Hindawi
Structural Control and Health Monitoring
Volume 2023, Article ID 6929861, 14 pages
https://doi.org/10.1155/2023/6929861

Research Article

Interval Prediction Model of Deformation Behavior for Dam Safety during Long-Term Operation Using Bootstrap-GBDT

Erfeng Zhao[1,2] **Yi Li**[1,2] **Jingmei Zhang**[2,3] **and Zhangyin Li**[1,2]

[1]*State Key Laboratory of Hydrology Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China*
[2]*National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing 210098, China*
[3]*College of Water Conservancy and Environmental Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China*

Correspondence should be addressed to Yi Li; 211302020013@hhu.edu.cn

Received 10 November 2022; Revised 24 March 2023; Accepted 28 April 2023; Published 12 May 2023

Academic Editor: Andrea Del Grosso

Accurate and reliable prediction on structural deformation is a powerful means to evaluate the safety state of dams during long-term operation. This work is a contribution to the quantification of the cognitive and stochastic uncertainties in prediction models for the dam performance using measured time series, through constructing the prediction interval (PI) by an innovative model combined with the gradient boosting decision tree (GBDT) and the bootstrap method. The constructed PI, improved by the kernel density estimation (KDA), consists of an upper bound and a lower bound of the interval to provide a confidence level for dam deformation prediction. The bootstrap method combined with multiple GBDT is utilized to estimate the variance of the model bias, mainly derived from the cognitive uncertainties. The variance of random noise can be furtherly estimated through training the combined model to fit the residuals so as to indicate the stochastic uncertainties. The effectiveness of the newly proposed model is validated employing measured data of the Jinping I arch Dam. The results indicate that the methodology can obtain high-quality PIs and accurate prediction values and thus can provide strong support for better appraising dam performance during long-term operation.

1. Introduction

Safety monitoring of dams is essential for identifying their performance and managing risks during long-term operation. Possible anomalies or potential hidden dangers can be diagnosed in time through appraising the practical status of dams [1, 2]. They are vital to eliminate abnormal behavior and prevent catastrophic accidents. Hence, according to the causal mechanism between measured effects and environmental quantities of dams, it has been a hot research topic to establish accurate prediction models with historical measured time series through multivariate statistical analysis, machine-learning methods, etc., to judge whether their performance normal or not. Through constantly integrating with new technologies in the field of big data mining and artificial intelligence, various prediction methods have emerged in recent years [3].

Among all the measured effects, dam deformation can indicate its operating performance effectively, so it is usually regarded as the most reliable indicator to predict the dam behavior [4, 5]. Traditional modeling methods for dam deformation include the hydrostatic-seasonal-time (HST) model and the hydrostatic-temperature-time (HTT) model [6–9]. Among machine-learning methods, BPNN [10], long- and short-term memory neural networks [11], extreme learning machine (ELM) [12], and radial basis function neural networks [13] are all used in dam safety modeling to indicate the nonlinear causal mechanism. However, these machine-learning models are a "black box," since the physical relationships between the measured effects and environmental variables cannot be expressed explicitly by them [14]. The gradient boosting decision tree (GBDT) is an iterative decision tree algorithm [15], which consists of

multiple decision trees. The conclusions of all trees are accumulated to make the appraisement results. Compared with other machine-learning methods, the GBDT can indicate the importance of each influence factor when modeling. Furthermore, because dam monitoring systems are often influenced by multiple factors, the multicollinearity problem [16, 17] referring to two or more highly correlated factors probably happens in the traditional HST and HTT models [18]. Dimensionality reduction can usually alleviate covariance through principal component analysis (PCA) [19]. In addition, the stochastic and uncertain response of dam structures under various factors can be indicated by interval prediction, such as the Bayesian method, mean squared deviation estimation, upper and lower bound estimation, and bootstrap method [20]. Chen et al. [13] combined a correlation vector machine for probability prediction, providing the confidence interval to quantify the uncertainty of the dam deformation behavior. Ren et al. [21] built an interval prediction model through integrated learning to indicate the uncertainty and the variability of the predicted deformation. Ren et al. [22] integrated nonparametric bootstrap, least squares support vector machine and artificial neural network algorithms to generate high-quality prediction intervals and identified data noises and model noises. However, in the Bayesian and mean square estimation methods, the sample distribution needs to be artificially assumed in advance [20]. By contrast, the upper and lower bound estimation can construct the intervals directly, though the reliability of its prediction results is insufficient [20]. The deficiency can also be improved by the bootstrap method, through constructing prediction intervals with high coverage and low width to obtain accurate prediction results [23].

Dam deformation has various uncertainties, divided into cognitive uncertainties and stochastic uncertainties [22, 24], which may weaken modeling accuracy. The uncertainties caused by the environmental factors and subjective assumptions when establishing prediction models can be classified as cognitive as the former. The latter is usually caused by random noises in the measured time series [22, 25]. Hence, the present work is a contribution to the quantification of these uncertainties for dams during long-term operation. First, the prediction interval (PI) is constructed and improved by the kernel density estimation (KDA) [26] to provide a confidence level for dam deformation prediction. Second, an innovative interval prediction model for dam deformation is established by integrating the bootstrap and GBDT methods with the improved PI to quantify the two uncertainties mentioned above. Finally, model validation is implemented on an arch dam to evaluate its development trend during long-term operation.

2. Improved PI for Dam Deformation

The PIs are effective for quantifying the uncertainties in prediction models for dam performance [22]. For a set of the measured time series, $\mathbf{D} = \{\mathbf{X}_i, y_i\}_{i=1}^{n}$, where $\mathbf{X}_i$ is a set of the environmental factors, y_i represents the measured deformation at the i^{th} moment, and n is the total monitoring number. The conventional PI consists of an upper bound and a lower bound to provide a confidence level for the prediction target:

$$I^{(\alpha)}(\mathbf{X}_i) = \left[L^{(\alpha)}[\mathbf{X}_i], U^{(\alpha)}[\mathbf{X}_i]\right], \tag{1}$$

where $I^{(\alpha)}(\mathbf{X}_i)$ is the PI generated by the interval prediction model, α is the significance level, and $U^{(\alpha)}[\mathbf{X}_i]$ and $L^{(\alpha)}[\mathbf{X}_i]$ are the upper and lower bounds of the i^{th} PI, respectively.

The upper and lower limits of the interval can be obtained as follows:

$$\begin{cases} L^{(\alpha)}[\mathbf{X}_i] = \widehat{y}_i - z_{1-\alpha/2}\sqrt{\sigma^2(\mathbf{X}_i)}, \\ U^{(\alpha)}[\mathbf{X}_i] = \widehat{y}_i + z_{1-\alpha/2}\sqrt{\sigma^2(\mathbf{X}_i)}, \end{cases} \tag{2}$$

where $\widehat{y}_i$ is the point prediction result, $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution, and $\sigma^2(\mathbf{X}_i)$ is the variance of the whole errors.

The whole errors $\sigma^2(\mathbf{X}_i)$ include the model bias and random noises. Assuming that they are independent of one another, the corresponding linear combination of variances between the errors can be constructed:

$$\sigma^2(\mathbf{X}_i) = \sigma_{\Delta}^2(\mathbf{X}_i) + \sigma_{\varepsilon}^2(\mathbf{X}_i), \tag{3}$$

where $\sigma_{\Delta}^2(\mathbf{X}_i)$ is the variance of the model bias, indicating the cognitive uncertainties, and $\sigma_{\varepsilon}^2(\mathbf{X}_i)$ is the variance of the random noises, indicating the stochastic uncertainties.

The residuals of the interval prediction model are usually assumed to obey the standard normal distribution, easily leading to some deviation from the actual distribution and probably reducing the validity of the constructed PI. Hence, KDA is taken to fit the residual distribution to accurately estimate the actual probability density function (PDF):

$$f(\varepsilon) = \frac{1}{th}\sum_{i=1}^{t} K\left(\frac{\varepsilon - \varepsilon_i}{h}\right), \tag{4}$$

where $f(\varepsilon)$ is the PDF of the actual residual distribution, t is the size of the residual, h is the window width, ε_i is the sample point, and $K(\cdot)$ is the kernel function.

The accuracy of KDA is dependent on $K(\cdot)$ and h. For the long-measured time series, the Gaussian kernel function is usually taken to estimate the kernel density:

$$K\left(\frac{\varepsilon - \varepsilon_i}{h}\right) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(\varepsilon - \varepsilon_i)^2}{2h^2}\right]. \tag{5}$$

According to equation (4), $z_{1-\alpha/2}$ can be replaced by the inverse of the PDF of the actual residual distribution. In this case, the upper and lower limits of the interval in equation (2) can be improved as follows:

$$\begin{cases} L^{(\alpha)}[\mathbf{X}_i] = \widehat{y}_i + q\sqrt{\sigma^2(\mathbf{X}_i)}, \\ U^{(\alpha)}[\mathbf{X}_i] = \widehat{y}_i + p\sqrt{\sigma^2(\mathbf{X}_i)}, \end{cases} \tag{6}$$

where p and q are the $1-\alpha/2$ upper percentile and the $1-\alpha/2$ lower percentile of the PDF, respectively.

