

Received 16 January 2024, accepted 31 March 2024, date of publication 11 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387858

## RESEARCH ARTICLE

# Adaptive Conformal Prediction Intervals Using Data-Dependent Weights With Application to Seismic Response Prediction

PARISA HAJIBABAE<sup>1</sup>, FARHAD POURKAMALI-ANARAKI<sup>2</sup>,  
AND MOHAMMAD AMIN HARIRI-ARDEBILI<sup>3,4</sup>

<sup>1</sup>Department of Data Science and Business Analytics, Florida Polytechnic University, Lakeland, FL 33805, USA

<sup>2</sup>Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80204, USA

<sup>3</sup>College of Computer, Mathematical, and Natural Sciences, University of Maryland, College Park, MD 20742, USA

<sup>4</sup>Department of Civil Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, CO 80309, USA


Corresponding author: Parisa Hajibabae (phajibabae@floridapoly.edu)

**ABSTRACT** Machine learning often lacks transparent performance indicators, especially in generating point predictions. This paper addresses this limitation through conformal prediction, a non-parametric forecasting technique seamlessly integrating with regression algorithms to produce prediction intervals at specified confidence levels. A crucial element in conformal prediction is the non-conformity score, traditionally based on absolute residual errors. In this work, we propose a novel approach, introducing data-dependent weights for computing non-conformity scores. This enhancement, considering the distances of training instances from the test sample, aims to improve overall algorithm performance. Empirical investigations across various real-world regression data sets, including scientific data, evaluate the efficiency and validity of prediction intervals from different uncertainty quantification methods. Results show that prediction intervals computed with data-dependent weights adapt to estimator uncertainty, offering more precise predictions in certain scenarios and appropriately conservative predictions in high uncertainty situations. Additionally, we compare predictive regions generated by conformal prediction with those from Gaussian Process Regression (GPR) for scientific data in structural engineering. To augment conformal prediction, we explore Conformalized Quantile Regression (CQR), a recent innovation combining conformal prediction with classical quantile regression, claiming full adaptability to heteroscedasticity. Our findings indicate that conformal prediction methods using data-dependent non-conformity scores achieve a 1% higher effective coverage level and a 15% reduction in prediction interval widths compared to other methods. The comparative analysis against GPR and CQR underscores the practical value of our approach in providing accurate prediction intervals in scientific and engineering domains.

**INDEX TERMS** Conditional coverage, conformal prediction, non-conformity scores, uncertainty quantification.

## I. INTRODUCTION

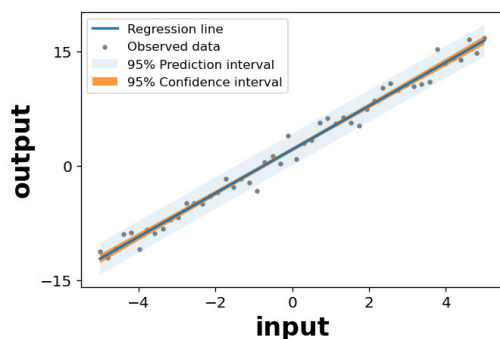
Machine learning models are valuable tools for predicting outcomes in many scientific application areas. However, they are mostly based on point prediction methods without any indication of their accuracy. These single-value-prediction-based

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu .

models provide no information about associated uncertainties or their level of reliability. This problem has become more challenging when noise and outliers are present in the data set or the outcomes are affected by inherent randomness [1]. When it comes to the decision-making process in real-world problems, practitioners and engineers are eager to consider several scenarios/solutions for uncertain conditions [2]. Providing the level of uncertainty associated with point

estimations would allow for more confident decision-making. The uncertainty of machine learning models has been a topic of interest in recent years, leading to the development of various methods for quantifying uncertainty and evaluating their effectiveness in various applications. For example, studies [3], [4], [5] discuss various approaches to quantifying uncertainty in machine learning models for engineering problems, including methods such as conformal prediction, Bayesian uncertainty estimation, and bootstrapping. They also discuss the challenges and benefits of using these methods in engineering applications and provide an overview of the current state of the field.

One solution to quantifying the uncertainty associated with machine learning models is to provide prediction intervals rather than point predictions. Prediction intervals and confidence intervals are often confused, but they are distinct concepts. While confidence intervals quantify the uncertainty in a population parameter, such as the mean or standard deviation of a probability distribution, prediction intervals represent the uncertainty of the predicted outcome for a single new/test sample. Prediction intervals are typically wider than confidence intervals because they take into account both the uncertainty associated with the population parameter and the variance of individual values at random [6]. That is, the prediction interval for a given new test sample with a user-predefined confidence level is expected to cover a *moving response*, whereas the confidence interval is only intended to cover *fixed response*. Fig. 1 shows a simple example of the distinction between a confidence interval and a prediction interval. In this figure, there is a 95% probability that the true best-fit line for the population lies within the confidence interval. Additionally, it is expected that 95% of the output values to be found for a certain input value will be within the prediction interval around the linear regression line.



**FIGURE 1. Visualization of prediction intervals and confidence intervals for a linear regression problem.**

There are various approaches to constructing prediction intervals, ranging from Bayesian and variational inference to frequentist methods. Using a Bayesian framework, the posterior credible set quantifies prediction uncertainty. In practice, exact Bayesian inference is computationally prohibitive for most machine learning and deep learning models. With variational inference, the true posterior distribution is

approximated by an ensemble of neural networks, similar to non-Bayesian ad hoc ensemble methods [7], [8], [9]. In these methods, the model disagreement in the ensemble can be used to measure prediction uncertainty. Nonetheless, a poor approximation of the posterior distribution would result in an incorrect quantification of prediction uncertainty. A frequentist coverage guarantee cannot generally be satisfied by these methods [10], [11]. Under certain additional assumptions regarding the functional parameters, hierarchical and empirical Bayes methods can construct reliable prediction regions asymptotically [12], [13]. However, it is challenging to confirm these assumptions in real-world situations, and some machine learning models do not satisfy these assumptions.

One of the most important uncertainty quantification methods that satisfy the frequentist coverage guarantee in finite sample regimes is conformal prediction [14]. Conformal prediction is a statistical technique that is used to make predictions while also providing a measure of the uncertainty of those predictions. It is based on the idea of using a set of “conformal” or “calibrated” models to make predictions about new data. Conformal prediction is a non-parametric forecasting technique that is distribution-free and based on minimal assumptions about the data. It can be combined with any regression algorithm to generate predictive regions to satisfy a given confidence level. It should be noted that two properties must be met for a prediction interval generation procedure to be effective. Firstly, it should be capable of providing valid coverage in finite samples, without imposing strong distributional assumptions, such as Gaussianity. This is referred to *validity* as an index of reliability. Secondly, the intervals at each point in the input space must be as narrow as possible, to ensure that the predictions are informative. In other words, the level of uncertainty associated with predictions should be as low as possible. This is referred to *efficiency* as an index of informativeness. The efficiency of the predictive region is also known as the sharpness of the predictive region. When comparing conformal prediction methods, the most important criteria are the tightness of the prediction intervals (efficiency) and the target coverage level (validity). In conformal prediction, the non-conformity score is an important component that is typically calculated based on the absolute residual error of the sample’s actual and predicted values. With this type of non-conformity score, conformal prediction methods produce conservative prediction intervals that are constant or weakly varying in width across an input space. Ideally, the size of the predictive regions should vary according to how difficult it is to predict. This paper aims to enhance the efficiency and adaptivity of the prediction intervals while maintaining their validity. The focus is on making the prediction intervals more sensitive to the difficulty of the inputs and their distributions, thereby ensuring that they provide accurate coverage.

### A. PRIOR WORK

The foundational work by Vovk et al. [14] introduced the fundamental framework of conformal prediction. Subsequent

research [15], [16] has highlighted the challenges of achieving conditional validity for prediction intervals without specific assumptions about model regularity and consistency. Conformal prediction with neural networks has been a subject of recent investigations, exploring coverage probability, prediction accuracy, and computational complexity [17], [18], [19].

Traditional conformal prediction, relying on estimating a conditional mean function with constant interval widths, assumes homoscedastic errors. To address heteroscedasticity, Romano et al. [20] proposed conformalized quantile regression. Kivaranovic et al. [21] explored adaptive and distribution-free prediction intervals using deep neural networks. A comparative study by Sesia and Candès [22] delved into conformal prediction based on quantile regression with different non-conformity scores. Despite the flexibility of regression quantiles, the quantile-crossing problem has been observed in simple linear quantile regression models. This issue becomes even worse in multivariate quantile regression. Several research studies have been conducted in recent decades on the quantile-crossing problem. For example, the authors in [23] proposed simultaneous quantile regression as a method of estimating quantiles by minimizing the pinball loss, whereas the target quantile is randomly sampled in every training iteration. The algorithm presented in [24] is designed to predict an arbitrary number of quantiles, which can maintain quantile monotonicity by restricting the partial derivatives of the quantile functions. The use of these approaches may alleviate the problem of quantile-crossing; however, they cannot eliminate it entirely.

## B. NOTATION

In this paper, we denote column vectors with lower-case bold letters and real values with lower-case letters. We consider a set of  $n$  independent and identically distributed (i.i.d.) data points  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  with  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  along with corresponding scalar-valued outputs denoted by  $y_1, \dots, y_n \in \mathbb{R}$ .

We assume the function  $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$  to be a regression model fitted on the training data set. Given the new test sample  $(\mathbf{x}_{n+1}, y_{n+1})$ , taken from the same distribution, the function  $\hat{f}$  can provide a predicted output value as  $\hat{f}(\mathbf{x}_{n+1})$ . However, in this study, we aim to construct a *prediction interval*  $\hat{C} : \mathbb{R}^d \rightarrow \mathbb{R}^2$  (here, 2 refers to lower and upper bounds of the interval) instead of *point estimation*  $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ . We assume a significance level  $\alpha$  or the target coverage level  $1 - \alpha$ , where  $\alpha \in (0, 1)$ . We let  $q_{\alpha,n}^+ \{v_i\}$  represent the upper bound of the interval such that for any values  $v_i, i = 1, \dots, n$ , we define  $q_{\alpha,n}^+ \{v_i\} =$  the  $\lceil (1 - \alpha)(n + 1) \rceil^{\text{th}}$  smallest value of  $v_i, i = 1, \dots, n$ . Similarly, we let  $q_{\alpha,n}^- \{v_i\}$  represent the lower bound of the interval such that for any values  $v_i, i = 1, \dots, n$ , we define  $q_{\alpha,n}^- \{v_i\} =$  the  $\lceil (1 - \alpha)(n + 1) \rceil^{\text{th}}$  largest value of  $v_i, i = 1, \dots, n$ .

## C. REGRESSION AND PERFORMANCE METRICS

This study focuses solely on regression problems, and we mostly use ridge regression, one of the most popular machine

learning methods, as an underlying (base) algorithm. In the following, we define some notions and performance metrics in a regression setting, which are mostly used by researchers in the area of uncertainty quantification [14].

- **Marginal Coverage:** Given target coverage level  $1 - \alpha$ , we aim to construct a prediction interval  $\hat{C}_{\alpha,n}$  for a new test sample  $\mathbf{x}_{n+1}$  such that

$$\mathbb{P} \left\{ y_{n+1} \in \hat{C}_{\alpha,n}(\mathbf{x}_{n+1}) \right\} \geq 1 - \alpha, \quad (1)$$

where the probability is taken over  $n$  training data samples and the unseen test sample  $(\mathbf{x}_{n+1}, y_{n+1})$ . Marginal coverage property of  $\hat{C}_{\alpha,n}$  is only defined as having  $1 - \alpha$  chance of being correct on average across all data points (marginalizing over the test sample). However, for some sub-regions in the data set, there may be no marginal coverage at all. Marginal coverage has the disadvantage of not being conditioned upon  $\mathbf{x}$ . This is where conditional coverage at  $1 - \alpha$  would be more appropriate.

- **Conditional Coverage:** The concept of adaptivity, which implies that the size of the prediction intervals can change according to the difficulty of the inputs, is typically formalized by asking for the conditional coverage property such that

$$\mathbb{P} \left\{ y_{n+1} \in \hat{C}_{\alpha,n}(\mathbf{x}_{n+1}) | \mathbf{x}_{n+1} = (\mathbf{x}) \right\} \geq 1 - \alpha, \quad (2)$$

for almost all  $\mathbf{x}$  or training samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . Here, we fix  $\mathbf{x}_{n+1}$  and the probability is taken over the training data points and  $y_{n+1}$ . The conditional coverage may exceed  $1 - \alpha$  at some values of  $\mathbf{x}_{n+1} = \mathbf{x}$  and may be less than  $1 - \alpha$  in other cases. In other words, we aim to return prediction regions with  $1 - \alpha$  coverage for every value of the training and test sets. For example, the coverage of the prediction intervals should be consistent for all inputs, regardless of their difficulty. This property is important for ensuring the reliability and informativeness of the prediction intervals.

- **Effective Coverage:** The test sample is considered covered if the true label falls within the prediction interval. Using an average over all samples in a test set, we can estimate the method's effective coverage for the data set. Effective coverage is the actual fraction of test points whose true values lie within the prediction intervals.
- **Prediction Interval Width:** The distance between the upper limit and the lower limit of a prediction interval represents the width of that prediction interval. In our analysis, we use an average of prediction interval widths over all samples in a test set.

