NSF) CAREER Award in 2017, INFORMS Wagner Prize Finalist in 2021, and the INFORMS Gaver Early Career Award for Excellence in Operations Research in 2022. She is currently an associate editor for *IEEE Transactions on Information Theory*, *IEEE Transactions on Signal Processing*, *Journal of the American Statistical Association*, *Theory and Methods*, *Sequential Analysis: Design Methods and Applications*, *INFORMS Journal on Data Science*, and an area chair of NeurIPS and ICML.