In equation (6), the key to constructing the PI is the calculation of the variance of the whole errors $\sigma^2(\mathbf{X}_i)$ and the percentiles p and q. They can be obtained by the bootstrap-GBDT model established in Section 3.3.

3. Interval Prediction Model Establishment for Dam Deformation

Dam performance is often affected by various internal and external factors during long-term operation, presenting complex uncertainties. An innovative interval prediction model will be constructed by integrating the bootstrap method and the GBDT, where the high-quality PI can be obtained with guaranteed prediction accuracy.

3.1. Estimation of the Overall Distribution by the Bootstrap Method. The bootstrap method can obtain the distribution characteristics of the samples only by repeatedly resampling the original data, without artificial assumptions. It has the advantages of high robustness and accuracy [20]. The nonparametric bootstrap method can be well for resampling the influence variables and effects from the original dataset [27].

For the original data $\mathbf{D} = \{\mathbf{X}_i, y_i\}_{i=1}^n$, the statistical estimate of overall Θ can be calculated by using the bootstrap method as follows.

Step 1. The original data are resampled with B times to obtain the B sets of the pseudodata set $\mathbf{D}^*$.

Step 2. The unknown parameters (including the mean and the variance) of these B sets of the pseudodata are calculated based on statistical theory.

Step 3. The calculated mean and variance are both taken to estimate the parameters of the overall distribution Θ.

3.2. Procedure of the GBDT for Prediction of Dam Deformation. The GBDT is an integrated learning method, whose results generated by all base learners can be integrated through certain processing methods [28]. The base learner can adopt the classification and regression tree (CRT) [29]. For the regression problem, the CRT model divides the input space into two subregions and determines the output value of each subregion and then constructs a binary decision tree. The input variables are divided, and the optimal shared variable s and the shared point j are selected to make the difference between the measured value and the output mean value minimized as follows:

$$\min_{s,j}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i - c_1)^2\right]. \tag{7}$$

We traverse the variable s and the point j to minimize equation (7). Selected (s, j) is used to determine the output value of the corresponding region:

$$\begin{aligned} R_1(s,j) &= \{x|x^s \le j\},\\ R_2(s,j) &= \{x|x^s > j\},\\ c_m &= \frac{1}{N_m}\sum_{x_i\in R_m(j,s)} y_i, \quad m = 1, 2, \end{aligned} \tag{8}$$

where R_m is a region generated by the (s, j), N_m is the total number of the samples in the m^{th} region, and c_m is the output value of the m^{th} region.

The specified conditions can be met through repeating the above operations for the subregion. The input space is assumed to be divided into K regions, denoted as $R_1, R_2 \ldots R_K$. Each region corresponds to the output variable c_k. Then, the regression tree is represented as follows:

$$T(x,\Theta) = \sum_{k=1}^{K} c_k I(x \in R_k), \tag{9}$$

where $T(x, \Theta)$ represents the decision tree, Θ is the decision tree's parameter, and $I(x \in R_k)$ is an indicator function:

$$I(x \in R_k) = \begin{cases} 1, & x \in R_k,\\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The GBDT model can be expressed as an additive model of the CRT:

$$f_m(x) = \sum_{m=1}^{M} T(x,\Theta_m), \tag{11}$$

where M represents the number of the decision trees.

The GBDT model adopts the forward distribution algorithm. The initial decision tree $f_0(x) = 0$ is determined, and the m^{th} decision tree model is

$$f_m(x) = f_{m-1}(x) + T(x,\Theta_m), \tag{12}$$

where $f_{m-1}(x)$ is a previous tree model.

The parameter Θ_m of the m^{th} tree model can be determined through prediction accuracy minimization $\arg\min_{\Theta_m}\sum_{i=1}^{N} L[y_i, f_{m-1}(x_i) + T(x_i,\Theta_m)]$. The squared error loss function S is generally adopted to indicate the prediction accuracy:

$$S = \sum_{i=1}^{N} L[y_i, f_{m-1}(x_i) + T(x_i,\Theta_m)] = \sum_{i=1}^{N}[y_i - f_{m-1}(x_i) - T(x_i,\Theta_m)]^2. \tag{13}$$

To avoid overfitting, the regularization parameter, i.e., the learning rate ν, is adopted:

$$f_m(x) = f_{m-1}(x) + \nu T(x, \Theta_m), \quad \nu \in [0, 1]. \tag{14}$$

The GBDT can appraise the importance of each influence factor for dam deformation and regard each factor as a cut-off variable to indicate its importance:

$$J_i = \frac{1}{N}\sum_{n=1}^{N} I_i(T_n) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{j-1} I_j(x_i), \tag{15}$$

where J_i is the overall degree of the importance of x_i, N is the number of the tree models, T_n is the n^{th} tree model, j is the number of internal nodes containing $j-1$ nonterminal nodes in the T_n tree model, $I_i(T_n)$ is the importance of x_i in the T_n tree model, and $I_j(x_i)$ denotes that if x_i is the cut variable of the j^{th} node, then it equals to 1; otherwise, it equals to 0.

3.3. The Establishment of the Combined Model. During long-term operation, the main influence factors of dam deformation are hydrostatic pressure, temperature change, aging, and other unknown factors. It is often divided into

$$\delta = \delta_H + \delta_T + \delta_\theta + \varepsilon, \tag{16}$$

where δ is the dam deformation, δ_H is the hydraulic component, δ_T is the temperature component, δ_θ is the aging component, and ε is the random residual.

According to traditional regression models for dam safety monitoring [30, 31], the environmental variables, regarded as the model input factors for an arch dam, are listed as follows:

$$\mathbf{X} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}] = \left[H, H^2, H^3, H^4, \sin\frac{2\pi t}{365}, \cos\frac{2\pi t}{365}, \sin\frac{4\pi t}{365}, \cos\frac{4\pi t}{365}, \theta, \ln\theta\right], \tag{17}$$

where H is the upstream water depth, t is the number of monitoring days, and $\theta = t/100$.

Since PCA can convert multiple correlated components into several uncorrelated PCs by the orthogonal exchange, it can ensure higher accuracy and be less likely to be pathological of the proposed model. Thus, the PCA method is first used to alleviate the multicollinearity issue of the environment variables. For $\mathbf{X}$ in equation (17), it is normalized to obtain $\mathbf{X}'$. The correlation matrix $\mathbf{R}$ of $\mathbf{X}'$ is

$$\mathbf{R} = \frac{1}{n+1}\left(\mathbf{X}'^{\mathrm{T}} \cdot \mathbf{X}'\right). \tag{18}$$

According to equation (18), the eigenvalues, eigenvectors, corresponding contribution rates, and cumulative contribution rates of $\mathbf{R}$ can all be calculated. Eigenvalues are usually sorted in descending order. Once the cumulative contribution rate of new components exceeds the threshold, they will be selected to form new principal elements. Their corresponding eigenvectors will form the principal component matrix $\mathbf{T}$. Then, the PCs $\mathbf{L}$ can be obtained as follows:

$$\mathbf{L} = \mathbf{X}' \cdot \mathbf{T}. \tag{19}$$

Principal components (PCs) can reduce the dimensionality for the dataset of the environmental factors. Then, the B pseudosamples $\mathbf{D}^*$ are constructed by the bootstrap method, and a total of B GBDT models are trained, respectively. The PDF of the residual distribution is estimated by the KDA to calculate p and q:

$$p = \frac{1}{B}\sum_{i=1}^{B} p_i, \tag{20}$$

$$q = \frac{1}{B}\sum_{i=1}^{B} q_i, \tag{21}$$

where p_i and q_i are the $1-\alpha/2$ upper percentile and the $1-\alpha/2$ lower percentile obtained in the i^{th} training, respectively.

The seagull optimization algorithm (SOA) [32] is taken to obtain optimal parameters of the GBDT, where the number of trees, the learning rate, and the tree depth are all optimization parameters. The predicted MSE is the fitness function:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{22}$$

where y_i is the measured value and $\hat{y}_i$ is the predicted value.

The expectation of the B pseudosample prediction result is regarded as the point prediction result in equation (23), and the output variance of the B GBDT model is taken to estimate the variance of the model errors in equation (24):

$$\overline{y}(\mathbf{X}_i) = \frac{1}{B}\sum_{l=1}^{B} y_l(\mathbf{X}_i), \tag{23}$$

$$\sigma_\Delta^2(\mathbf{X}_i) = \frac{1}{B-1}\sum_{l=1}^{B}\left[y_l(\mathbf{X}_i) - \overline{y}(\mathbf{X}_i)\right]^2, \tag{24}$$

where $y_l(\mathbf{X}_i)$ is the prediction result of the l^{th} model and $\overline{y}(\mathbf{X}_i)$ is the predicted result of the B models, i.e., the point prediction result.