## D. MAIN CONTRIBUTIONS

Researchers in conformal prediction are actively innovating methods to construct enhanced prediction intervals, extend applications across domains, explore intersections with other machine learning areas, incorporate domain-specific

knowledge, and tailor approaches for specific data types. Despite its strengths, conformal prediction often leans towards conservative intervals, even in situations of high estimator certainty. This study concentrates on refining the regression non-conformity measure, specifically addressing absolute residuals, with the primary goal of improving efficiency and enhancing conditional coverage.

In the following, we provide a list of our contributions.

- We propose a new non-conformity measure that considers the relevance between the training instances and the test sample in regression conformal prediction. Non-conformity scores are calculated by assigning weights to the training points based on their distance from the test sample. In this way, the constructed prediction intervals are more adaptive than those derived from absolute residuals.
- We present a thorough empirical evaluation of the proposed and existing non-conformity measures using several real-world data sets. To illustrate the merits and limitations of these uncertainty quantification techniques in the presence of heteroscedastic or not identically distributed data, we report effective coverage, conditional coverage, and prediction interval width for different target coverage levels.
- We show that the proposed non-conformity measure is suitable for quantifying uncertainty in scientific applications. Developing machine learning-based surrogate models for advanced computer simulations of scientific models has received significant attention [25], [26], [27], [28], [29], [30], [31], [32]. This paper studies a regression data set stemming from computational models of structural and earthquake engineering. The objective is to quantify uncertainty using conformal prediction methods to construct prediction intervals of the structural response, e.g., shear forces.
- As part of our study, we compare the predictive regions produced by conformal prediction methods with those produced by Gaussian Process Regression (GPR) and Conformalized Quantile Regression (CQR), which are two of the most popular Bayesian and frequentist machine learning approaches, respectively. Specifically, our proposed method achieves narrower prediction intervals with higher effective and conditional coverage rates than both GPR and CQR, demonstrating its suitability for practical uncertainty quantification in scientific applications.

## E. PAPER ORGANIZATION

The remainder of the paper is outlined as follows: In Section II, we start with some background information on conformal prediction methods. Section III explains the proposed idea of using data-dependent weights in leave-one-out/out-of-fold non-conformity scores when constructing prediction intervals. Next, we present our numerical experiments and discussions based on the synthetic data set and

standard benchmark data sets in Section IV. Furthermore, in Section V, we outline the scientific simulation model in earthquake engineering and demonstrate the effectiveness of proposed non-conformity scores for uncertainty quantification. Finally, Section VI presents concluding remarks and future research directions.

## II. BACKGROUND

### A. CONFORMAL PREDICTION

This section provides a brief description of conformal prediction; the interested reader is referred to the book [14] for a more detailed description. There is only one model fitting step required for split conformal prediction, but this comes at the expense of statistical efficiency. In contrast, full conformal prediction is more computationally intensive and requires many steps for model fitting but is statistically more efficient.

- **Split Conformal Prediction:** This method is also called inductive conformal prediction which begins by partitioning the  $n$  available labeled data points into two disjoint subsets including a training set of size  $m$  ( $m < n$ ) and a holdout set of size  $n - m$ . We let training set as  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$  and holdout set as  $(\mathbf{x}_{m+1}, y_{m+1}), \dots, (\mathbf{x}_n, y_n)$ . To obtain a fitted model  $\hat{f}$ , the regression algorithm is run on the training set. Finally, the prediction interval on the new test sample  $(\mathbf{x}_{n+1}, y_{n+1})$  is defined as

$$\hat{C}_{\alpha,n}(\mathbf{x}_{n+1}) = \hat{f}(\mathbf{x}_{n+1}) \pm \hat{q}_{holdout}, \quad (3)$$

where  $\hat{q}_{holdout}$  is defined as  $\lceil (1 - \alpha)(n - m + 1) \rceil$  - th smallest value of the residuals in holdout set  $|y_{m+1} - \hat{f}(\mathbf{x}_{m+1})|, \dots, |y_n - \hat{f}(\mathbf{x}_n)|$ . Although split conformal offers both computation efficiency and distribution-free coverage, its accuracy may be compromised due to the loss of sample size resulting from the subdivision of the data set.

- **Full Conformal Prediction:** This method is also called transductive conformal prediction, and it is extremely computationally intensive since it makes use of all the training data available for model fitting. Here, we assume every possible label of the new test sample as  $(\mathbf{x}_{n+1}, y)$ , where  $y \in \mathbb{R}$ . Then we fit the regression model on all  $n$  available labeled data points along with the new test sample as  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), (\mathbf{x}_{n+1}, y)$ . Finally, the prediction interval on the new test sample is defined as

$$\hat{C}_{\alpha,n}(\mathbf{x}_{n+1}) = \left\{ y \in \mathbb{R} : |y - \hat{f}(\mathbf{x}_{n+1})| \leq q^y \right\}, \quad (4)$$

where  $q^y$  is  $\lceil (1 - \alpha)(n + 1) \rceil$  - th smallest value of the residuals  $|y_1 - \hat{f}(\mathbf{x}_1)|, \dots, |y_n - \hat{f}(\mathbf{x}_n)|, |y - \hat{f}(\mathbf{x}_{n+1})|$ . In addition to the distribution-free coverage guarantee, full conformal prediction is statistically more efficient than split conformal prediction because it does not require splitting the training data. However,

the computational cost of this method is high; the prediction interval can only be determined by running the regression algorithm for every possible real value of  $y$  infinitely many times. To alleviate this problem, recent work [33] developed discretized conformal prediction algorithms that are guaranteed to cover the true value with the target coverage level. This method uses only a finite grid of finely spaced values for the response variable  $y$  to offer a trade-off between computational cost and prediction accuracy.

### B. JACKKNIFE CONFORMAL PREDICTION

There is a method known as jackknife prediction that lies between the computational complexity and statistical efficiency of the full conformal and split conformal methods. This method defines prediction intervals based on the quantiles of leave-one-out residuals. The “leave one out” procedure, similar to the cross-validation technique, was first proposed in [34]. First, the idea of the “leave one out” procedure was used to estimate the bias of an estimator. Then, [35] extended the use of this idea and named it as a jackknife estimate of standard error. The jackknife method can be used to estimate the parameter systematically. Given the data set with size  $n$ , we create a subsample of size  $(n - 1)$  each time by removing one data point and estimating the parameter i.e., the population mean of random variables, over the remaining data points. Here, we have  $n$  jackknife replicates which can be considered as an approximation of the distribution of the mean. Finally, the jackknife estimator can be obtained by aggregating these calculations and taking the average and variance of these  $n$  replicates. In comparison with the split conformal method, the jackknife method utilizes more of the training data to construct the absolute residuals and corresponding quantiles. Therefore, jackknife methods often produce shorter prediction intervals than split conformal methods.

In the following, we briefly review jackknife and its recently introduced modification, jackknife+, as well as CV and CV+, which are cross-validation versions.

- Jackknife (J): Given  $n$  training data points and target coverage level  $1 - \alpha$ , the jackknife (J) method considers a leave-one-out construction to find error margin. That is, for each training sample  $i = 1, \dots, n$ , we remove  $i$ -th sample point and then fit the regression function  $\hat{f}_{-i}$  to the remaining training data and compute prediction interval  $\hat{C}_{\alpha,n}$  on the new test sample as  $\hat{f}(\mathbf{x}_{n+1}) \pm$  ( the  $(1 - \alpha)$  quantile of  $\left| y_1 - \hat{f}_{-1}(\mathbf{x}_1) \right|, \dots, \left| y_n - \hat{f}_{-n}(\mathbf{x}_n) \right|$  ). The J technique should, on average, have the right coverage qualities since it avoids overfitting.

We denote a prediction interval with J method as  $\hat{C}_{\alpha,n}^J$ . On the new test sample, it can be computed as

$$\hat{C}_{\alpha,n}^J(\mathbf{x}_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}(\mathbf{x}_{n+1}) - R_i^{\text{LOO}} \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}(\mathbf{x}_{n+1}) + R_i^{\text{LOO}} \right\} \right], \quad (5)$$

where  $\hat{q}_{\alpha,n}^-$  and  $\hat{q}_{\alpha,n}^+$  denote the lower and upper  $(1 - \alpha)$  quantile of the distribution. The *leave-one-out non-conformity score*  $R_i^{\text{LOO}}$  can be defined as

$$R_i^{\text{LOO}} = \left| y_i - \hat{f}_{-i}(\mathbf{x}_i) \right|, i = 1, \dots, n. \quad (6)$$

- CV: CV prediction interval is cross-conformal prediction method [15], [36]. We should split the training data points into  $K$  disjoint subsets  $[n] = S_1 \cup \dots \cup S_K$  each of size  $m = \frac{n}{K}$ .  $\hat{f}_{-S_k}$  is as the fitted model when the  $k$ -th fold  $S_k$  is removed from the  $n$  training data points. The CV prediction interval is defined as

$$\hat{C}_{\alpha,n,K}^{\text{CV}}(X_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}(\mathbf{x}_{n+1}) - R_i^{\text{CV}} \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}(\mathbf{x}_{n+1}) + R_i^{\text{CV}} \right\} \right], \quad (7)$$

where  $q_{\alpha,n}^-$  and  $q_{\alpha,n}^+$  denote the lower and upper  $(1 - \alpha)$  quantile of the distribution.  $R_i^{\text{CV}}$ , the residuals from this  $K$ -fold process, is defined as

$$R_i^{\text{CV}} = \left| Y_i - \hat{f}_{-S_{k(i)}}(\mathbf{x}_i) \right|, i = 1, \dots, n, \quad (8)$$

where  $k(i) \in \{1, \dots, K\}$  identifies the subset that contains  $i$ , i.e.,  $i \in S_{k(i)}$ . CV requires fewer models to be calculated when we choose a smaller  $K$  ( $K$  models instead of  $n$  models in the J method), but at the expense of slightly wider intervals, since the models  $\hat{f}_{-S_k}$  are fitted with fewer samples i.e.,  $(n - n/K)$ , thus resulting in larger residuals.

- Jackknife+ (J+): As the name suggests, J+ is a simple modification of the J method to construct prediction interval [37]. Although both variants (J and J+) use the leave-one-out residuals, the distinction is that in the J method, we center our interval on the predicted value  $\hat{f}(\mathbf{x}_{n+1})$  fitted on the entire training data, whereas in J+, we use the leave-one-out predictions  $\hat{f}_{-i}(\mathbf{x}_{n+1})$  for the test sample point.

It should be noted that two methods, J and J+, should produce almost identical prediction intervals if the leave-one-out fitted functions  $\hat{f}_{-i}$  are similar to  $\hat{f}$ , fitted on the entire training data. However, the result may be quite different in cases where the regression algorithm is highly sensitive to the training data, eliminating one data point might significantly impact the predicted value at  $(\mathbf{x}_{n+1})$ . We denote a prediction interval with J+ method as  $\hat{C}_{\alpha,n}^{\text{J+}}$ . On the new test sample, it can be computed as

$$\hat{C}_{\alpha,n}^{\text{J+}}(\mathbf{x}_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}_{-i}(\mathbf{x}_{n+1}) - R_i^{\text{LOO}} \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}_{-i}(\mathbf{x}_{n+1}) + R_i^{\text{LOO}} \right\} \right], \quad (9)$$

where  $q_{\alpha,n}^-$  and  $q_{\alpha,n}^+$  denote the lower and upper  $(1 - \alpha)$  quantile of the distribution.  $R_i^{\text{LOO}}$  is defined in Eq. 6.

- CV+: CV+ is a  $K$ -fold cross-validation version of jackknife+ [37]. We should split the training data points into  $K$  disjoint subsets  $[n] = S_1 \cup \dots \cup S_K$  each of size

$m = \frac{n}{K}$ . In this sense, J+ can be considered a special case of CV+, with  $K = n$ . By choosing a smaller  $K$ , we compute only  $K$  instead of  $n$  models.  $\hat{f}_{-S_k}$  is as the fitted model when the  $k$ -th fold  $S_k$  is removed from the  $n$  training data points. Both CV and CV+ use out-of-fold non-conformity scores; however, the only modification is that CV centers our interval on the predicted value  $\hat{f}(\mathbf{x}_{n+1})$  fitted to the entire training data, whereas CV+ uses the out-of-fold predictions  $\hat{f}_{-i}(\mathbf{x}_{n+1})$  for the test sample point. The CV+ prediction interval is defined as

$$\hat{C}_{\alpha,n,K}^{CV+}(X_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}_{-S_{k(i)}}(\mathbf{x}_{n+1}) - R_i^{CV} \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}_{-S_{k(i)}}(\mathbf{x}_{n+1}) + R_i^{CV} \right\} \right], \quad (10)$$

where  $q_{\alpha,n}^-$  and  $q_{\alpha,n}^+$  denote the lower and upper  $(1 - \alpha)$  quantile of the distribution.  $R_i^{CV}$  is defined in Eq. 8. Compared to J+, it is likely that CV+ results in slightly wider intervals because the models  $\hat{f}_{-S_k}$  are fitted with fewer samples, i.e.,  $(n - n/K)$ , resulting in slightly larger residuals.

