IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 21, 2021, accepted January 31, 2021, date of publication February 2, 2021, date of current version February 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056588

A Novel Seepage Behavior Prediction and Lag Process Identification Method for Concrete Dams Using HGWO-XGBoost Model

KANG ZHANG, CHONGSHI GU, YANTAO ZHU, SIYU CHEN, BO DAI, YANGTAO LI, AND XIAOSONG SHU

State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China
College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing 210098, China

Corresponding authors: Chongshi Gu (csgu@hhu.edu.cn) and Yantao Zhu (zyt50@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51739003, Grant 51909173 and Grant U2040223, in part by the Free Exploration Project of Hohai University under Grant B200201058, in part by the Basic Research Project Funded of National Key Laboratory under Grant 20195025912, in part by the Open Foundation of Changjiang survey, planning, design and Research Co., Ltd under Grant CX2019K01.

ABSTRACT Seepage monitoring is a vital task in the risk management of concrete dams. Considering the lag effect of input factors, this paper presents a novel seepage monitoring model for concrete dams and proposes an effective identification method of lag process. Firstly, extreme gradient boosting (XGBoost) were adopted to predict the dam seepage. Hybridizing grey wolf optimization (HGWO) which integrates differential evolution (DE) into grey wolf optimization (GWO) and five–fold cross validation were utilized to optimize the hyper–parameters of XGBoost. Secondly, under the same search range and four evaluation indicators, the models optimized respectively by HGWO and three other algorithms were compared to confirm the global optimization capability of HGWO. Six state–of–art methods were also introduced to verify the effectiveness and feasibility of the proposed model. Then, based on the computation method of factor importance in decision tree models, we evaluated the relative importance of each component in the proposed model. Finally, according to the factor importance, the lag process of upstream water level and rainfall was identified, meanwhile a new equivalent water level calculation method is proposed. Monitoring data from three piezometric tubes on a concrete dam were taken as the experimental object. The results show that the improved HGWO has stronger global optimization ability, and the HGWO–XGBoost model achieves satisfactory prediction for seepage in concrete dams. Compared with the traditional trial–and–error method, the lag process computation method proposed in this paper provides a better recognition effect, which is of great value to the seepage monitoring and control of concrete dams.

INDEX TERMS Dam seepage prediction, lag process identification, extreme gradient boosting, hybridizing grey wolf optimization, factor importance.

I. INTRODUCTION

Dam safety monitoring is an essential means to control risk and understand the operational status of a dam. A large number of instruments are embedded in dams to monitor all aspects of daily behavior [1]. Seepage control is one of the most important tasks in dam surveillance [2], [3]. Therefore, in order to grasp the seepage law of dams, it is essential to establish a reasonable monitoring model on the basis of massive data.

Generally, the seepage is related to the loads which the dam received, and the size of dam seepage directly reflects the anti–seepage and drainage performance of dams. Previous research results have indicated that water level, rainfall, temperature and aging are main factors influencing dam seepage [4], [5]. However, the water pressure transfer and dissipation of an unsaturated body take a certain time. Compared with dam displacement, the lag effect of factors makes

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

the dam seepage model more complicated, which causes the seepage more difficult to predict.

Traditional statistical models mostly adopt multiple regression or stepwise regression methods, assuming that the variables are independent, to solve the mapping relation between dam seepage and explanatory variables [6]–[8]. But this assumption is often invalid due to the existence of multicollinearity among variables, which may severely limit the robustness and accuracy of the model [9]–[11]. Furthermore, considering the lag effect, a mass of previous water level and rainfall factors should be included in the statistical model, so as to make the model prone to ill–conditioned problems.

For solving those problems, Wu and Gu greatly simplified factors by exploiting the previous segmental average values or equivalent values of the upstream reservoir water level and rainfall based on the lag influence function, to establish different forms of seepage monitoring models [12]. But even so, the information loss in the reduction process, as well as the nonlinear relationship between explanatory variables and dependent variables, still restrict the accuracy of the model to a certain extent.

With the rapid development of computer technology, plenty of data–driven machine learning algorithms have been proposed constantly. Although large amounts of computation are usually required, high accuracy attracts scholars to gradually shift their research focus to these methods [13]–[17].