The noise variance can be estimated by equation (25). In the residual dataset $\mathbf{D}_{r^2} = \{(\mathbf{X}_i, r^2(\mathbf{X}_i)), i = 1, 2, \ldots, n\}$, the residual $r^2(\mathbf{X}_i)$ can be determined by equation (26):

$$\sigma_\varepsilon^2(\mathbf{X}_i) \simeq E\left\{\left[y - \overline{y}(\mathbf{X}_i)\right]^2\right\} - \sigma_\Delta^2(\mathbf{X}_i), \tag{25}$$

$$r^2(\mathbf{X}_i) = \max\left(\left[y - \overline{y}(\mathbf{X}_i)\right]^2 - \sigma_\Delta^2(\mathbf{X}_i), 0\right). \tag{26}$$

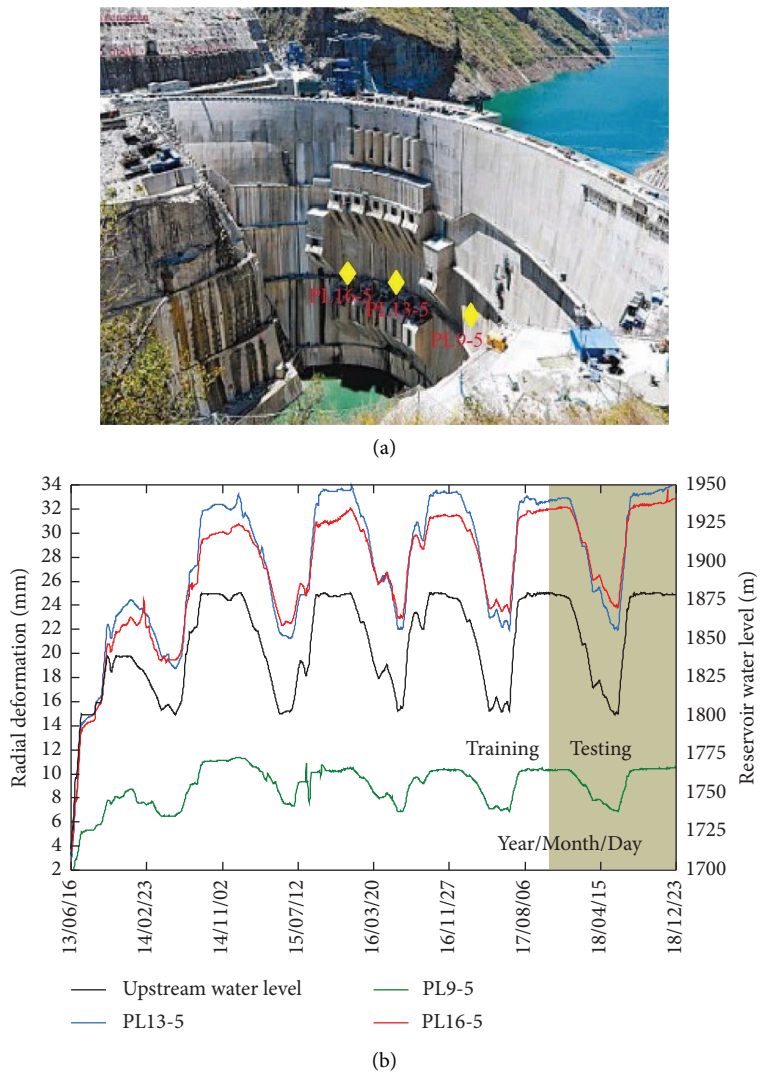

FIGURE 1: The measured time series of dam radial deformation and the reservoir water level: (a) location of the three monitoring points; (b) measured time series.

The $B+1^{st}$ GBDT model is trained to fit the residuals. C_{B+1} (omitting the constant term) is regarded as the loss function to train the $B+1^{st}$ GBDT model to estimate the noise variance:

$$P\left(r^2(\mathbf{X}_i), \sigma_\varepsilon^2(\mathbf{X}_i)\right) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2(\mathbf{X}_i)}} \exp\left(-\frac{r^2(\mathbf{X}_i)}{2\sigma_\varepsilon^2(\mathbf{X}_i)}\right)$$
$$C_{B+1} = \frac{1}{2}\sum_{i=1}^{N}\left[\frac{r^2(\mathbf{X}_i)}{\sigma_\varepsilon^2(\mathbf{X}_i)} + \ln\left(\sigma_\varepsilon^2(\mathbf{X}_i)\right)\right]. \quad (27)$$

Finally, the PI for dam deformation can be constructed using the combined model. The variance of the model errors $\sigma_\Delta^2(\mathbf{X}_i)$ can be obtained by the B GBDT model in equation (24). The variance of the random noise $\sigma_\varepsilon^2(\mathbf{X}_i)$ can be estimated by training the $B+1^{st}$ GBDT model in equation (25). The $1-\alpha/2$ percentile p and q can be calculated by the KDA in equations (20) and (21). On this basis, the improved PI can be obtained using equation (6).

4. Validation and Application

The proposed model is validated and applied into the Jinping I Arch Dam, located on the Yalong River in southwest China. Its dam foundation elevation is 1580.0 m, and the maximum dam height is 305.0 m. The normal water level of the reservoir is 1880.0 m. The dam consists of 26 dam sections. As shown in Figure 1(a), three monitoring points PL9-5, PL13-5, and PL16-5 in the perpendicular line monitoring system are taken for illustration. The measured time series of the three points and the reservoir water level are shown in Figure 1(b). The measured values of the three points increase and decrease accordingly when the reservoir water level rises and falls, and the effect of the temperature change is also included, indicating significant correlations with the environmental factors. The measured time series from June 16, 2013, to December 23, 2018, with data length 2017, are selected for modeling, setting June 16, 2013, to November 1, 2017, as

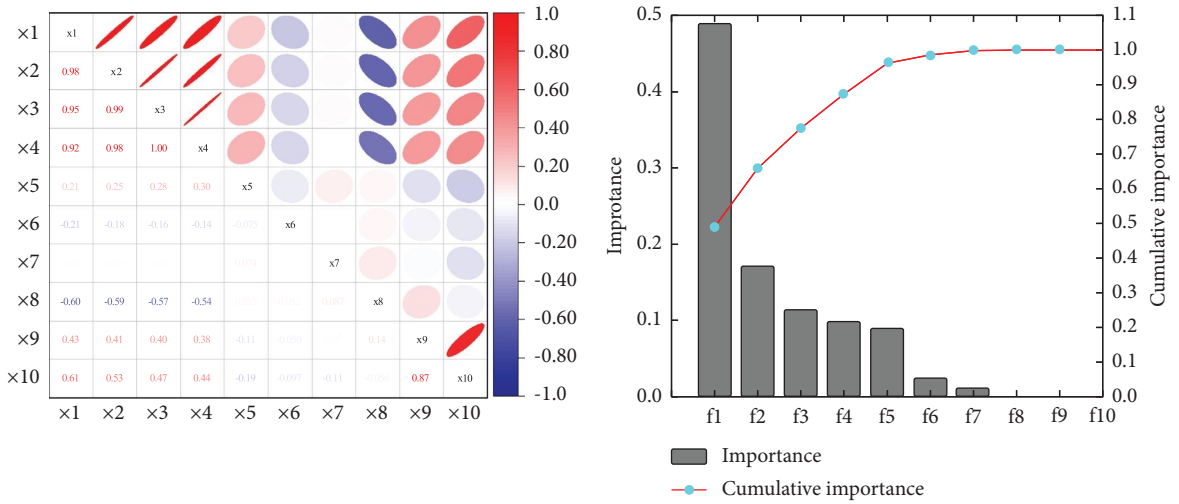

Figure 2: Principal component analysis results: (a) thermodynamic diagram of the features; (b) feature importance ranking.

the training set and November 2, 2017, to December 23, 2018, as the testing set, respectively.

4.1. Calculating the Input Vector of the Combined Model. Taking the monitoring point PL13-5 for illustration, in order to indicate the multicollinearity among the influence factors in equation (17), the correlation coefficients among them are calculated using the Pearson correlation coefficients, and the thermodynamic diagram is shown in Figure 2(a). The correlation coefficients among these factors are large, indicating significant multicollinearity, which probably lead to poor accuracy of the established model and weakness of modelling effectiveness. According to equation (18), the contribution rate and the cumulative contribution rate of each PC are calculated and shown in Figure 2(b). The contribution rate of the first 5 PCs ($f1$, $f2$, $f3$, $f4$, and $f5$) calculated by equation (19) reaches 96.2%, which has exceeded a threshold value of 95%. Hence, these PCs are selected to replace the original factors as the input vector to reduce data dimensionality. According to Figure 1(b), PCs are divided into the training set and test set, as shown in Figure 3(a).