### III. PREDICTION INTERVALS WITH DATA-DEPENDENT WEIGHTS

In conformal prediction, the non-conformity score is a measure of how closely a sample complies with the training set. It is used to determine the likelihood that the sample belongs to the same distribution as the training set. In regression settings, the non-conformity score is typically calculated based on the absolute residual error between the actual and predicted values of the sample (L2-loss). However, using the L2-loss as the non-conformity measure can result in conservative prediction intervals that are constant or weakly varying in width across the input space. For example, in J and J+ methods, for any future unseen test samples, we will have the same set of non-conformity scores calculated in Eq. 6. This may be undesirable from the perspective of uncertainty quantification, as the size of the predictive regions should vary depending on the difficulty of predicting each sample in the test set. For example, more challenging test samples should have larger prediction intervals, while easier test samples should have smaller intervals.

To address this issue, researchers have been exploring ways to improve the adaptivity of conformal prediction methods. This includes developing new non-conformity measures that can better capture the difficulty of predicting individual samples, as well as methods for adapting the prediction intervals to specific types of data, such as time series data or data with complex dependencies. This work is ongoing and is an active area of research within the field of conformal prediction [18], [20], [38], [39].

On the other hand, the main assumption behind conformal prediction is that data are exchangeable and distributed identically. However, this assumption is often violated in practice, as the distribution of data may vary depending on the specific context or domain. In these circumstances, all

training data points should not be treated equally. Some training points should have been considered more relevant and given greater weights to modify the non-conformity scores that are used to generate the prediction intervals. In this approach, the non-conformity scores are calculated based on the absolute residual error between the actual and predicted values of the sample. However, instead of using a fixed weight for all samples, data-dependent weights are used that are based on the distances of the training instances to the test sample. This approach allows the prediction intervals to be adaptive and to vary in width according to the difficulty of predicting each sample in the test set. Using data-dependent weights to modify the non-conformity scores can improve the adaptivity of conformal prediction methods and provide more accurate and reliable prediction intervals. It allows the method to better capture the variations in the distribution of data and to generate intervals that are more informative and relevant to the specific context or domain.

To be formal, we consider the input data set  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  comprising  $n$  samples in the ambient space  $\mathbb{R}^d$ . We compute the similarity between each training data point and the test sample using the metric  $\rho$ , which is inversely proportional to the distance between them

$$\rho(\mathbf{x}_i, \mathbf{x}_{n+1}) := \frac{1}{\|\mathbf{x}_i - \mathbf{x}_{n+1}\|_2}. \quad (11)$$

This step requires just a single pass over the data stream, and the computational cost scales linearly with the data size  $n$ . Using the computed distances in Eq. (11), we propose the following weighted score for each data point with respect to the test sample  $\mathbf{x}_{n+1}$ :

$$w_i := \frac{\rho(\mathbf{x}_i, \mathbf{x}_{n+1})}{\frac{1}{n} \sum_{i'=1}^n \rho(\mathbf{x}_{i'}, \mathbf{x}_{n+1})}, i = 1, \dots, n. \quad (12)$$

In our method, we normalize the similarity of each training sample to the test sample by using the average value of  $\rho$  in the denominator (as shown in Eq. (12)). This ensures that the sum of weights for all  $n$  training samples is equal to  $n$ , similar to methods J and J+. However, our method differs in that the weights are not all equal to 1, but rather are determined based on the proximity of each training sample to the test sample. To further explain this point, a hypothetical scenario would be that we have 100 samples, all of which are located at the same distance from the test sample. In this case, methods J+ and WJ+ should work identically and  $w_i$  must be equal to 1 for all samples. In other words, the numerator and denominator in Eq. (12) are equal and  $w_i = 1, i = 1, \dots, n$ .

Hence, using the corrective weighting scheme, we assign greater weights to data points that are closer to the test sample, and lower weights to data points that are farther away. In the following, we present how we can use data-dependent weights to compute non-conformity scores and construct prediction intervals. Our study only focuses on J+ and CV+; however, this can be applied to any other conformal prediction methods.

**A. J+ WITH WEIGHTED NON-CONFORMITY SCORE (WJ+)**

With WJ+, we use data-dependent weights in computing non-conformity scores in the J+ method. That is, we compute the weight of each data point based on the distance between that data point and the future unseen test sample. In this method, higher weights are given to data points closest to the test point, and lower weights are given to data points farther away from the test point. This means that residuals resulting from data points close to the test point will have a greater effect on the prediction interval.

According to what we have for J+ in the Section II, we denote a prediction interval using J+ with weighted leave-one-out non-conformity score method as  $\hat{C}_{\alpha,n}^{WJ+}$ . On the new test sample, it can be defined as

$$\hat{C}_{\alpha,n}^{WJ+}(\mathbf{x}_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}_{-i}(\mathbf{x}_{n+1}) - (w_i R_i^{LOO}) \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}_{-i}(\mathbf{x}_{n+1}) + (w_i R_i^{LOO}) \right\} \right], \quad (13)$$

where  $R_i^{LOO}$  and  $w_i$  are shown in Eq. 6 and Eq. 12.

**B. CV+ WITH WEIGHTED NON-CONFORMITY SCORE (WCV+)**

According to what we have for CV+ in the Section II, we denote a prediction interval using CV+ with weighted out-of-fold non-conformity score (WCV+) method as  $\hat{C}_{\alpha,n,K}^{WCV+}$ . For the new test sample, it can be defined as

$$\hat{C}_{\alpha,n,K}^{WCV+}(X_{n+1}) = \left[ q_{\alpha,n}^- \left\{ \hat{f}_{-S_{k(i)}}(\mathbf{x}_{n+1}) - (w_i R_i^{CV}) \right\}, \right. \\ \left. \times q_{\alpha,n}^+ \left\{ \hat{f}_{-S_{k(i)}}(\mathbf{x}_{n+1}) + (w_i R_i^{CV}) \right\} \right], \quad (14)$$

where  $k(i) \in \{1, \dots, K\}$  identifies the subset that contains  $i$ , i.e.,  $i \in S_{k(i)}$ . Finally,  $R_i^{CV}$  and  $w_i$  are defined in Eq. 8 and Eq. 12.

Given the challenging nature of real-world data or data sets coming from engineering, there is likely no equal variance across the levels of independent variables. This is known as heteroscedasticity, which violates one of the most fundamental assumptions of regression analysis, homoscedasticity, where we assume that the variance of the error term (the residual term) in a regression model is constant. Furthermore, the main assumption of conformal prediction regarding the identical distribution of data is often violated. Under such circumstances, it is necessary not to treat all observations equally. Some observations in the training set are more relevant to the test sample. In contrast to assuming that all data points are treated equally, this method seems to be more robust to distributional changes in different regions of the training set.

In assessing the performance of our adaptive conformal prediction intervals, we focus on three key metrics: Effective Coverage, Conditional Coverage, and Prediction Interval Width. Effective Coverage quantifies the proportion of instances where the true value falls within our predicted intervals across all predictions, offering a measure of our

model’s overall accuracy. Conditional Coverage evaluates the adaptability of our prediction intervals, ensuring they adjust appropriately to the difficulty of each prediction scenario. Lastly, Prediction Interval Width measures the range within which we expect the true value to fall, with narrower intervals indicating more precise predictions and wider intervals reflecting greater uncertainty. These metrics collectively provide a comprehensive view of our method’s reliability and precision in uncertainty quantification.

**IV. EXPERIMENTAL RESULTS**

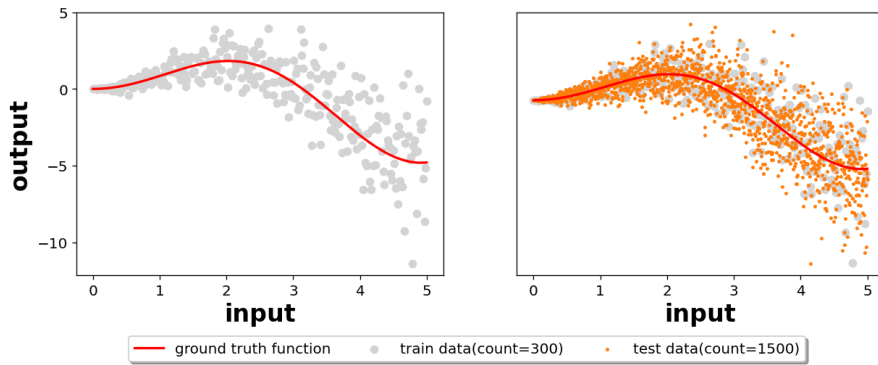
**A. SYNTHETIC DATA SET**

We generate 1-dimensional (1D) heteroscedastic noisy data when the actual signal is defined by a ground truth function  $f(\mathbf{x}) = \mathbf{x} \sin(\mathbf{x})$ . The output variable is generated independently from  $y = f(\mathbf{x}) + \epsilon$ , where  $\epsilon$  is the heteroscedastic noise that is assumed to increase linearly with input values. However, it is important to note that this is just one specific example, and in general, heteroscedastic noise can occur in a variety of forms and may not necessarily increase linearly with input values.

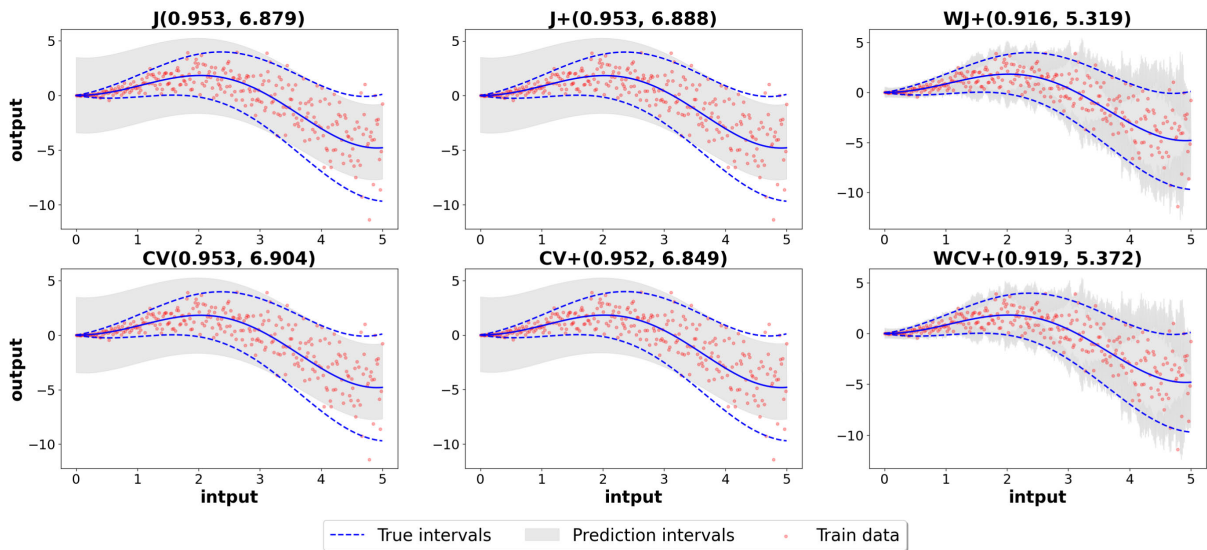
To be specific, we generate 300 train points and 1500 test points in the interval (0, 5) as shown in Fig. 2.

We use the ridge regression algorithm with polynomial features of degree 2 and compare the prediction intervals obtained from the methods including J, J+, WJ+, CV, CV+, and WCV+ at the target coverage level 0.95. Fig. 3 represents the prediction interval bounds as well as the true interval bounds of the test points. Here, to obtain a 95% confidence for our prediction intervals, we set  $\alpha = 0.05$ . In each sub-figure, the title shows the method used along with the effective coverage and average prediction interval width of that method in parentheses. This figure shows that the effective coverage levels of WJ+ and WCV+ are slightly lower than those of the other methods, which may not be desirable at first glance. However, they result in smaller average prediction interval widths. The interesting observation here is that WJ+ and WCV+ offer solutions that adjust the prediction intervals to the local noise while the prediction intervals obtained from the other four methods appear to be constant over different regions in the input space. As noise levels are low, the prediction intervals derived from WJ+ and WCV+ are tight, and as noise levels are high, the prediction intervals derived from these methods are wide. This means that WJ+ and WCV+ methods closely follow the true width and data heteroscedasticity is taken into account with these methods. Hence, it can be conferred that the prediction intervals of WJ+ and WCV+ are much more accurate than those of the other methods. While the constant high prediction interval widths in J, J+, CV, and CV+ might contribute to high effective coverage, their conditional coverage may not be as reliable as those of WJ+ and WCV+.

Fig. 4 compares these six methods when the value of the input variable and the corresponding noise increase. The input interval (0,5) is divided into five bins as shown on the



**FIGURE 2.** Synthetic 1D data uniformly distributed from the given function and heteroscedastic noise that is assumed to increase linearly with input values.



**FIGURE 3.** Comparison of six methods, including J, J+, WJ+, CV, CV+, and WCV+ on test points at the 0.95 target coverage level. The effective coverage and prediction interval width of each method are shown in parentheses. Effective coverage refers to the fraction of test points whose true values are within the prediction intervals. Prediction interval width refers to the average of prediction interval widths over all samples in the test set.