In recent years, some research on the application of machine learning methods has been carried out to analyse the dam seepage, and positive results have been achieved. Sharghi *et al*. [18] respectively adopted neural network, support vector machine and adaptive neural fuzzy inference system to establish a monitoring model of piezometric heads in a earthfill dam, and integrated the three methods to improve the prediction performance. Shi *et al*. [19] used radial basis function neural network which optimized by genetic algorithm to predict the seepage discharge of the concrete face rockfill dam. Based on the data mining method, Hu and Ma established a zoned safety monitoring model to estimate the uplift pressures of concrete dams [20]. Nourani *et al*. [21] proposed a statistical model of the piezometric tube water level by using feed–forward back–propagation and radial basis function neural network. Chen *et al*. [22] developed kernel extreme learning machine to study the dam leakage, and employed global sensitivity analysis to evaluate the importance of each input factor. Salazar *et al*. [23] analyzed the leakage problem of the La Baells Dam with the application of boosted regression trees and explored the interpretability of the model. Wang *et al*. [24] decomposed piezometric tube water level and reservoir water level into the form of base value plus daily variation, then the seepage statistical model is proposed by combining linear regression and support vector regression (SVR).

It can be seen from the literature that, compared with traditional statistical methods, machine learning algorithms can deeply explore the implicit relationship among variables and have a better performance in prediction. But at the same time, the interpretability of models is usually sacrificed. Although some scholars have made some attempts to improve the interpretability of dam seepage monitoring model (such as [22]–[24]), most of them merely evaluated the importance of the previous segmental average values of input factors. However, average value is a relatively general concept which cannot directly reflect the specific lag process of relevant variables.

In order to accurately evaluate the lag behavior of dam seepage, this study takes the piezometric tubes of a dam as an example. Based on extreme gradient boosting (XGBoost) which has been recognized as one of the most advanced algorithms in the field of machine learning in recent years [25]–[28], a monitoring model of the uplift pressure of concrete dam is established with consideration of the time lag effect. Under the premise of cross–validation, hybridizing grey wolf optimization (HGWO) is applied to optimize the hyper–parameters in the XGBoost model. Then four quantitative evaluation indicators are utilized to comprehensively evaluate the advantages of HGWO and proposed model by introducing six state–of–art baseline models. According to the growth process of decision tree [29], [30], the lag process of upstream water level and rainfall are identified, then a new computation method of equivalent water level is proposed.

The structure of this paper is as follows: Section II briefly describes the relevant methods and theories involved in this study. In Section III, combined with a specific engineering case, the research design, implementation of the model and experimental results in detail are introduced. Finally, main conclusions and future work are drawn in Section IV.

II. METHODOLOGY

A. VARIABLES CONSIDERED FOR THE SEEPAGE MONITORING MODEL OF CONCRETE DAMS

The measured data shows that the uplift pressure of concrete dam foundation is mainly affected by upstream and downstream water level. The slope seepage induced by rainfall also has a certain influence on the uplift pressure. The fissure size of bedrock will change as a result of temperature fluctuation. In addition, given the degradation of impermeable material performance, the aging component should also be selected. Therefore, the traditional hydraulic, precipitation, temperature, and time effect (HPTT) statistical model of piezometric tube water level can be expressed as [2], [12]:

$$Y = Y_{Hu} + Y_{Hd} + Y_p + Y_T + Y_\theta \tag{1}$$

where Y is the piezometric tube water level at the monitoring point; Y_{Hu} and Y_{Hd} respectively denote the upstream and downstream water level components; Y_p, Y_T and Y_θ refer to the rainfall component, temperature component and aging component, respectively.

1) UPSTREAM WATER LEVEL COMPONENT Y_{Hu}

$$Y_{Hu} = f\,(H_1, H_2, \cdots, H_i, \cdots, H_n, \cdots \cdots, w_1, w_2, \cdots, w_i, \cdots, w_n) \tag{2}$$

where $H_i(i = 1, 2, \cdots, n)$ represents the ith previous water level; w_i is the weight of H_i and $\sum_{i=1}^{n} w_i = 1$, thus the upstream water level component can be obtained as:

$$Y_{Hu} = a_u \sum_{i=1}^{n} w_i H_i = a_u \bar{H}_u \tag{3}$$

where a_u is the regression coefficient; $\bar{H}_u$ denotes the equivalent value of the upstream reservoir water level. It is generally believed that the influence process of upstream water level on dam seepage basically follows a normal distribution (see Figure 1). The weight of upstream water level on the tth day before the monitoring day can be expressed as [2], [12]:

$$w(t) = \frac{1}{\sqrt{2\pi}x_2} \exp\left(\frac{-(t - x_1)^2}{2x_2^2}\right) \tag{4}$$

where x_1 is the number of lag days of the upstream water level on uplift pressure; x_2 is the number of influence days of the upstream water level, both of them need to be calculated by trial [31], [32]; $w(t)$ obeys $\int_{-\infty}^{0} w(t)dt = 1$. The equivalent value of the upstream reservoir water level can be expressed as:

$$\bar{H}_u = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}x_2} \exp\left(\frac{-(t - x_1)^2}{2x_2^2}\right) H(t)dt \tag{5}$$

Generally, there is only one measured value of upstream water level in a monitoring day. Therefore, in practical application, continuous integral can be changed into discrete integral, and the integral interval can be taken as $2 \sim 3$ times of x_2.

2) DOWNSTREAM WATER LEVEL COMPONENT Y_{Hd}

The downstream water level also has lag effect on uplift pressure. But since the downstream water level is generally measured fewer times and fluctuated gently, only the downstream water level of the monitoring day is taken as a factor [2]. The downstream water level component can be written as:

$$Y_{Hd} = a_d H_d \tag{6}$$

where a_d is the regression coefficient; H_d is the downstream water level of the monitoring day.

3) RAINFALL COMPONENT Y_{Hp}

During the course of rainfall, part of the precipitation causes the surface runoff which will flow into reservoir to change the water level, the rest of them produces groundwater. Then the groundwater percolates through the bedrock joints and fissures to affect the uplift pressure of dam foundation. There is obvious lag effect between the uplift pressure and rainfall. Similar to the upstream water level component, the expression form of rainfall component can be obtained as:

$$\begin{aligned} Y_p &= d_i \bar{P} \\ &= d_i \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}x_4} \exp\left(\frac{-(t - x_3)^2}{2x_4^2}\right) [P(t)]^{2/5} dt \end{aligned} \tag{7}$$

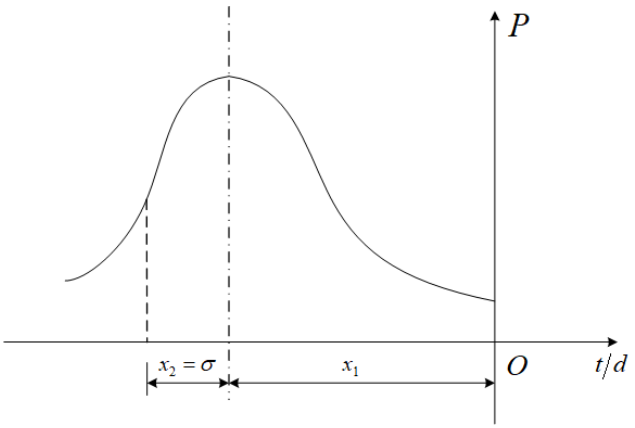


FIGURE 1. The influence process of upstream water level on dam seepage.

where d_i is the regression coefficient; $\bar{P}$ denotes the equivalent value of rainfall; x_3 and x_4 are the number of lag days and influence days of rainfall, respectively.

4) TEMPERATURE COMPONENT Y_{HT}

The bedrock fissures are affected by the temperature fluctuation, thus the uplift pressure of the dam foundation changes correspondingly. The temperature of bedrock basically cycle with the annual period. In the absence of measured bedrock temperature, the simple harmonic wave of multi–period can be used as the temperature variables [2], [12]. The downstream water level component can be expressed as:

$$Y_T = \sum_{i=1}^{n} \left(b_{1i} \sin\frac{2\pi it}{365} + b_{2i} \cos\frac{2\pi it}{365}\right) \tag{8}$$

where $n = 1$ or 2; t is the cumulative days from the initial monitoring day; $b_{1\,i}$ and $b_{2\,i}$ are the regression coefficients.