In addition, as shown in Figure 3(b), PCA-MLR and the MLR are both built to further indicate the effect of the adopted PCA on model performance. The MSE and the calculation time of the PCA-MLR model are 2.92 and 2.31 s, while those of the MLR model are 3.01 and 4.53 s. The results indicate that PCA can not only improve prediction accuracy but also enhance computational efficiency.

4.2. Comparison of Different Point Prediction Methods. To compare with the point prediction results of the bootstrap-GBDT method, other 4 prediction models are also constructed by using multiple linear regression (MLR) [33], BPNN, ELM, and SOA-GBDT, respectively. The input factors in each model are the 5 PCs mentioned above. The following evaluation indexes are taken to compare the prediction accuracy of these established models, including MAE, MAPE, and RMSE:

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \\ \mathrm{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right|}, \\ \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}. \end{aligned} \tag{28}$$

The measured radial deformation of PL13-5 is still taken for illustration. SOA, the grey wolf optimizer (GWO) [34], and particle swarm optimization (PSO) [35] are all utilized to search the optimal parameters of the GBDT, respectively, including the number m of trees, the tree depth h, and the learning rate v. The search space of these parameters is [50, 400], [1, 4], and [0, 1], respectively. The maximum iteration number and the population size of the three algorithms are set to 400 and 30. The convergence curves of the GBDT optimized by these algorithms are shown in Figure 4, where the fitness value quickly reaches the minimum value of 0.201 of SOA, which is better than that of the other two algorithms. The optimal parameter values are finally obtained by SOA, $m = 189$, $h = 2$, and $v = 0.157$. In the BPNN model, the number of hidden layers, the learning rate, the amount of training, and the training objective for the MSE are set as 10, 0.01, 1000, and 10^{-3}, respectively. For the ELM model, the number of the hidden layers is taken as 30. The sigmoid function is regarded as an activation function for these models. For these neural network models, the number of nodes in the input layer and the number of nodes in the output layer are set as 5 and 1, respectively. Since the weights are randomly initialized, the averages of these two models trained five times are taken as their final results.

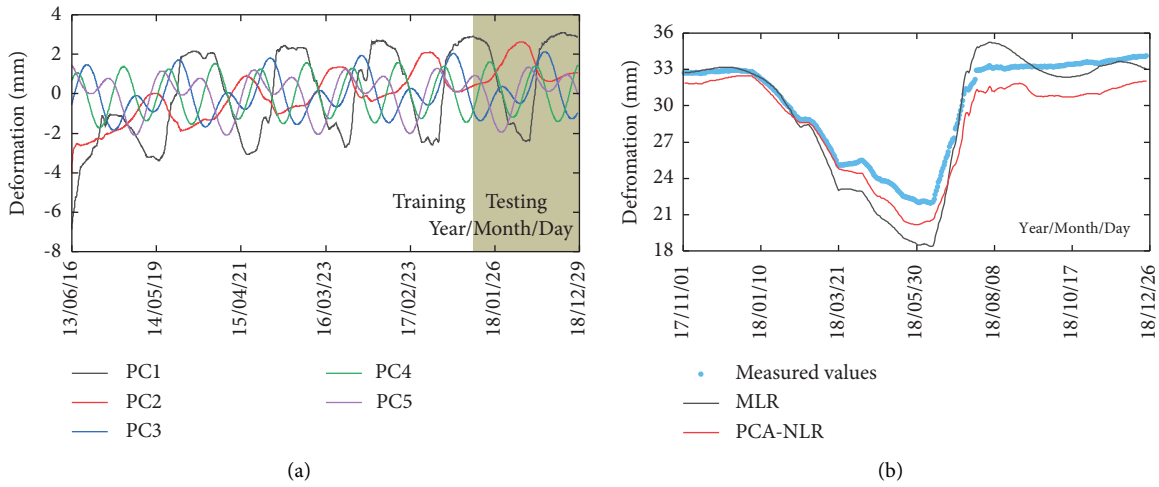

Figure 3: The comparison results whether adopting the PCA or not: (a) time series of the PCs; (b) comparison of PCA-MLR and MLR.

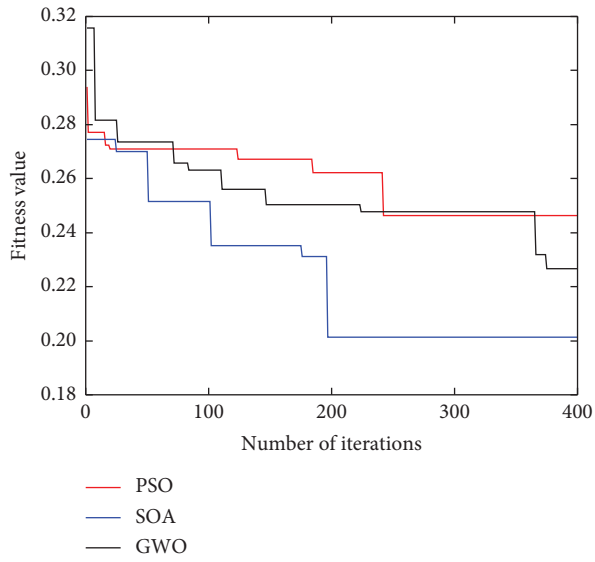

Figure 4: The convergence curves of different optimization algorithms.

The point prediction results of each model are shown in Figure 5, in which the bootstrap-GBDT model is shown in Figure 5(a), the SOA-GBDT model is shown in Figure 5(b), the GBDT model is shown in Figure 5(c), the ELM model is shown in Figure 5(d), the BPNN model is shown in Fig. 5(e), and the MLR model is shown in Figure 5(f). Table 1 lists the quantitative evaluation results. Compared with MLR, the corresponding MAE, MAPE, and RMSE of the GBDT model are reduced by 54.4%, 59.4%, and 60.6%, respectively, indicating the effectiveness of the GBDT model. The predicted MAE, MAPE, and RMSE indexes of the SOA-GBDT model can provide reasonably better results, indicating that the accuracy of the optimized GBDT model by SOA has been improved. The prediction accuracy of the bootstrap-GBDT model is highest in Table 1, and the values of MAE, MAPE, and RMSE are 0.2975, 1.06%, and 0.4170, respectively. They all indicate the superiority of the proposed model.

Moreover, the GBDT model has not only high prediction accuracy but also high explanatory ability. As shown in Figure 6, the importance degree of the input 5 PCs is 53.6%, 20%, 12.9%, 9.2%, and 4.3%, respectively. The results indicate that the physical relationships between the measured effects and the environmental factors can be expressed explicitly through the GBDT model.

4.3. Comparison of Different Interval Prediction Methods

4.3.1. Comparative Analysis of Different Methods. According to Section 3.3, the bootstrap-GBDT interval prediction model is constructed. The 10 pseudosamples are obtained by the bootstrap method to train the GBDT models. To calculate the $1-\alpha/2$ percentile of the PDF of training residual distribution p and q, its distribution is first tested whether obeying the normal distribution or not. The first training is taken as an example. The distributions of the training residuals are shown in Figure 7(a). The quantile-quantile (Q-Q) plot technique is applied to measure the normality of distributions, as shown in Figure 7(b). In Figure 7(b), the residual distribution deviates from the normal distribution at the tail. The quality of the PI is extremely sensitive to the tail deviation of the PDF. For the bandwidth h, the lager h is chosen, the smoother the fitted PDF is, but the lower the accuracy is. Conversely, the smaller h is chosen, the higher the fitting accuracy is, but the lower the smoothness of the PDF is. Hence, the PDF of the residual distribution is fitted by the KDA with different bandwidths (including $h = 0.1$, 0.5, and 1). In Figure 7(a), the bandwidth setting of 1 is more appropriate. The B pseudosamples are trained in turn to obtain their respective residuals. The results indicate that the residuals do not obey the normal distribution with a zero mean, so the KDA is utilized to fit their actual distribution, where h is set to 1. The upper and lower percentiles are listed in Table 2. p and q are calculated as 0.94 and 0.82, respectively.