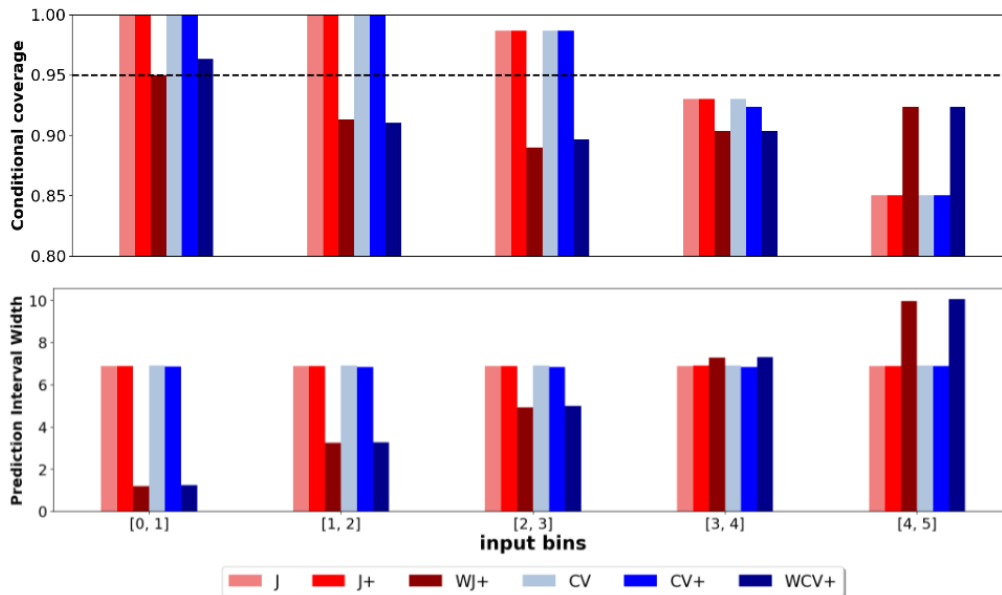
x-axis. In the top sub-figure, the y-axis shows the conditional coverage, and the dashed horizontal line refers to the target coverage level of 0.95. In the bottom sub-figure, the y-axis presents the average prediction interval width. According to Fig. 4, J, J+, CV, and CV+ provide the same prediction interval width over all different bins. In the first two bins, J, J+, CV, and CV+ provide the highest coverage level (1), even greater than the desired coverage (0.95). However, it should be noted that this high confidence comes at the expense of very wide prediction intervals, which is unnecessary. Note that the main advantage of quantifying uncertainty is providing the prediction intervals as tight/informative as they can be while keeping the desired coverage level.

Looking at Fig. 3 and Fig. 4, when the input range is low, the noise is low and there is no need to have very wide prediction intervals which shows the high uncertainty in the first bins. As can be seen from Fig. 4, the prediction intervals obtained from WJ+ and WCV+ are much tighter

than those obtained from the other methods while keeping the conditional coverage more or less 0.95 in the first two bins. We also noticed that when the input value and consequently the noise are increased, the prediction interval widths of WJ+ and WCV+ have been increased leading to higher conditional coverage levels. For example, in the last bin which is  $4 < \mathbf{x} < 5$ , where the noise is high, WJ+ and WCV+ perform better than other methods.

Whenever we construct prediction intervals, there is a trade-off between validity and efficiency. As prediction intervals  $\hat{C}_{\alpha,n}$  (obtained from any uncertainty quantification method with the target coverage level  $1 - \alpha$ ) can always be set to be infinitely large to satisfy the validity condition  $\mathbb{P} \left\{ y_{n+1} \in \hat{C}_{\alpha,n}(\mathbf{x}_{n+1}) \right\} \geq 1 - \alpha$ , which is undesirable. Ideally, we want to reduce the size of the predictive region (e.g., the width) as much as possible, provided that the validity condition is met. In this example, the proposed methods WJ+ and WCV+ produce more adaptive intervals





**FIGURE 4.** Six methods, J, J+, WJ+, CV, CV+, and WCV+, were assessed based on conditional coverage and prediction interval width across different input bins (0.5). Bins, depicted on the x-axis, reveal variations in noise levels. In the top sub-figure, the y-axis indicates conditional coverage, with a dashed line denoting the 0.95 target. For the (0,1) interval, representing minimal noise, J, J+, CV, and CV+ achieve full coverage (level 1) surpassing the 0.95 target. However, this comes at the cost of unnecessarily wide prediction intervals. In contrast, WJ+ and WCV+ nearly hit the target coverage, maintaining small average prediction interval widths in the first bin. The bottom sub-figure displays consistent prediction intervals for J, J+, CV, and CV+ across all bins, while WJ+ and WCV+ adapt, widening from the first to the last bin with increased data noise. This illustrates their capacity for adaptive prediction intervals across the input space. In the last bin with the highest data noise, WJ+ and WCV+ exhibit wider average prediction interval widths compared to other methods.

than the other methods; therefore, using data-dependent weights in non-conformity scores should be preferred in the presence of heteroscedastic noise.

### B. BENCHMARK DATA SETS

Six real-world regression data sets that we consider in this section are listed in Table 1, describing the data size, the number of (used) features, and the skewness and kurtosis (Pearson) of the response variables. All data sets are publicly available and from DELVE [40], UCI [41], and KEEL [42]. For each data set, we standardize the features to have zero means and unit variances and rescale the responses by dividing them by their mean values.

Fig. 5 represents the histogram of the response variables by counting the number of observations that fall within discrete bins. It also shows smooth curves obtained by using kernel density estimates to provide complementary information about the shape of the distributions. This figure also shows three vertical dashed lines referring to the 10th, 50th, and 90th percentile of the response variables in each data set. As can be seen in Table 1, the skewness and kurtosis of response variables in all data sets are reported. In essence, skewness measures distribution symmetry, while kurtosis measures distribution tail heaviness (heavy-tailed or light-tailed). The data are perfectly symmetrical if skewness = 0; however, having a skewness of exactly zero in real-world data is quite unlikely. As a rule of thumb, the distribution

can be considered approximately symmetric if the skewness is between  $-0.5$  and  $0.5$  (Kinematics, Energy, Wizmir). The distribution of response variables of remaining data sets (Communities, Treasury, and Mortgage) can be called highly skewed (skewness  $> 1$ ). It is possible that the tail region of skewed data may appear as an outlier to the statistical model, and outliers adversely affect the model’s performance, particularly in regression-based models. When kurtosis is high (Communities, Treasury, and Mortgage), distributions tend to have heavy tails or outliers, whereas when kurtosis is low (Kinematics, Energy, Wizmir), distributions tend to have light tails or few outliers.

All data sets used in this study are divided into train and test sets in an 80-20 ratio. As mentioned earlier, the base regression algorithm to quantify uncertainty is ridge regression with polynomial features. In each data set, the degree of the polynomial features and regularization strength is tuned using the hyperparameter optimization technique. This information is provided in Table 1. The goal of the experiments is to evaluate different uncertainty quantification methods discussed in Sections II and III when the target coverage level is set to  $1 - \alpha = 0.95$ .

Table 2 summarizes the effective and conditional coverage levels, as well as the associated prediction interval widths while setting the target coverage level at 0.95. Over the test set, effective coverage levels indicate the proportion of test points whose true values lie within the prediction intervals. To report conditional coverage, the test set is

TABLE 1. Overview of the publicly available data sets.

Data set	# of samples	# of features	Skewness	Kurtosis	Polynomial degree	Regularization strength	Source
Kinematics	8192	8	0.09	2.46	5	10	DELVE
Communities	1994	100	1.521	4.82	2	1	UCI
Energy	768	8	0.36	1.75	3	0.001	UCI
Wizmir	1461	9	0.07	1.76	2	1	KEEL
Treasury	1049	15	1.33	4.74	2	1	KEEL
Mortgage	1049	15	1.02	3.5	2	1	KEEL

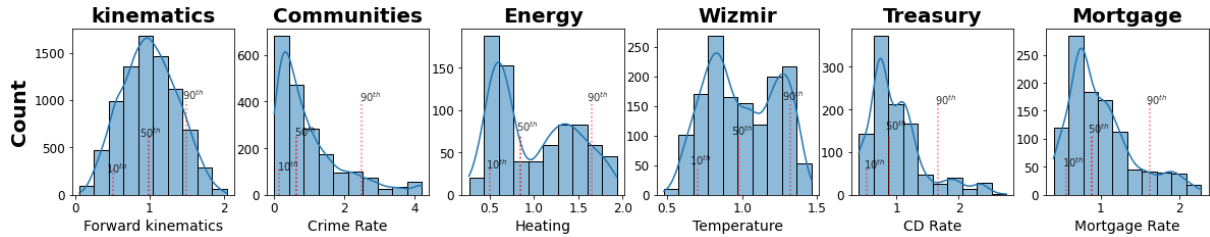


FIGURE 5. Histogram of data sets showing the distribution of the response variables. Kernel density estimation curves are also visualized to show how each distribution is shaped. Vertical dashed lines represent 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentiles of the distributions.

TABLE 2. Reporting the effective coverage and average prediction interval width over the test set. For each data set, the true responses are divided into 10 bins by quantiles, in the test set. The conditional coverage levels (generally the higher the better) and the average prediction interval widths (generally the lower the better) of each bin are presented. For each bin of each data set, the contour color of a column varies from red (worst case) to green (best case). However, each uncertainty quantification method should consider both coverage and width at the same time in order to be effective. For each data set, the most effective methods considering both coverage levels and widths are shown in bold.

Data set	method	Effective test set		Conditional																			
				bin1		bin2		bin3		bin4		bin5		bin6		bin7		bin8		bin9		bin10	
		cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid	cov	wid
Kinematics	J	0.95	0.485	0.93	0.485	0.96	0.485	0.94	0.485	0.97	0.485	0.96	0.485	0.97	0.485	0.97	0.485	0.97	0.485	0.96	0.485	0.96	0.485
	J+	0.95	0.485	0.93	0.485	0.96	0.485	0.94	0.485	0.97	0.485	0.96	0.485	0.97	0.485	0.97	0.485	0.97	0.485	0.96	0.485	0.96	0.485
	<b>WJ+</b>	<b>0.96</b>	<b>0.491</b>	0.93	0.499	<b>0.95</b>	0.495	0.95	0.493	0.97	0.491	0.96	0.489	0.98	0.489	0.97	0.488	0.98	0.489	0.97	0.488	0.96	0.487
	CV	0.96	0.495	0.93	0.495	0.96	0.495	0.94	0.495	0.97	0.495	0.96	0.495	0.98	0.495	0.97	0.495	0.98	0.495	0.97	0.495	0.96	0.495
	<b>WCV+</b>	<b>0.96</b>	<b>0.508</b>	0.94	0.513	0.96	0.51	0.96	0.512	0.98	0.508	0.96	0.509	0.98	0.506	0.97	0.504	0.98	0.503	0.97	0.505	0.97	0.512
Communities	J	0.96	2.543	1	2.543	1	2.543	1	2.543	1	2.543	1	2.543	0.948	2.543	0.976	2.543	0.93	2.544	1	2.544	0.75	2.544
	J+	0.96	2.544	1	2.544	1	2.543	1	2.544	1	2.544	1	2.544	0.948	2.544	0.976	2.543	0.93	2.544	1	2.544	0.75	2.544
	<b>WJ+</b>	<b>0.97</b>	<b>2.373</b>	1	2.093	1	2.09	1	2.269	1	2.246	1	2.3	1	2.317	0.976	2.357	0.91	2.626	0.97	2.589	0.8	2.82
	CV	0.96	2.49	1	2.49	1	2.49	1	2.49	1	2.49	0.948	2.49	0.976	2.49	0.93	2.49	0.97	2.49	0.97	2.49	0.75	2.49
	<b>WCV+</b>	<b>0.97</b>	<b>2.671</b>	1	2.488	1	2.488	1	2.497	1	2.488	1	2.492	1	2.542	1	2.563	1	2.856	0.97	2.897	0.75	3.394
Energy	J	0.90	0.086	1	0.086	1	0.086	0.93	0.086	1	0.086	0.8	0.086	0.8	0.086	1	0.086	0.8	0.086	0.8	0.086	0.75	0.086
	J+	0.92	0.086	1	0.086	1	0.086	0.93	0.086	1	0.086	0.8	0.086	0.8	0.086	1	0.086	0.93	0.086	0.8	0.086	0.81	0.086
	<b>WJ+</b>	<b>0.98</b>	<b>0.15</b>	1	0.095	1	0.089	0.93	0.097	1	0.113	0.86	0.131	1	0.211	1	0.166	1	0.167	1	0.203	1	0.23
	CV	0.90	0.086	1	0.086	1	0.086	0.93	0.086	1	0.086	0.8	0.086	0.8	0.086	1	0.086	1	0.086	0.8	0.086	0.75	0.086
	<b>WCV+</b>	<b>0.94</b>	<b>0.095</b>	1	0.094	1	0.094	0.93	0.094	1	0.094	0.93	0.094	0.8	0.095	1	0.095	1	0.095	0.87	0.095	0.81	0.095
Wizmir	J	0.96	0.072	0.93	0.072	0.93	0.072	0.89	0.072	0.96	0.072	1	0.072	0.97	0.072	1	0.072	0.93	0.072	1	0.072	0.93	0.072
	J+	0.96	0.072	0.93	0.072	0.93	0.072	0.89	0.072	0.96	0.072	1	0.072	0.97	0.072	1	0.072	0.93	0.072	1	0.072	0.93	0.072
	<b>WJ+</b>	<b>0.98</b>	<b>0.086</b>	0.96	0.08	0.96	0.083	0.93	0.084	0.96	0.082	1	0.079	1	0.085	1	0.089	1	0.089	1	0.095	0.96	0.091
	CV	0.96	0.072	0.93	0.72	0.93	0.72	0.89	0.72	0.96	0.72	1	0.72	0.97	0.72	1	0.72	0.93	0.72	1	0.72	0.93	0.72
	<b>WCV+</b>	<b>0.96</b>	<b>0.074</b>	0.93	0.074	0.93	0.075	0.93	0.076	0.96	0.074	1	0.074	0.97	0.074	1	0.074	0.96	0.074	1	0.074	0.93	0.074
Treasury	J	0.94	0.109	1	0.109	1	0.109	1	0.109	1	0.109	1	0.109	0.95	0.109	1	0.109	1	0.109	0.81	0.109	0.67	0.109
	J+	0.94	0.109	1	0.109	1	0.11	1	0.109	1	0.109	1	0.109	0.95	0.109	1	0.109	1	0.109	0.81	0.109	0.67	0.109
	<b>WJ+</b>	<b>0.96</b>	<b>0.109</b>	1	0.069	1	0.095	0.95	0.083	0.95	0.08	0.95	0.101	0.95	0.101	0.95	0.112	1	0.103	0.91	0.148	0.95	0.199
	CV	0.94	0.109	1	0.109	1	0.109	1	0.109	1	0.109	0.95	0.109	0.95	0.109	1	0.109	1	0.109	0.81	0.109	0.66	0.109
	<b>WCV+</b>	<b>0.97</b>	<b>0.125</b>	1	0.081	1	0.11	1	0.094	0.95	0.095	1	0.123	0.95	0.116	0.95	0.131	1	0.119	0.91	0.165	0.95	0.218
Mortgage	J	0.95	0.044	1	0.044	1	0.044	1	0.044	1	0.044	1	0.044	0.86	0.044	1	0.044	1	0.044	0.76	0.044	0.86	0.044
	J+	0.95	0.044	1	0.044	1	0.044	1	0.044	1	0.044	1	0.044	0.86	0.044	1	0.044	1	0.044	0.76	0.044	0.86	0.044
	<b>WJ+</b>	<b>0.97</b>	<b>0.049</b>	1	0.039	0.95	0.037	1	0.033	1	0.042	1	0.041	0.86	0.038	1	0.042	1	0.055	0.95	0.072	0.95	0.094
	CV	0.95	0.045	1	0.045	1	0.045	1	0.045	1	0.045	1	0.045	0.86	0.045	1	0.045	1	0.045	0.76	0.045	0.86	0.045
	<b>WCV+</b>	<b>0.95</b>	<b>0.047</b>	1	0.047	1	0.047	1	0.047	1	0.047	1	0.047	0.86	0.047	1	0.047	1	0.047	0.76	0.047	0.86	0.05

divided into 10 bins by quantiles based on the true values of the response variables. Our next step is calculating the average coverage and prediction interval width per bin. Within each bin of a data set, the contour color varies from red (the worst case) to green (the best case). When evaluating uncertainty quantification methods, both coverage and width should be considered simultaneously. For each data set, the best methods considering both coverage levels and widths are shown in bold. For example, looking at Communities, in the first five bins, all six methods provide coverage 1 meaning that all test samples in these bins are covered.

However, the interesting point is that the corresponding prediction interval widths obtained by WJ+ and WCV+ are far narrower than those obtained by other methods. In bin 6, the coverage of these methods remains the same (1), while having tighter widths compared to other methods. Even though the prediction interval widths of other methods are wider in bin 6, their coverage rate drops from 1 to 0.95, but CV does not change. CV, however, cannot compete with WJ+ and WCV+ in this bin due to its large width. The trend can be seen up to the last bin, where WJ+ and WCV+ have better performance than other methods. Compared with other

methods, these methods have relatively wider widths, which is essential for this bin to achieve better conditional coverage levels (almost 0.8).

As can be seen from Table 2, across all bins, the average prediction interval widths obtained from the J method and its cross-validation version CV method are the same. Compared to J and CV methods, J+ and CV+ methods provide slightly different prediction intervals in different bins; however, they are not as adaptive as WJ+ and WCV+. The prediction interval widths of WJ+ and WCV+ can differ significantly over different bins based on the level of uncertainty associated with each bin. To confirm these claims, we select two data sets, Kinematics and Treasury, and plot the conditional coverage and interval width on different bins separated by quantiles. In Kinematics, the distribution of the response variable is symmetric (low skewness and low kurtosis), while in Treasury, the distribution of the response variable is asymmetric with heavy tails due to high skewness and high kurtosis.

Fig. 6 illustrates the true values vs. prediction intervals of response variables of test points. Two data sets, namely, Kinematics and Treasury, and six methods including J, J+, WJ+, CV, CV+, and WCV+ at the target coverage level of 0.95 are used. The top and bottom sub-figures refer to Kinematics and Treasury data sets, respectively. Effective coverage and average prediction interval width of each method are shown in parentheses. As Kinematics is a relatively large data set, we only plot 8% of the observations in order not to crowd the plot. According to Fig. 5, the distribution of the response variable in the Kinematics data set can be considered a Gaussian distribution. This may lead to a situation in which symmetric regions can be found in the data set. Under such circumstances, there should be no noticeable differences among the methods used in this study. As can be seen from Fig. 6, in this data set, the prediction intervals obtained by J, J+, and WJ+ are approximately the same, while the cross-validation versions, CV, CV+, and WCV+ provide slightly wider prediction intervals, resulting in higher effective coverage levels.

In the Treasury data set, the distribution of the response variable appears to be highly skewed, as shown in Fig. 5. This may lead to the occurrence of dense and sparse regions in different portions of the data set. It may be possible to achieve better uncertainty quantification in this scenario by using data-dependent weights when constructing non-conformity scores. According to Fig. 6, we observe that the prediction intervals for the WJ+ and WCV+ methods are more adaptable than the intervals for the other methods in the Treasury data set. The prediction intervals seem to increase as the response variable (CD Rate) increases. It is also noteworthy that while J, J+, and WJ+ have the same average prediction interval width of 0.109, the effective coverage of WJ+ is higher than those of J, J+, indicating that WJ+ provides superior performance. The same observation can be seen in the results of CV, CV+, and WCV+. Looking at this figure, it is interesting to note that the WJ+ and WCV+

methods result in fewer red points lying outside prediction intervals than those using other methods.

Fig. 7 shows conditional coverage and prediction interval width on test points using six methods at the target coverage level of 0.95. The top and bottom sub-figures refer to Kinematics and Treasury data sets, respectively. As mentioned earlier, in the test set, the true values of the response variables are divided into 10 bins by quantiles. In each sub-figure, the x-axis shows these 10 bins. In the top and bottom sub-figures, the y-axes represent conditional coverage and prediction interval width. As previously discussed, all six methods perform similarly across all bins in the Kinematics data set. In different bins, almost all methods appear to have conditional coverage at the target coverage (0.95). Furthermore, the prediction intervals obtained from these methods are nearly identical across bins.

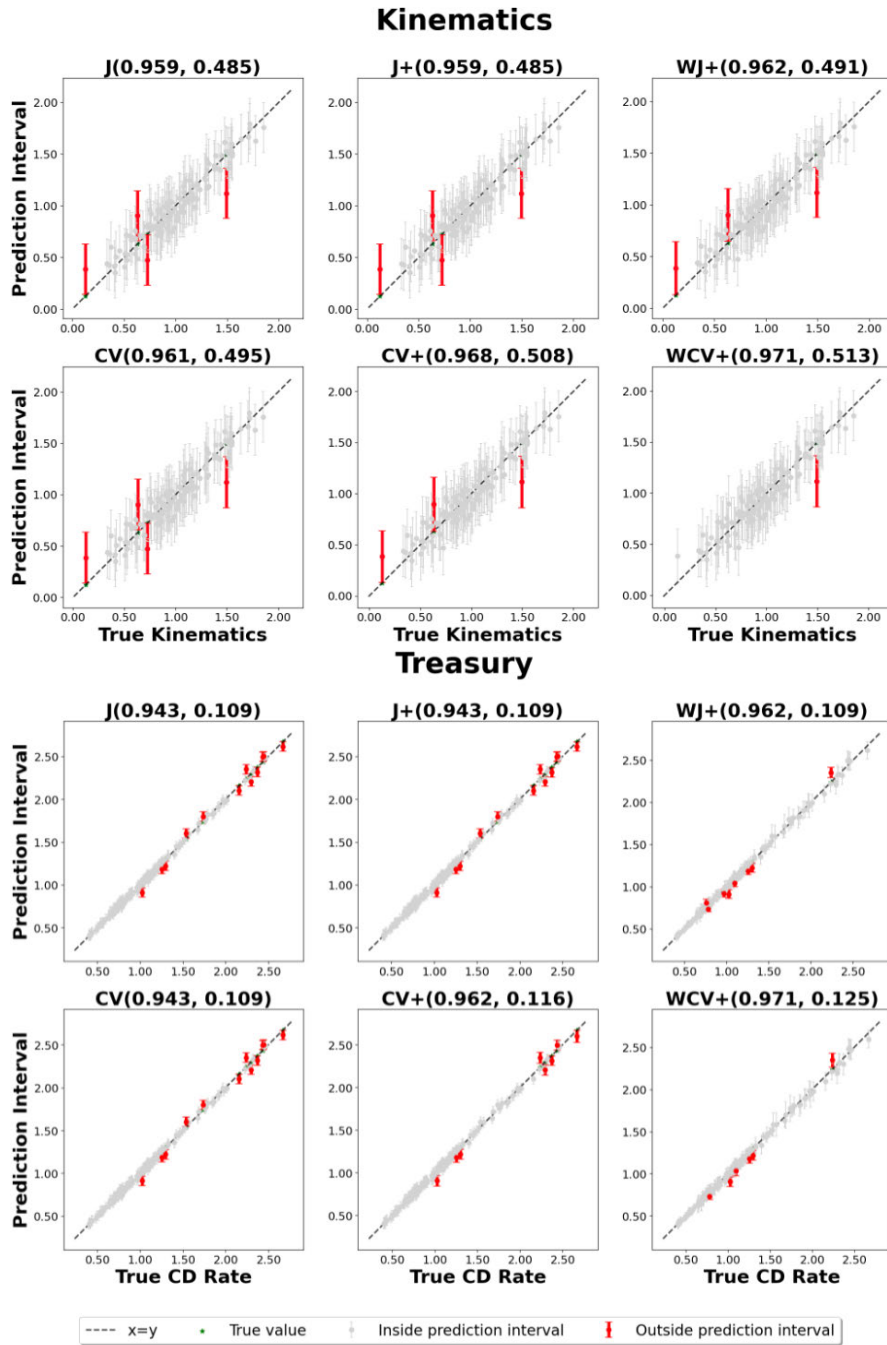
On the other hand, in the Treasury data set, the effectiveness of using data-dependent weights in non-conformity scores in the WJ+ and CV+ methods is demonstrated well. Within the first eight bins, all methods have conditional coverage equal to or above the desired coverage (0.95), while WJ+ and WCV+ have much tighter prediction intervals than other methods. In other words, using these methods, the prediction intervals tend to be shorter when the estimator is more certain. In contrast, in the last two bins, where there is significantly more uncertainty, the conditional coverage of WJ+ and CV+ methods is significantly higher than the other methods. It is noteworthy that, while the prediction interval width of all four methods is the same for all bins, the prediction interval widths of WJ+ and CV+ differ for different bins depending on how uncertain the bins are.

So far, we have shown how using data-dependent weights in non-conformity scores help to construct valid prediction intervals. The experiments show that WJ+ and WCV+ outperform other methods in terms of their efficiency and adaptability over the input space.

## V. APPLICATION TO SCIENTIFIC SIMULATION

In addition to benchmarks, we consider a high-rise telecommunication tower as a case study [43], [44], [45], [46]. The height is over 400 meters, made of reinforced concrete. The concrete shaft is the main load-carrying structure of the tower that transfers the lateral and gravitational loads to the foundation. We consider several modeling aspects, including material non-linearities (i.e., cracking, crushing, and damage), and geometric non-linearities.