The improved PI is constructed based on the bootstrap-GBDT model to quantify the uncertainties for the arch dam

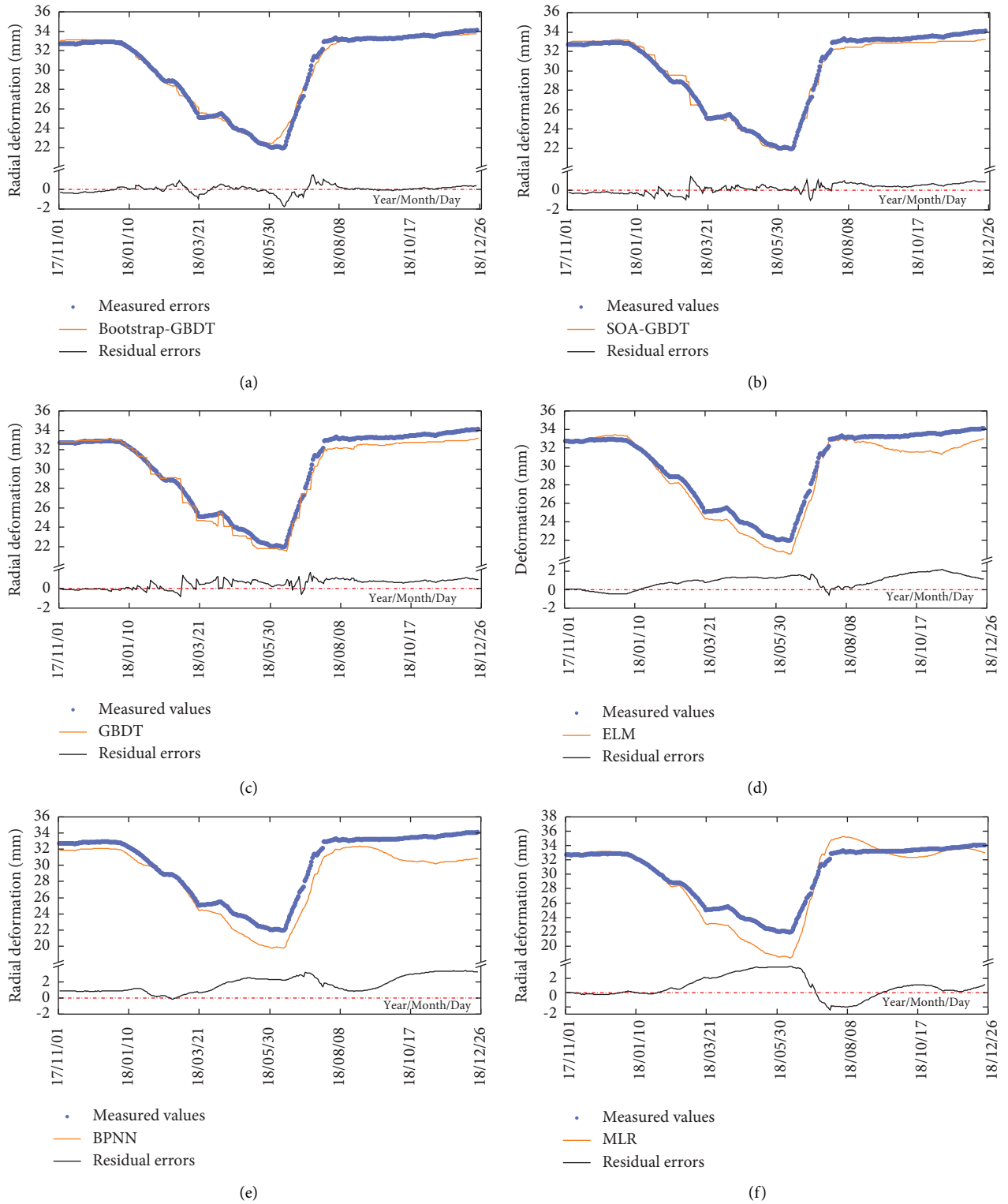

Figure 5: Point prediction results of different models: (a) bootstrap-GBDT model; (b) SOA-GBDT model; (c) GBDT model; (d) ELM model; (e) BPNN model; (f) MLR.

during long-term operation. PICP, MPIW, and CWC are all taken to evaluate the effectiveness of interval prediction [21]. The confidence level α is taken as 5%, which means that the measured values in Figure 1(b) are considered to have 95% probability to fall within the prediction interval. The penalty parameter η is set as 10 in this work. The PI constructed by different methods is shown in Figure 8, including the LUBE-BPNN model [36], the LUBE-GBDT model using the

Table 1: Evaluation results of the six models.

Evaluation indexes	MLR	BPNN	ELM	GBDT	SOA-GBDT	Bootstrap-GBDT
MAE	1.2716	1.0661	0.8720	0.5797	0.4013	0.2975
MAPE (%)	4.83	3.64	2.91	1.96	1.38	1.06
RMSE	1.7355	1.4652	1.0889	0.6844	0.5021	0.4170

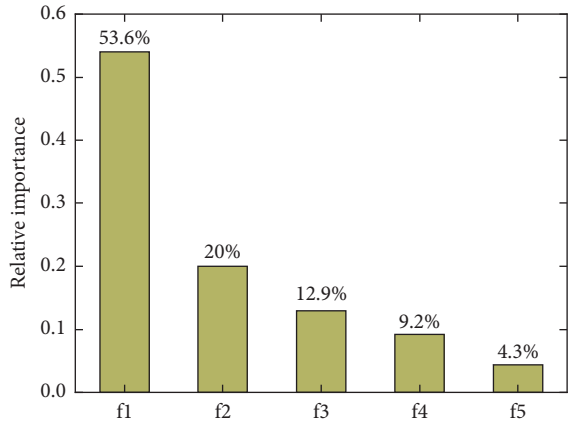

Figure 6: Feature importance evaluation results.

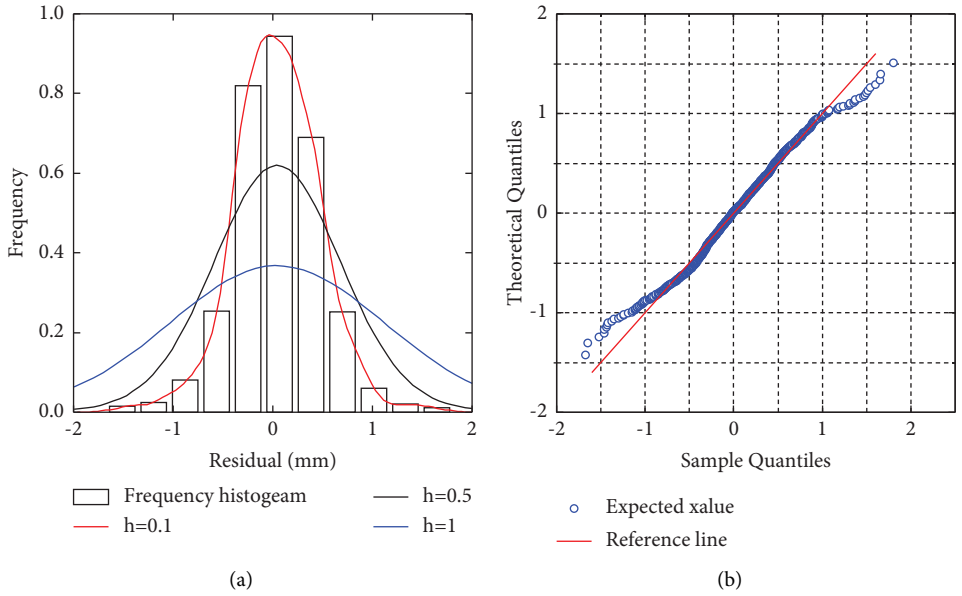

Figure 7: The normal distribution test of the residuals: (a) histogram and different bandwidths; (b) Q-Q plot.

Table 2: The upper and lower percentile.

Number of training	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth
p_i	0.93	0.91	0.98	1.01	0.88	1.00	1.02	0.98	0.95	0.75
q_i	−0.84	−0.83	−0.78	−0.70	−0.91	−0.81	−0.72	−0.95	−0.81	−0.83

GBDT instead of the BPNN, the bootstrap-ELM model, and the bootstrap-GBDT model. The PI constructed in Figure 8(a) is wider and cannot contain most of the measured values. The PI in Figure 8(b) is narrower, but its PI width is fixed, which is not conducive to the uncertainty analysis. In contrast, the PI constructed by the bootstrap method is more reasonable. As shown in Figures 8(c) and 8(d), the interval width of the bootstrap-GBDT model is narrower than that of the