We develop a 2D finite element model of the tower, including the head structure, shaft, and transition. A total of 10 random models are generated using Latin Hypercube Sampling (LHS), to consider the epistemic variability in 18 material/modeling parameters (concrete, steel, and system level). Moreover, 100 ground motions are used to account for aleatory uncertainty. Since a ground motion record has a temporal nature, a series of scalar meta-features should be extracted to be used in the context of machine learning regression [47]. For each ground motion, we extract

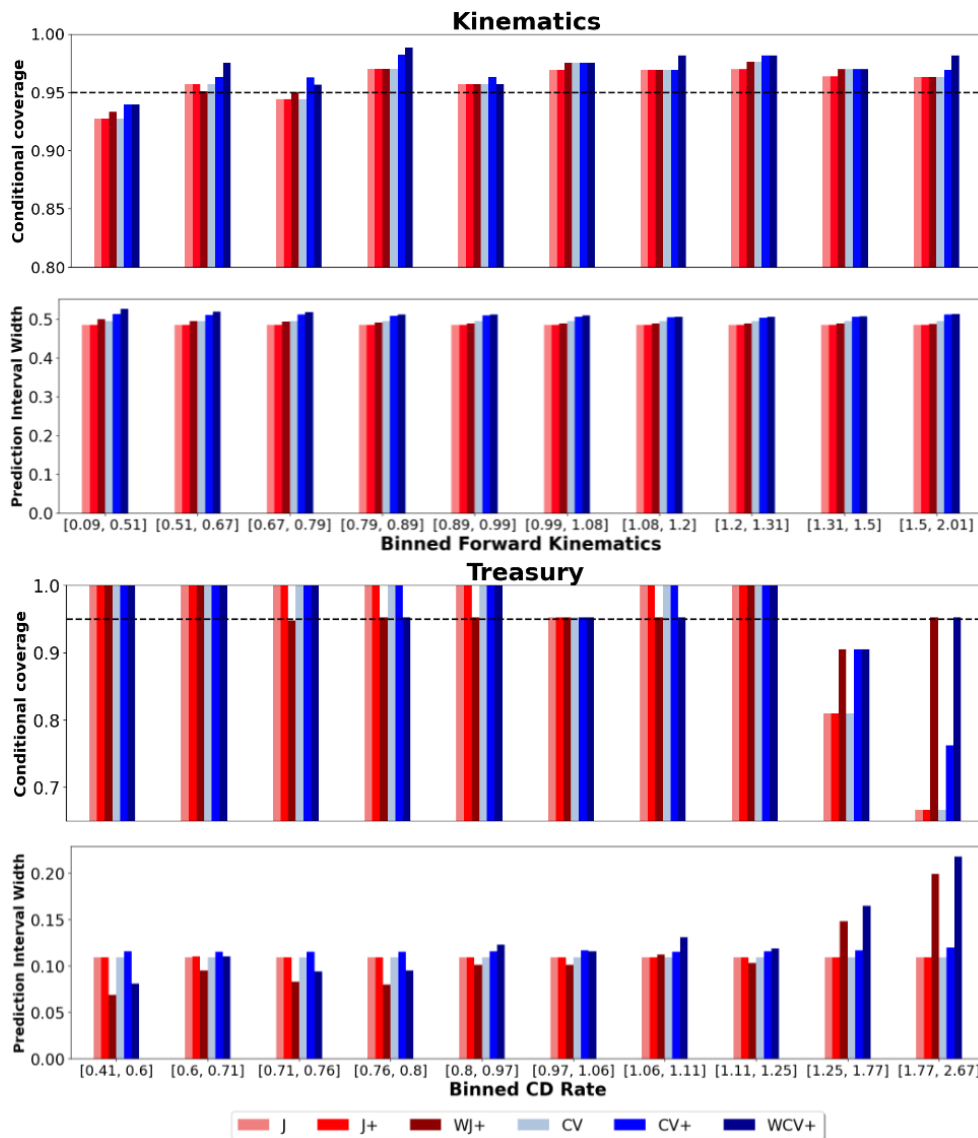


**FIGURE 6.** True values vs. prediction intervals of response variables of test points using six methods, including J, J+, WJ+, CV, CV+, and WCV+ at the target coverage level 0.95. The top and bottom sub-figures refer to Kinematics and Treasury data sets, respectively. For the Kinematics data set, only 8% of the observations are used for the visibility of the plot. The effective coverage and prediction interval width of each method are shown in parentheses.

31 intensity measure parameters, including all peak values (e.g., peak ground acceleration - PGA), intensity-, frequency-, and duration-dependent parameters. The combination of 100 ground motions and 10 modeling samples yields 1,000 unique assessments. To account for higher seismic intensity levels (and possibly failure mechanism), three scale factors are also considered. Overall, we create a data set containing  $n = 3,000$  simulations with  $d = 49$

attributes. The output space for the regression analysis represents a structural response, base shear. We present a schematic 3D finite element model and the histogram of structural response in Fig. 8.

We compare the predictive regions produced by conformal prediction methods with those produced by the Gaussian Process Regression algorithm (GPR) [48], which is one of the most popular Bayesian machine learning approaches.

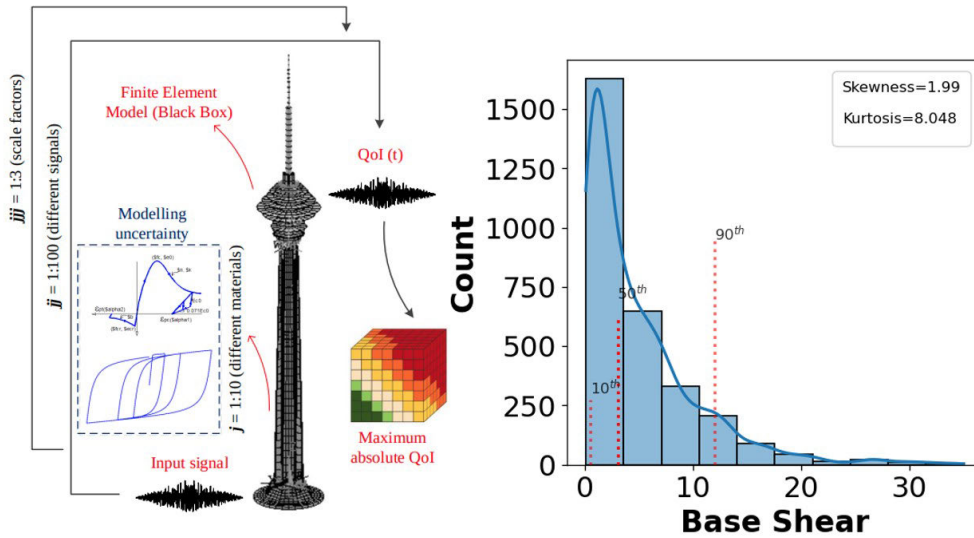


**FIGURE 7.** Conditional coverage and prediction interval width of six methods, including J, J+, WJ+, CV, CV+, and WCV+ at target coverage level of 0.95 on different bins of test points. The top and bottom sub-figures refer to Kinematics and Treasury data sets, respectively.

GPR models the distribution of possible functions that could have generated the data, rather than just the mean or median of this distribution. This allows it to provide uncertainty estimates for its predictions by modeling the full distribution of the prediction. To further explore the efficiency of our proposed method, we also compared it with Conformalized Quantile Regression (CQR) [20]. CQR combines conformal prediction with classical quantile regression and offers the benefits of both approaches. Our comparative analysis shows that our proposed method tends to produce shorter intervals than CQR. We present the results of our comparison by first reviewing GPR and then comparing its performance with conformal prediction methods on the scientific data set. Subsequently, we discuss CQR and compare the performance of our proposed method, WCV+, with CQR on the same data set.

**A. REVIEW OF GPR**

From a conventional regression perspective  $y = \hat{f}(x, \beta) + e$ , we estimate  $\beta$ , a vector of unknown parameters, using several tools, e.g., ordinary least squares (OLS), based on the observed input-output pairs. In this case, we end up with a set of fixed parameters  $\beta$  resulting in a fixed function  $\hat{f}$ . However, from the GP perspective, not only function  $\hat{f}$  is not fixed but also it is unknown/stochastic and considered a major source of uncertainty. GP extends the idea of a Gaussian distribution over discrete random variables to the concept of a Gaussian distribution over continuous functions, with inference occurring directly in the function space. To put it another way, Gaussian distribution is over random vectors, while GP is over random functions. For more details, please see [49] and [50]. GP is used in uncertainty quantification since it can take both the mean



**FIGURE 8.** Schematic diagram of tower model and a histogram of the response variable, i.e., base shear. The skewness and kurtosis of base shear are also shown in the legend. Vertical dashed lines represent 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentiles of the distributions.

and covariance of the distribution into account. GP has been extensively used in many machine-learning tasks, including regression, classification, and clustering. However, in this paper, we mainly focus on the use of GP in regression settings. In the following, we show how GPR works.

To be formal, given  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , we describe  $\hat{f}(\mathbf{x})$  by GP as  $\hat{f}(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ , where a mean function  $m(\mathbf{x})$  and a covariance (kernel) function  $\kappa(\mathbf{x}, \mathbf{x}')$  are defined as  $m(\mathbf{x}) = \mathbb{E}[\hat{f}(\mathbf{x})]$  and  $\kappa(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(\hat{f}(\mathbf{x}) - m(\mathbf{x}))(\hat{f}(\mathbf{x}') - m(\mathbf{x}'))]$ , respectively. The prior expected function value is often set to  $m(\mathbf{x}) = 0$ , since we assume that we have no initial knowledge about all functions in the distribution. So, the kernel function  $\kappa(\mathbf{x}, \mathbf{x}')$  which reflects the relationships between function values at each input pair  $(\mathbf{x}, \mathbf{x}')$ , only define the GP. Here, we have  $\hat{f}(\mathbf{x}) \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ .

All assumptions about underlying functions and generalization properties of a GP model lie in the specifying of the kernel function. Kernel structures can determine modeling assumptions in terms of smoothness, linearity, or periodicity expected in the data. The appropriate selection of kernel function plays an important role in the performance of GPR. Some powerful and widely used kernels in GPR with a smoothing parameter  $\sigma_b > 0$  (called bandwidth) are shown below:

- **constant kernel** is defined as  $\kappa(\mathbf{x}, \mathbf{x}') = \sigma_b^2$
- **linear kernel** is defined as  $\kappa(\mathbf{x}, \mathbf{x}') = \sigma_b^2 \mathbf{x}^\top \mathbf{x}'$
- **squared exponential or radial basis function (RBF) kernel** is defined as  $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\nu^2(\mathbf{x}, \mathbf{x}'))$ , where  $\nu(\mathbf{x}, \mathbf{x}') := \|\mathbf{x} - \mathbf{x}'\|_2 / \sigma_b$ .

In the GPR setting, we have  $\hat{y} = \hat{f}(\mathbf{x}) + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . Here,  $\epsilon$  is independent and identically distributed random errors with zero mean and unknown variance  $\sigma^2$ . Since the GP prior function  $\hat{f}$  over observations is assumed to

be a Gaussian process, the GP posterior function is a Gaussian process too.

Assume we want to predict the value of output  $y_{n+1}$ , for the test sample  $\mathbf{x}_{n+1}$ , given  $n$  training data points. Here, we can have  $\hat{y}_{n+1} \sim \mathcal{N}(\hat{f}(\mathbf{x}_{n+1}), \sigma^2(\mathbf{x}_{n+1}))$ . Considering the distribution of function value at sample point is Gaussian with mean  $m(\mathbf{x}_{n+1})$  and variance  $\sigma^2(\mathbf{x}_{n+1})$ :

$$m(\mathbf{x}_{n+1}) = \kappa(\mathbf{x}_{n+1}, \mathcal{X}) (\sigma^2 \mathbf{I} + \kappa(\mathcal{X}, \mathcal{X}))^{-1} \mathbf{y}, \quad (15)$$

$$\begin{aligned} \sigma^2(\mathbf{x}_{n+1}) &= \kappa(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}) \\ &\quad - \kappa(\mathbf{x}_{n+1}, \mathcal{X}) (\sigma^2 \mathbf{I} + \kappa(\mathcal{X}, \mathcal{X}))^{-1} \kappa(\mathcal{X}, \mathbf{x}_{n+1}), \end{aligned} \quad (16)$$

where  $\kappa(\mathbf{x}_{n+1}, \mathcal{X})$  is a kernel matrix between test sample  $\mathbf{x}_{n+1}$  and  $n$  training points  $\mathcal{X}$ ; and  $\kappa(\mathcal{X}, \mathcal{X})$  denotes the kernel matrix of the  $n$  training points  $\mathcal{X}$ . Although GPR can provide us a measure of uncertainty for the prediction, the time complexity of constructing the inversion of kernel matrix  $\kappa(\mathcal{X}, \mathcal{X})$  would be  $\mathcal{O}(n^3)$ , which is computationally intractable for large high-dimensional data on platforms that have limited computing resources. Furthermore, storing the computed kernel matrix needs  $\mathcal{O}(n^2)$  storage space, posing a significant challenge for large data sets.

### B. COMPARISON OF CONFORMAL PREDICTION METHODS WITH GPR

In this section, we compare the predictive regions produced by conformal prediction with GPR methods in our scientific data. In conformal prediction, a ridge regression algorithm with polynomial features is employed as an underlying regression algorithm. The degree of the polynomial features and regularization strength are set to 2 and 10, respectively, using hyperparameter optimization. We also apply GPR

with two different choices of kernel functions. In the following, GPR refers to the use of the default kernel in `scikit-learn` package (Version 1.0.1), whereas, in GPR\* the kernel hyperparameters are optimized.