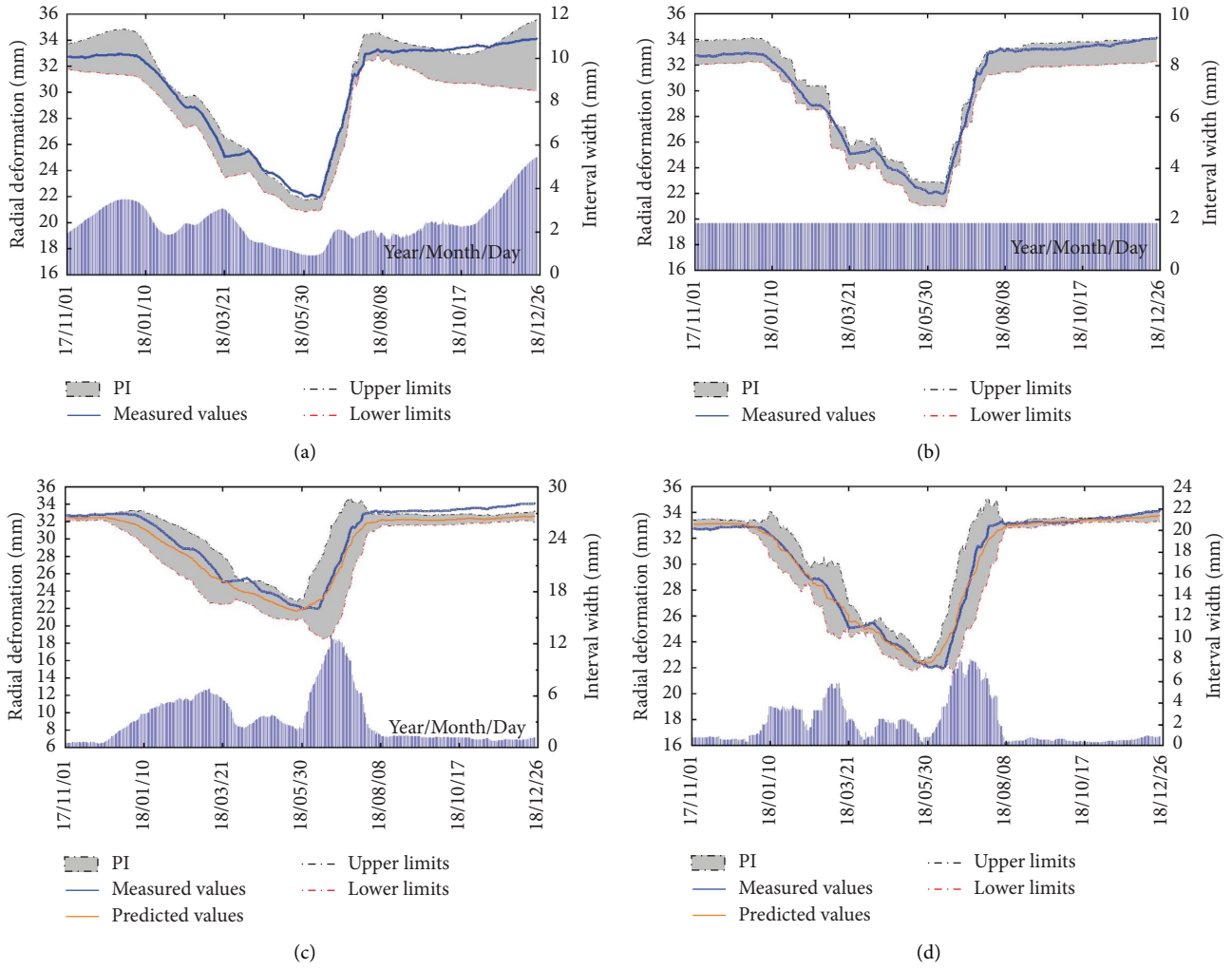

Figure 8: Prediction results for different interval prediction methods: (a) LUBE-BPNN model; (b) LUBE-GBDT model; (c) bootstrap ELM; (d) bootstrap GBDT.

bootstrap-ELM model, mainly due to different approximation levels of the function of different algorithms, also indicating the differences in the cognitive uncertainties. The results indicate that the proposed model can better estimate the variance of model bias. It is worth noting that the MAE, MAPE, and RMSE of the bootstrap-ELM model are 0.7142, 2.46%, and 0.9280, respectively. Combining Figure 5(d) and Table 1, it can be observed that its point prediction performance is better than that of the ELM model.

The quantitative indexes and the computation cost for the interval prediction of different models are listed in Table 3. The PICP of the LUBE-BPNN model, the LUBE-GBDT model, and the bootstrap-ELM model is 68.8%, 82.2%, and 67.8%, respectively. Their PIs cannot cover most of the measured values. However, the coverage of the bootstrap-GBDT model for the measured values reached 100%, and its CWC is the smallest, 1.22; i.e., the bootstrap-GBDT model can quantify the above uncertainties by the PI effectively. Furthermore, as shown in Figure 8(d), during the rising and falling periods of the measured time series, caused by the environmental factors, the width of the PI is always larger than that of other periods. The variation in the PI can indicate the effect of the change in the reservoir water level and other factors of the arch dam; i.e., the above uncertainties are well characterized by the PI through its widths.

The interval prediction for the whole selected period of the arch dam is shown in Figure 9. During the periods of drastic changes of the dam radial deformation and fluctuating periods, the PI widths are larger, indicating considerable influence of various uncertainties. Moreover, the reliability of the proposed model can also quantify these uncertainties during the smooth periods of the measured time series.

4.3.2. Exploration of the Stochastic Uncertainties. In order to explore the stochastic uncertainties of the arch dam, the bootstrap-GBDT model is trained again by adding the Gaussian noise into the measured time series. The results are further taken to investigate the robustness of the proposed model. As shown in Figure 10, the testing set remains unchanged when only the Gaussian noise is added into the training set to reduce the influence of other factors.

Table 3: Evaluation results for different interval prediction methods.

Evaluation indexes	LUBE-BPNN model	LUBE-GBDT model	Bootstrap-ELM model	Bootstrap-GBDT model
PICP (%)	68.8	82.2	67.8	100
PINAW	1.439	1.15	1.99	1.22
CWC	13.39	5.28	8.85	1.22
Time (min)	17.5	16.3	11.7	18.6

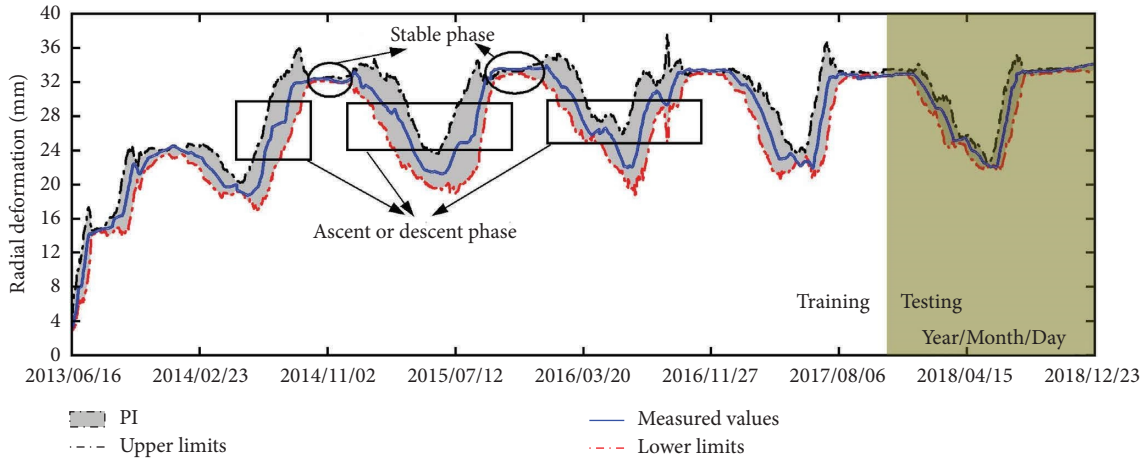

Figure 9: The interval prediction results for the whole measured time series.

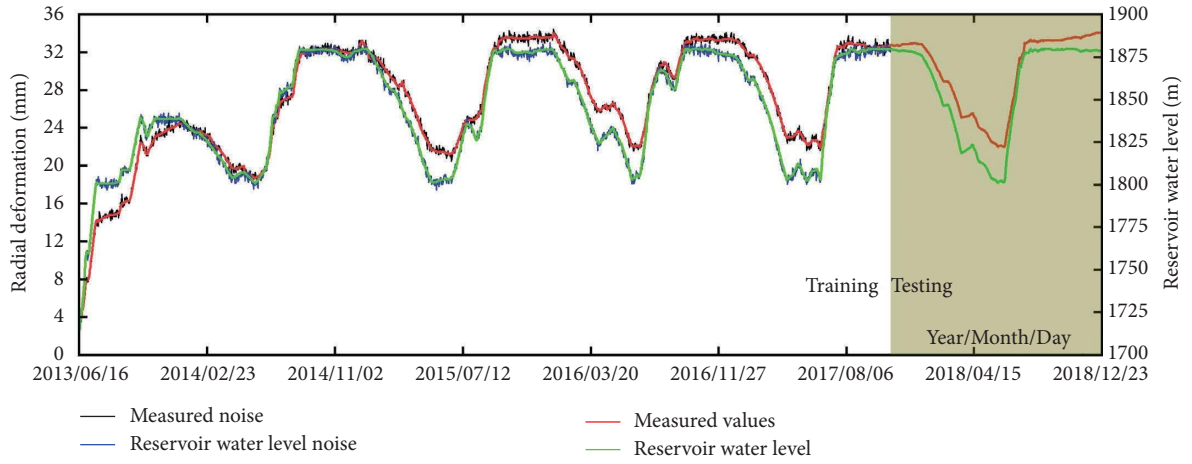

Figure 10: The measured time series after adding the Gaussian noises.

Following the aforementioned parameter settings, the bootstrap-GBDT model is trained using the noise data to implement the interval prediction for dam deformation, as shown in Figures 11(a) and 11(b). The width of the PI after adding noises has a certain increase, while the point prediction results do not change visibly. When the noise data have increased the stochastic uncertainties, the certain change in CWC indicates that added low noises can be identified by the interval prediction. The constructed model can also estimate the variance of added noises, which is feasible.

Evaluation indexes of the interval prediction after adding the Gaussian noises are listed in Table 4. Comparison with Table 1, the difference is not significant in the point prediction results whether there are the noise data or not. The results indicate that the constructed model on the point prediction is not sensitive to these noises; i.e., the point prediction fails to indicate the negative effect of additional noises. Since PIs can offer more information than the point prediction, the reasonable results can be achieved through combining the point prediction and the interval prediction to appraise the dam performance during long-term operation.

4.3.3. Exploration of the Cognitive Uncertainties. The degree of the cognitive uncertainties varies for different dams. Hence, the cognitive uncertainties are further explored by the interval

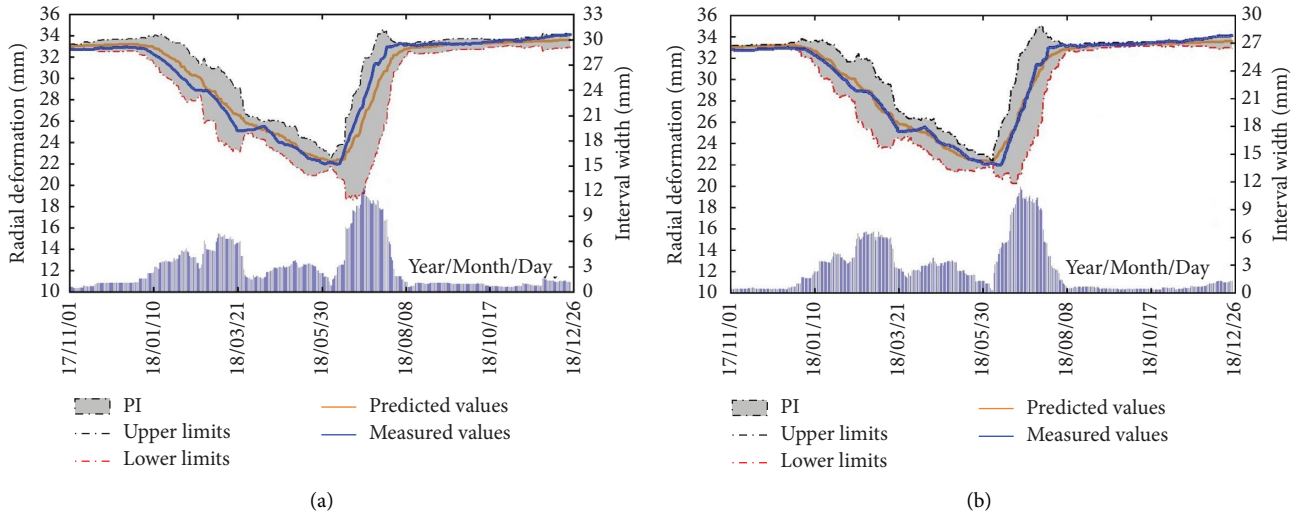

Figure 11: The interval prediction results after adding the Gaussian noises: (a) noisy deformation time series; (b) noisy upstream water level.

Table 4: Evaluation indexes of the interval prediction after adding the Gaussian noises.

Evaluation indexes	PICP (%)	PINAW	CWC	MAE	MAPE (%)	RMSE
Dam deformation	100	1.63	1.63	0.52	1.8	0.73
Reservoir water level	98.7	1.53	1.53	0.31	1.1	0.39

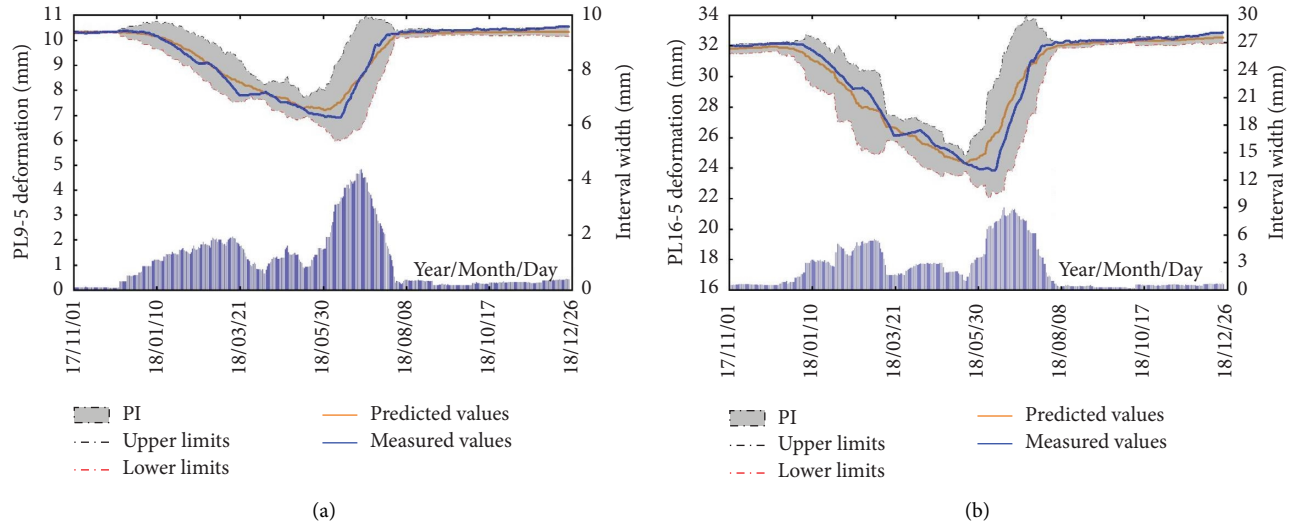

Figure 12: The interval prediction results of other monitoring points with different locations: (a) PL9-5; (b) PL16-5.

prediction results of PL9-5 and PL16-5 with different locations. The results of their interval prediction are shown in Figures 12(a) and 12(b), respectively. The evaluation indexes are listed in Table 5. Combined with Figure 8(d), it can be observed that the PI of the two points is similar to that of PL13-5, which can cover most of the original measured values and obtain smaller CWC to construct a high-quality PI. In the interval prediction of PL9-5, PICP reaches 100%, indicating full coverage of the measured values, and CWC is 0.70, which is smaller than all other points due to the closeness to rock foundation. In these areas, the effect of the above uncertainties is smaller, since the change in the reservoir water level and temperature have relatively less influence than the other two points. The interval prediction results of PL16-5 are similar to those of PL13-5 due to semblable cognitive uncertainties. The above results show that the interval prediction model can reasonably estimate the cognitive uncertainties of the whole arch dam.

Table 5: The evaluation indexes of the other monitoring points with different locations.

Monitoring points	PICP (%)	PINAW	CWC	MAE	MAPE (%)	RMSE
PL9-5	100	0.63	0.63	0.14	1.7	0.19
PL16-5	100	1.12	1.12	0.41	1.4	0.59

5. Conclusions

The present work explored the impact of uncertainties in prediction models on the structural deformation behavior of dams using the measured time series, by proposing an innovative interval prediction model that combines the GBDT and the bootstrap method. The GBDT model is suitable for the prediction of dam deformation behavior since it can accurately indicate the nonlinear relationship between the measured effects and the environmental factors. In particular, compared with other prediction models, the GBDT model can also quantify the importance of input variables. Different from the regression tree approach, the GBDT model adopts the forward distribution algorithm, and the learning rate is taken to avoid overfitting. The established model is general and can successfully quantify the uncertainties in the form of the improved PIs, where high-quality PIs can be obtained including the majority of the measured values and with smaller PINAW. They can also provide accurate point prediction results.

The methodology was validated and applied herein to the Jinping I Arch Dam, whose performance was characterized via the structural deformation of a set of monitoring points in the perpendicular line monitoring system. Their measured time series were postprocessed by PCA to reduce multicollinearity and data dimensionality. Structural performance prediction was carried out using different point prediction methods and interval prediction methods, in order to investigate the impact of stochastic and cognitive uncertainties of the arch dam. SOA was taken to search the optimal parameters of the GBDT. The results show that the variation in the improved PI can indicate the effect of the change of the environmental factors. During periods of dramatic changes in the measured time series, the width of PIs is always larger than that of other periods; thus, the PI can reasonably quantify the uncertainties caused by these complex factors. Furthermore, the PI based on the bootstrap-GBDT model plays a key role in quantifying the uncertainties for the arch dam during long-term operation, since it indicates the model errors and random noises. The results of the case study proved that the proposed model can appropriately account separately for the stochastic or cognitive uncertainties, which are reflected in predicted values for each monitoring point closer to the measured values. The increased level of the stochastic or cognitive uncertainties can be better identified by the constructed PI. In future works, since uncertainty quantification is very practical and of current interest during dam operation, panel data models or other suitable models are still required to improve the generalization ability of the bootstrap-GBDT model.