We use 80% of data for the training set (size = 2,400) and 20% of data for the test set (size = 600) and provide the results based on the test set. Fig. 9 illustrates the effective coverage at three target coverage levels (0.9, 0.95, 0.99), as well as the prediction interval width for test points using three conformal prediction methods including CV, CV+, and WCV+ and the two GPR methods. To improve readability, we only show the CV family in this figure and not the J family. Across all target coverage levels, GPR without kernel optimization performs the worst among all methods. While GPR\* with an optimized kernel performed better than GPR, it is still inferior to the six conformal prediction methods. Despite the shorter prediction intervals obtained by GPR than by conformal prediction methods, they do not meet the target coverage levels. Accordingly, GPR methods are less reliable than conformal prediction methods.

Besides, users may believe that the choice of a kernel function determines almost all the generalization properties of GPs. However, it is important to note that we are dealing with a black box model. This means that the user may not be an expert, or may not have a deep understanding of the data or the modeling challenge. In these cases, the option to select the proper kernel function requires an extensive hyperparameter optimization step, which can be a very time-consuming task. The WCV+ method achieves higher effective coverage levels than other methods, even when their prediction intervals are smaller. Results indicate that the proposed method WCV+ is more effective at quantifying uncertainty than other methods. Using weighted non-conformity scores when constructing prediction intervals allows us to be more accurate/valid (higher coverage) while being more efficient (smaller prediction intervals).

In the following, figures are presented only when the target coverage level is set to 0.95 in order to save space and avoid repetition. Fig. 10 shows true values vs. prediction intervals of test points using different conformal prediction and GPR methods. Elements in the parentheses indicate the effective coverage, average prediction interval width, and the number of red missed points outside of prediction intervals for each method. This figure clearly illustrates the effectiveness of the proposed methods WJ+ and WCV+ in comparison with all other methods. According to this figure, with WCV+, the number of missed points whose prediction intervals do not cover the true values is less than the number of missed points for other methods. Additionally, it is interesting to note that the average prediction interval widths obtained by WJ+ and WCV+ are significantly smaller than those obtained by other methods except GPR, which is not reliable at all. Despite the tightness of prediction intervals obtained by these methods, effective coverage is not sacrificed at all. Using weighted non-conformity scores when constructing prediction intervals

allows us to be more accurate/valid (higher coverage) while being more efficient (shorter prediction intervals).

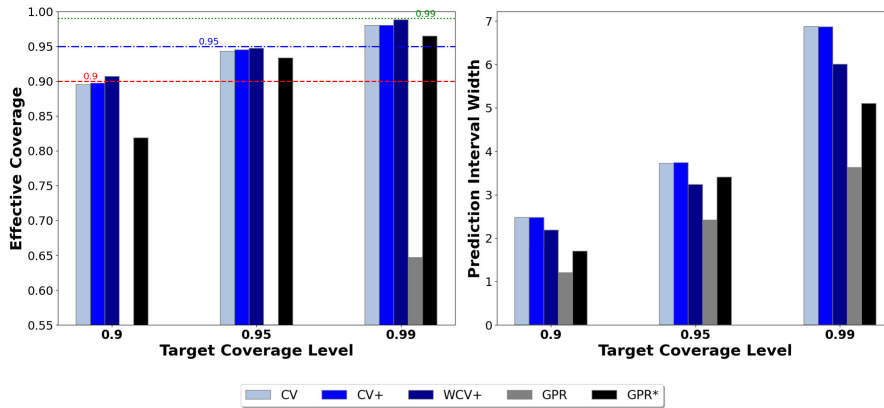
Fig. 11 shows conditional coverage and prediction interval width on test points using different conformal prediction and GPR methods. For better readability, only the CV family is shown in this figure and not the J family. Again, quantiles are used to divide the true values of the response variable in the test set into ten bins. This figure illustrates that almost all conformal prediction and GPR\* methods have acceptable conditional coverage (0.95 or higher), except for the last two bins. However, the WCV+ method has proven to be highly effective when data-dependent weights are applied to non-conformity scores. Compared to the other methods, this method produces much tighter prediction intervals. Looking at Fig. 11, as we move from bin 1 to bin 10, we can observe an increase in the widths of the prediction intervals for WCV+. Hence, when there is less uncertainty in some regions of the data set (here, first few bins), the prediction intervals obtained from WCV+ are as tight and informative as possible. The increasing trend of prediction interval widths using WCV+ over 10 bins shows how adaptive these methods are when dealing with uncertainty. In the last bin, when uncertainty is greatest, this method results in the widest prediction intervals and the highest conditional coverage.

It should be noted that the predictive regions produced by GPR are not valid and therefore they are unreliable if the correct prior is not available. Looking at this figure and Fig. 9, the tightness of prediction intervals of GPR comes at the cost of low effective/conditional coverage rates. This problem can be alleviated by using the appropriate kernel function. However, choosing the right kernel is time-consuming, and there is no guarantee that it will be the most optimized. On the contrary, as shown in this section, conformal prediction methods produce valid predictive regions even without the need to optimize any hyperparameters. These methods can also be applied to any underlying regression algorithm. According to the results of the experiments conducted in this section, conformal prediction methods using data-dependent non-conformity scores are superior to other commonly used methods in terms of validity and efficiency.

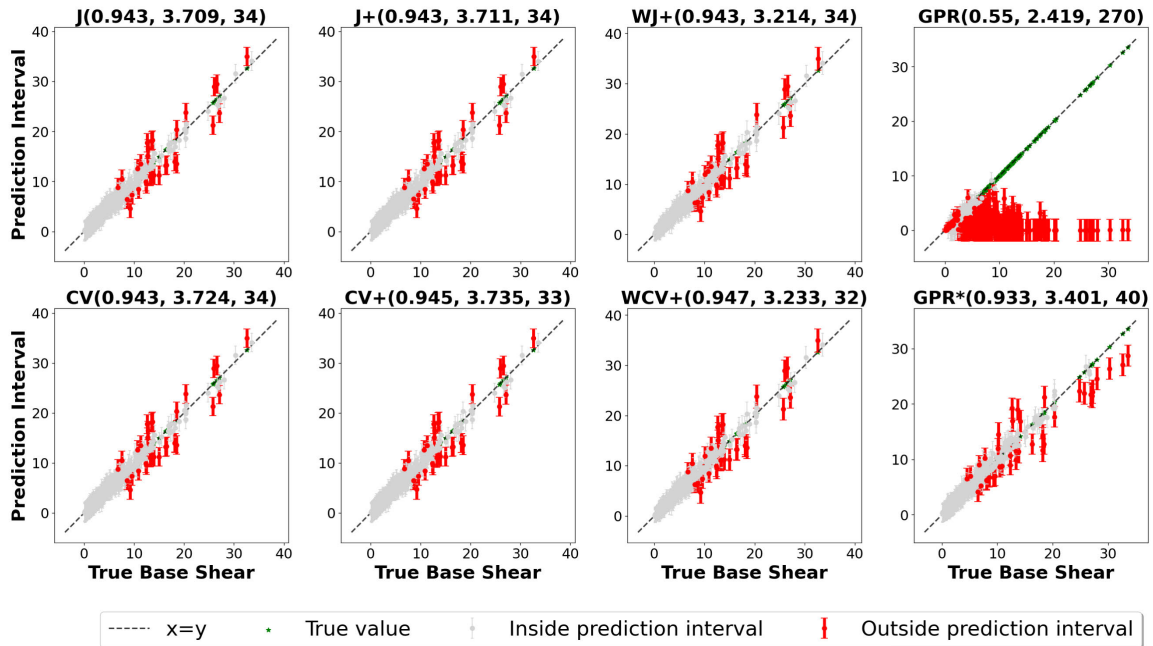
### C. REVIEW OF CQR

The conformalized quantile regression method (CQR) [20] is a prediction interval method that takes into account heteroscedasticity in the data to construct narrower and more accurate prediction intervals. CQR uses quantile regression to estimate the prediction bounds and the residuals from this method are used to create the guaranteed coverage value.

To use CQR to construct a prediction interval for a new test point  $\mathbf{x}_{n+1}$ , we first fit a quantile regression model to the calibration set, with quantile levels  $\alpha_{lo}$  and  $\alpha_{hi}$  corresponding to the lower and upper prediction bounds, respectively. We then predict the quantiles of the response variable at  $\mathbf{x}_{n+1}$  using the fitted quantile regression models, denoted by  $\hat{q}_{\alpha_{lo}}(\mathbf{X}_{n+1})$  and  $\hat{q}_{\alpha_{hi}}(\mathbf{X}_{n+1})$ , respectively.



**FIGURE 9.** Effective coverage and prediction interval width on test points using three conformal prediction methods including CV, CV+, and WCV+ and two GPR methods at three target coverage levels (0.9, 0.95, 0.99). GPR without kernel optimization performs the worst among all methods across all target coverage levels. GPR\* with an optimized kernel performed better than GPR; however, it is still inferior to the six conformal prediction methods. While GPR prediction intervals are shorter than those obtained from conformal prediction methods, they do not meet the required coverage levels. This makes GPR methods less reliable than conformal prediction methods. Despite the tight prediction intervals, WCV+ achieves a higher effective coverage level than other conformal prediction methods at all target coverage levels.



**FIGURE 10.** True values vs. prediction intervals on test points using six conformal prediction methods including J, J+, WJ+, CV, CV+, and WCV+ and two GPR methods at the target coverage level of 0.95. Effective coverage, prediction interval width, and number of red missed test points lying outside of the prediction intervals of each method are shown in parentheses.

The prediction interval is then constructed as follows:

$$\hat{C}_{\alpha,n}^{CQR}(\mathbf{x}_{n+1}) = [\hat{q}_{\alpha_{lo}}(\mathbf{x}_{n+1}) - Q_{1-\alpha}(E_{low}, \mathcal{I}_2), \hat{q}_{\alpha_{hi}}(\mathbf{x}_{n+1}) + Q_{1-\alpha}(E_{high}, \mathcal{I}_2)], \quad (17)$$

where  $Q_{1-\alpha}(E, \mathcal{I}_2)$  is the  $(1 - \alpha)(1 + 1/|\mathcal{I}_2|)^{th}$  empirical quantile of the set of residuals  $E_i : i \in \mathcal{I}_2$  and  $\mathcal{I}_2$  is the set of indices corresponding to the residuals of the quantile regression estimator fitted on the calibration set. In the symmetric CQR method,  $E_{low}$  and  $E_{high}$  are equal.

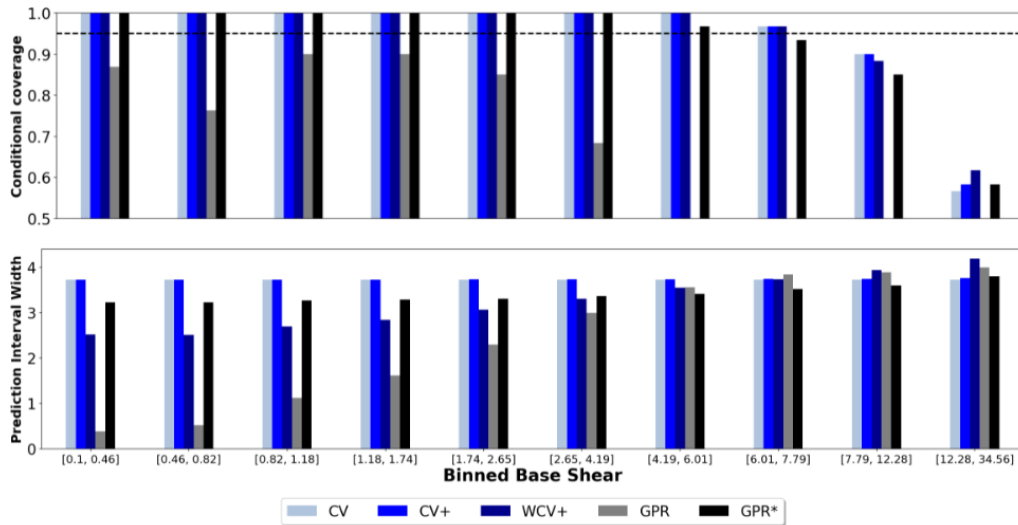
The idea behind CQR is to use the residuals of the quantile regression method as a proxy for the heteroscedasticity in

the data. By using these residuals to adjust the prediction interval, CQR can produce narrower and more accurate prediction intervals than traditional methods that do not take into account heteroscedasticity.

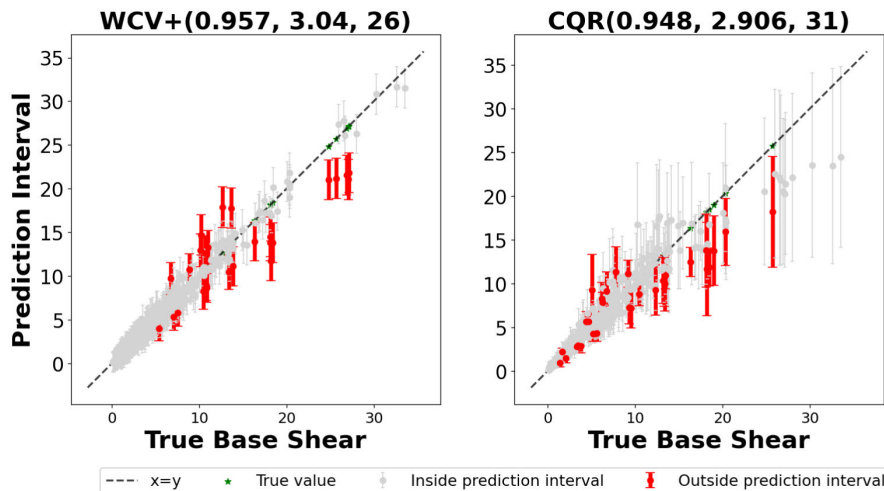
#### D. COMPARISON OF WCV+ WITH CQR USING LIGHTGBM

In this section, we compare the performance of two conformal prediction methods, WCV+ and CQR, using LightGBM as the base regressor. The MAPIE package is used to implement CQR, as detailed in their documentation available online at <https://mapie.readthedocs.io/>. While Ridge regression with





**FIGURE 11.** Conditional coverage and prediction interval width of six conformal prediction methods including CV, CV+, and WCV+ and two GPR methods at target coverage level of 0.95 on different bins of test points.