Data Availability

The research data used to support the findings of this study have not been made available because the data are confidential for the project.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Erfeng Zhao conceptualized the study, designed the methodology, prepared the software, wrote the original draft of the manuscript, and acquired the funding. Yi Li curated the data, prepared the software, wrote the original draft of the manuscript, and validated the results. Jingmei Zhang revised the manuscript, addressed reviewers' comments, and acquired the funding. Zhangyin Li supervised the study and revised the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (52079046), the Belt and Road Special Foundation of the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (2021492211), and the Science and Technology Project of Zhejiang Water Resources Department (RC2244).

References

[1] H. Su, X. Yan, H. Liu, and Z. Wen, "Integrated multi-level control value and variation trend early-warning approach for deformation safety of arch dam," *Water Resources Management*, vol. 31, pp. 2025–2045, 2017.

[2] S. Zheng, C. Shao, C. Gu, and Y. Xu, "An automatic data process line identification method for dam safety monitoring data outlier detection," *Structural Control and Health Monitoring*, vol. 29, no. 7, Article ID e2948, 2022.

[3] E. Zhao and C. Wu, "Centroid deformation-based nonlinear safety monitoring model for arch dam performance evaluation," *Engineering Structures*, vol. 243, Article ID 112652, 2021.

[4] H. Su, X. Li, B. Yang, and Z. Wen, "Wavelet support vector machine-based prediction model of dam deformation," *Mechanical Systems and Signal Processing*, vol. 110, pp. 412–427, 2018.

[5] M. Li, Y. Shen, Q. Ren, and H. Li, "A new distributed time series evolution prediction model for dam deformation based on constituent elements," *Advanced Engineering Informatics*, vol. 39, pp. 41–52, 2019.

[6] S. Gamse, W. Zhou, F. Tan, K. Yuen, and M. Oberguggenberger, "Hydrostatic-season-time model updating using Bayesian model class selection," *Reliability Engineering and System Safety*, vol. 169, pp. 40–50, 2018.

[7] P. Milillo, D. Perissin, J. T. Salzer et al., "Monitoring dam structural health from space: insights from novel InSAR techniques and multi-parametric modeling applied to the Pertusillo dam Basilicata, Italy," *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, pp. 221–229, 2016.

[8] D. Yuan, C. Gu, B. Wei, X. Qin, and H. Gu, "Displacement behavior interpretation and prediction model of concrete gravity dams located in cold area," *Structural Health Monitoring*, Article ID 147592172211223, 2022.

[9] S. Wang, C. Xu, C. Gu, H. Su, and B. Wu, "Hydraulic-seasonal-time-based state space model for displacement monitoring of high concrete dams," *Transactions of the Institute of Measurement and Control*, vol. 43, no. 15, pp. 3347–3359, 2021.

[10] Y. Zhu, C. Gu, E. Zhao, J. Song, and Z. Guo, "Structural safety monitoring of high arch dam using improved ABC-BP model," *Mathematical Problems in Engineering*, vol. 2016, Article ID 6858697, 9 pages, 2016.

[11] D. Yang, C. Gu, Y. Zhu et al., "A concrete dam deformation prediction method based on LSTM with attention mechanism," *IEEE Access*, vol. 8, pp. 185177–185186, 2020.

[12] B. Dai, C. Gu, E. Zhao, K. Zhu, W. Cao, and X. Qin, "Improved online sequential extreme learning machine for identifying crack behavior in concrete dam," *Advances in Structural Engineering*, vol. 22, no. 2, pp. 402–412, 2019.

[13] S. Chen, C. Gu, C. Lin, E. Zhao, and J. Song, "Safety monitoring model of a super-high concrete dam by using RBF neural network coupled with kernel principal component analysis," *Mathematical Problems in Engineering*, vol. 2018, Article ID 1712653, 13 pages, 2018.

[14] F. Salazar, R. Morán, M. Toledo, and E. Oñate, "Data-based models for the prediction of dam behaviour: a review and some methodological considerations," *Archives of Computational Methods in Engineering*, vol. 24, no. 1, p. 21, 2015.

[15] Z. Jun, M. Jian, X. Jiemin, and M. Zuohong, "Displacement prediction model for concrete dam based on PSO-GBDT," *IOP Conference Series: Earth and Environmental Science*, vol. 358, no. 5, Article ID 052043, 2019.

[16] W. Lei and J. Wang, "Dynamic stacking ensemble monitoring model of dam displacement based on the feature selection with PCA-RF," *Journal of Civil Structural Health Monitoring*, vol. 12, no. 3, pp. 557–578, 2022.

[17] S. Gokmen, R. Dagalp, and S. Kilickaplan, "Multicollinearity in measurement error models," *Communications in Statistics - Theory and Methods*, vol. 51, no. 2, pp. 474–485, 2020.

[18] Y. Hu, C. Shao, C. Gu, and Z. Meng, "Concrete dam displacement prediction based on an ISODATA-GMM clustering and random coefficient model," *Water*, vol. 11, no. 4, p. 714, 2019.

[19] H. Yu, Z. Wu, T. Bao, and L. Zhang, "Multivariate analysis in dam monitoring data with PCA," *Science China Technological Sciences*, vol. 53, no. 4, pp. 1088–1097, 2010.

[20] N. Dewolf, B. D. Baets, and W. Waegeman, "Valid prediction intervals for regression problems," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 577–613, 2023.

[21] Q. Ren, M. Li, and Y. Shen, "A new interval prediction method for displacement behavior of concrete dams based on gradient boosted quantile regression," *Structural Control and Health Monitoring*, vol. 29, no. 1, Article ID e2859, 2022.

[22] Q. Ren, M. Li, R. Kong, Y. Shen, and S. Du, "A hybrid approach for interval prediction of concrete dam displacements under uncertain conditions," *Engineering with Computers*, vol. 39, no. 2, pp. 1285–1303, 2021.

[23] Y. Xu, C. Mi, Q. Zhu, J. Gao, and Y. He, "An effective high-quality prediction intervals construction method based on parallel bootstrapped RVM for complex chemical processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 171, pp. 161–169, 2017.

[24] J. Ma, H. Tang, X. Liu et al., "Probabilistic forecasting of landslide displacement accounting for epistemic uncertainty: a case study in the Three Gorges Reservoir area, China," *Landslides*, vol. 15, no. 6, pp. 1145–1153, 2018.

[25] J. Liu, X. Qin, Y. Sun, and Q. Zhang, "Interval early warning method for state of engineering structures based on structural health monitoring data," *Structural Control and Health Monitoring*, vol. 29, no. 8, Article ID e2935, 2022.

[26] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[27] J. Cara, "Modal identification of structures from input/output data using the expectation-maximization algorithm and uncertainty quantification by mean of the bootstrap," *Structural Control and Health Monitoring*, vol. 26, no. 1, Article ID e2272, 2019.

[28] Z. Zhang and C. Jung, "GBDT-MO: gradient-boosted decision trees for multiple outputs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3156–3167, 2021.

[29] N. Kumar Bhagat, A. Mishra, R. Singh, C. Sawmliana, and P. Singh, "Application of logistic regression, CART and random forest techniques in prediction of blast-induced slope failure during reconstruction of railway rock-cut slopes," *Engineering Failure Analysis*, vol. 137, Article ID 106230, 2022.

[30] J. Hu and F. Ma, "Comparison of hierarchical clustering based deformation prediction models for high arch dams during the initial operation period," *Journal of Civil Structural Health Monitoring*, vol. 11, no. 4, pp. 897–914, 2021.

[31] C. Gu and E. Zhao, *Dam Safety Monitoring Theory and Methods*, p. 117, Hohai University Press, Nanjing, China, 2006.

[32] G. Dhiman and V. Kumar, "Seagull optimization algorithm: theory and its applications for large-scale industrial engineering problems," *Knowledge-Based Systems*, vol. 165, pp. 169–196, 2019.

[33] Y. Yu, X. Liu, E. Wang, K. Fang, and L. Huang, "Dam safety evaluation based on multiple linear regression and numerical simulation," *Rock Mechanics and Rock Engineering*, vol. 51, no. 8, pp. 2451–2467, 2018.

[34] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[35] S. Jiang, L. Zhao, and C. Du, "Structural deformation prediction model based on extreme learning machine algorithm and particle swarm optimization," *Structural Health Monitoring*, vol. 21, no. 6, pp. 2786–2803, 2022.

[36] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower upper bound estimation method for construction of neural network-based prediction intervals," *IEEE Transactions on Neural Networks*, vol. 22, no. 3, pp. 337–346, 2011.