**FIGURE 12.** True values vs. prediction intervals on test points using WCV+ and CQR methods using LightGBM as the base regressor at the target coverage level of 0.95. Effective coverage, prediction interval width, and number of red missed test points lying outside of the prediction intervals of each method are shown in parentheses.

polynomial features was used as the base algorithm in section V-B and the previous parts of this paper, here we use LightGBM, a gradient-boosting ensemble method based on decision trees.

Fig. 12 displays the true values versus prediction intervals of test points using WCV+ and CQR methods. The values in parentheses represent the effective coverage, average prediction interval width, and the number of red missed points outside of prediction intervals for each method. Compared to Fig. 10, it is clear that the proposed methods WCV+ with base regressor LightGBM outperform WCV+ with Ridge regression with polynomial features. WCV+ with LightGBM achieves higher effective coverage while providing tighter prediction intervals and fewer missed test points.

However, as noted in [20], CQR can sometimes be overly conservative, leading to unnecessarily wide prediction

intervals. According to Fig. 12, while the average prediction interval width is 2.906 (slightly less than the 3.04 obtained with WCV+), CQR’s prediction interval widths can be too wide, especially in the tails of the data where the Base Shear is greater than 20. This can be attributed to the fact that CQR relies on quantile regression, which estimates the conditional quantiles of a response variable. While this technique is designed to provide a prediction interval that includes a certain proportion of the data points, typically 95%, it can be overly uncertain in areas where there is a lot of variation in the tails of the data. As a result, the prediction intervals can be unnecessarily wide, which can lead to lower effective coverage rates and more red intervals. In other words, while the prediction intervals provided by CQR may contain the true values of the test samples, they can be too wide to be informative.

## VI. CONCLUSION

In this paper, we proposed a modification to conformal prediction methods to account for violations of the exchangeability and homoscedasticity assumptions. We introduced data-dependent weights to the non-conformity scores, which allowed us to treat relevant training points differently and achieve improved accuracy in predicting seismic response. Our numerical experiments showed that our proposed method outperforms other commonly used uncertainty quantification methods in terms of both validity and efficiency. We evaluated the effective coverage, conditional coverage, and average prediction interval width on different intervals of response variables divided by quantiles. In particular, for the highly skewed and heavy-tailed seismic response data, our method achieved a 1% higher coverage level and a 15% decrease in average prediction interval width compared to other methods for all three target coverage levels (0.9, 0.95, 0.99). This demonstrates the practical value of our proposed method in scientific and engineering domains.

Furthermore, we compared our conformal prediction methods to GPR and found that our proposed method is superior in terms of reliability and ease of use. Our proposed method requires less prior knowledge and optimization than Gaussian process regression, making it more accessible to practitioners.

In summary, our proposed modification to conformal prediction methods, using data-dependent weights to adjust the non-conformity scores, provides a more accurate and reliable approach for quantifying uncertainty in machine learning models. Our numerical experiments demonstrate the effectiveness of our proposed method, including its comparison with CQR, and our comparison to GPR highlights its practical value. Future studies will explore the potential of our method to enhance trust in machine learning models, particularly for tasks with rare occurrences but significant consequences. Additionally, we plan to broaden our approach to encompass other well-regarded regression methods. Our contributions push the boundaries of existing knowledge in uncertainty quantification within machine learning, offering substantial benefits to both practitioners and scholars in the field.

## REFERENCES

- [1] M. A. Hariri-Ardebili and F. Pourkamali-Anaraki, "Matrix completion for cost reduction in finite element simulations under hybrid uncertainties," *Appl. Math. Model.*, vol. 69, pp. 164–180, May 2019.
- [2] B. Sudret, S. Marelli, and J. Wiart, "Surrogate models for uncertainty quantification: An overview," in *Proc. 11th Eur. Conf. Antennas Propag. (EUCAP)*, Mar. 2017, pp. 793–797.
- [3] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021.
- [4] R. Alizadehsani et al., "Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020)," *Ann. Oper. Res.*, pp. 1–42, 2021.
- [5] K. B. Hansen and C. Borch, "The absorption and multiplication of uncertainty in machine-learning-driven finance," *Brit. J. Sociol.*, vol. 72, no. 4, pp. 1015–1029, Sep. 2021.
- [6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. Cham, Switzerland: Springer, 2013.
- [7] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
- [8] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
- [9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
- [10] D. D. Cox, "An analysis of Bayesian inference for nonparametric regression," *Ann. Statist.*, vol. 21, no. 2, pp. 903–923, Jun. 1993.
- [11] M. J. Bayarri and J. O. Berger, "The interplay of Bayesian and frequentist analysis," *Stat. Sci.*, vol. 19, no. 1, pp. 58–80, Feb. 2004.
- [12] B. Szabó, A. W. van der Vaart, and J. H. van Zanten, "Frequentist coverage of adaptive nonparametric Bayesian credible sets," *Ann. Statist.*, vol. 43, no. 4, pp. 1391–1428, Aug. 2015.
- [13] J. Rousseau and B. Szabo, "Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors," 2016, *arXiv:1609.05067*.
- [14] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Cham, Switzerland: Springer, 2005.
- [15] V. Vovk, "Conditional validity of inductive conformal predictors," in *Proc. Asian Conf. Mach. Learn.*, 2012, pp. 475–490.
- [16] J. Lei and L. Wasserman, "Distribution-free prediction bands for non-parametric regression," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 76, no. 1, pp. 71–96, Jan. 2014.
- [17] H. Papadopoulos, "Inductive conformal prediction: Theory and application to neural networks," in *Tools in Artificial Intelligence*. Princeton, NJ, USA: Citeseer, 2008.
- [18] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1094–1111, Jul. 2018.
- [19] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Inf. Inference, J. IMA*, vol. 10, no. 1, pp. 455–482, Oct. 2020.
- [20] Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–8.
- [21] D. Kivaranovic, K. D. Johnson, and H. Leeb, "Adaptive, distribution-free prediction intervals for deep networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 4346–4356.
- [22] M. Sesia and E. J. Candès, "A comparison of some conformal quantile regression methods," *Stat.*, vol. 9, no. 1, p. e261, Jan. 2020.
- [23] N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 6417–6428.
- [24] A. Brando, B. S. Center, J. Rodriguez-Serrano, and J. Vitria, "Deep non-crossing quantiles through the partial derivative," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 7902–7914.
- [25] M. A. Hariri-Ardebili and B. Sudret, "Polynomial chaos expansion for uncertainty quantification of dam engineering problems," *Eng. Struct.*, vol. 203, Jan. 2020, Art. no. 109631.
- [26] M. A. Hariri-Ardebili and F. Salazar, "Engaging soft computing in material and modeling uncertainty quantification of dam engineering problems," *Soft Comput.*, vol. 24, no. 15, pp. 11583–11604, Aug. 2020.
- [27] M. A. Hariri-Ardebili and F. Pourkamali-Anaraki, "Structural uncertainty quantification with partial information," *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116736.
- [28] F. Pourkamali-Anaraki and M. A. Hariri-Ardebili, "Neural networks and imbalanced learning for data-driven scientific computing with uncertainties," *IEEE Access*, vol. 9, pp. 15334–15350, 2021.
- [29] Y. Xie, M. E. Sichani, J. E. Padgett, and R. DesRoches, "The promise of implementing machine learning in earthquake engineering: A state-of-the-art review," *Earthq. Spectra*, vol. 36, no. 4, pp. 1769–1801, Nov. 2020.
- [30] Y. Xie, M. E. Sichani, J. Padgett, and R. DesRoches, "Machine learning applications in earthquake engineering: Literature review and case studies," in *Proc. 17th World Conf. Earthq. Eng.*, 2020.
- [31] H. Salehi and R. Burgueño, "Emerging artificial intelligence methods in structural engineering," *Eng. Struct.*, vol. 171, pp. 170–189, Sep. 2018.
- [32] H.-T. Thai, "Machine learning for structural engineering: A state-of-the-art review," *Structures*, vol. 38, pp. 448–491, Apr. 2022.

- [33] W. Chen, K. Chun, and R. F. Barber, "Discretized conformal prediction for efficient distribution-free inference," *Stat.*, vol. 7, no. 1, Jan. 2018, Art. no. e173.
- [34] M. H. Quenouille, "Approximate tests of correlation in time-series 3," *Math. Proc. Cambridge Phil. Soc.*, vol. 45, no. 3, pp. 483–484, Jul. 1949.
- [35] J. Tukey, "Bias and confidence in not quite large samples," *Ann. Math. Statist.*, vol. 29, p. 614, 1958.
- [36] V. Vovk, "Cross-conformal predictors," *Ann. Math. Artif. Intell.*, vol. 74, nos. 1–2, pp. 9–28, Jun. 2015.
- [37] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife," *Ann. Statist.*, vol. 49, no. 1, pp. 486–507, Feb. 2021.
- [38] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan, "Distribution-free, risk-controlling prediction sets," *J. ACM*, vol. 68, no. 6, pp. 1–34, Dec. 2021.
- [39] F. Svensson, N. Aniceto, U. Norinder, I. Cortes-Ciriano, O. Spjuth, L. Carlsson, and A. Bender, "Conformal regression for quantitative structure–activity relationship modeling-quantifying prediction uncertainty," *J. Chem. Inf. Model.*, vol. 58, no. 5, pp. 1132–1140, 2018.
- [40] C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. (1996). *The Delve Manual*. [Online]. Available: <http://www.cs.toronto.edu/~delve>
- [41] A. Frank and A. Asuncion. (2010). *UCI Machine Learning Repository*. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [42] J. Derrac, S. Garcia, L. Sanchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Valued Log. Soft Comput.*, vol. 17, pp. 255–287, Jun. 2015.
- [43] M. A. Hariri-Ardebili, H. Rahmani-Samani, and M. Mirtaheri, "Seismic stability assessment of a high-rise concrete tower utilizing endurance time analysis," *Int. J. Struct. Stability Dyn.*, vol. 14, no. 6, Aug. 2014, Art. no. 1450016.
- [44] P. Hajibabae, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "An empirical evaluation of the t-SNE algorithm for data visualization in structural engineering," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 1674–1680.
- [45] P. Hajibabae, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "Dimensionality reduction techniques in structural and earthquake engineering," *Eng. Struct.*, vol. 278, Mar. 2022, Art. no. 115485.
- [46] P. Hajibabae, F. Pourkamali-Anaraki, and M. A. Hariri-Ardebili, "Kernel matrix approximation on class-imbalanced data with an application to scientific simulation," *IEEE Access*, vol. 9, pp. 83579–83591, 2021.
- [47] M. A. Hariri-Ardebili and S. Barak, "A series of forecasting models for seismic evaluation of dams based on ground motion meta-features," *Eng. Struct.*, vol. 203, Jan. 2020, Art. no. 109657.
- [48] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, vol. 32. Cambridge, MA, USA: MIT Press, 2006, p. 68.
- [49] D. J. MacKay, "Introduction to Gaussian processes," *NATO ASI Ser. F Comput. Syst. Sci.*, vol. 168, pp. 133–166, Dec. 1998.
- [50] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Cham, Switzerland: Springer, 2003, pp. 63–71.



**PARISA HAJIBABAE** received the master's degree in industrial engineering from West Virginia University and the Ph.D. degree in computer science from the University of Massachusetts Lowell. She is currently an Assistant Professor with the Department of Data Science and Business Analytics, Florida Polytechnic University. Her main research interests include the intersection of statistics, optimization, uncertainty quantification, and applied machine learning.



**FARHAD POURKAMALI-ANARAKI** received the Ph.D. degree in electrical engineering from CU Boulder, in 2017. He is currently an Assistant Professor with the Department of Mathematical and Statistical Sciences, University of Colorado Denver. Previously, he was an Assistant Professor of computer science with the University of Massachusetts Lowell, from 2018 to 2022. His main research interests include transitioning machine learning models from controlled laboratory environments to real-world settings involving unpredictable and changing conditions, such as accelerating the design and discovery of new materials using cost-effective and uncertainty-aware machine learning models.



**MOHAMMAD AMIN HARIRI-ARDEBILI** is currently a Researcher with the National Institute of Standards and Technology (NIST). He is a Faculty Member with the University of Maryland, College Park, MD, USA, and the University of Colorado Boulder. His main research interests include performance-based earthquake assessment of infrastructures, coupled systems mechanics, risk and resilience, climate change and aging, uncertainty quantification, optimization, and scientific machine learning.

• • •