5) AGING COMPONENT Y_θ

The aging component is an important component of uplift pressure. It changes rapidly at the initial impounding stage, and then becomes stable with time. The common expression is as follows:

$$Y_\theta = c_1\theta + c_2 \ln\theta \tag{9}$$

where θ is the cumulative days from the initial monitoring day divided by 100; c_1 and c_2 are the regression coefficients. To sum up, the traditional statistical model of the piezometric tube water level at a single monitoring point located in dam foundation can be expressed as follows:

$$\begin{aligned} Y = {} & a_0 + a_u \bar{H}_u + a_d H_d + d_i \bar{P} \\ & + \sum_{i=1}^{n} \left(b_{1i} \sin\frac{2\pi it}{365} + b_{2i} \cos\frac{2\pi it}{365}\right) + c_1\theta + c_2 \ln\theta \end{aligned} \tag{10}$$

where a_0 is the constant term, the remaining symbols have the same meaning as above.

B. THE METHODOLOGY OF XGBoost

1) THE BOOSTING TREE METHOD

The boosting method based on decision trees is called the boosting tree algorithm, which is a binary classification tree for classification problems and a binary regression tree for regression problems. It is considered as one of the methods with the best performance in statistical learning [33].

The purpose of machine learning is usually using training data to find the mapping relationship between independent variables and response variables. Different from the traditional machine learning methods, the optimization of boosting tree algorithm is carried out in the function space. In order to find the optimal mapping function, an objective function $L(y, F(x))$ is usually defined, to solves $F(x)$ by minimizing the expected value of the objective function on the joint distribution of (x, y):

$$\begin{aligned}\arg\min_F \Phi(F) &= \arg\min_F E_{y,x} L(y, F(x)) \\ &= \arg\min_F E_X \left[E_y(L(y, F(x))) \mid x\right] \end{aligned} \quad (11)$$

or

$$\arg\min_F \phi(F(x)) = \arg\min_F E_y[L(y, F(x)) \mid x] \quad (12)$$

As for a given sample combination:

$$\begin{aligned} T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \\ (x_i \in \chi \subseteq R^n, y_i \in Y \subseteq R) \end{aligned} \quad (13)$$

where $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ik}\}$ and y_i are the eigenvector and response variable of the ith sample, respectively. When the joint distribution of (x, y) is estimated by finite data samples $\{x_i, y_i\}_1^N$, instead of the above parameterized form for accurate computation, $E_y[\cdot \mid x]$ can only be obtained by optimizing the objective function through the following parameterized form:

$$\begin{aligned}(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + \beta h(x_i; a)\right) \\ (m = 1, 2, \cdots, M) \end{aligned} \quad (14)$$

where $h(x; a)$ denotes the weak learner which is the decision tree in this case; β is the weight of $h(x; a)$; a refers to the parameters of decision trees; $F_{m-1}(x)$ is the model of step $m-1$.

To prevent overfitting, Friedman [34] proposed a shrinkage method which helped the algorithm gradually approximate the result by applying a small weight to each decision tree, similar to β of equation (14). Due to the existence of the shrinkage method, after the completion of each iteration, there will be enough space left for the boosting of the subsequent decision trees.

Therefore, the boosting tree method can be expressed as the additive model of decision trees:

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \quad (15)$$

where ρ denotes the learning rate, that is, the weight of weak learners.

2) XGBoost

XGBoost model can be expressed as:

$$\hat{y}_i = \phi(X_i) = \sum_{k=1}^{K} \rho_k f_k(X_i, a_k) \quad (16)$$

where $\hat{y}_i$ is the predicted value of the ith sample; f_k denotes the kth decision tree; ρ and a have the same meaning as the corresponding symbols in equation (15).

In order to get the optimal model, the objective function can be written as:

$$\begin{aligned} L(\phi) = \sum_{i=1}^{N} l\left(\hat{y}_i, y_i\right) + \sum_{k=1}^{K} \Omega(f_k) \\ \text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{aligned} \quad (17)$$

where y_i is the measured value of the ith sample; T is the number of leaf nodes in a single tree; w refers to the vector of scores on leaf nodes of single tree; l is the loss function between the predicted value and the measured value of the sample; Ω denotes the regularization term to prevent overfitting; γ and λ are coefficients.

The boosting tree method adopts forward stagewise algorithm. Assuming $F_0(x) = f_0(x)$ as the initial value of the model, each time a decision tree is obtained by solving the objective function to update the current model, and the optimal model $F(x)$ is obtained after several iterations.

When the structural form of the objective function is relatively simple, equation (14) is easy to solve. Under the circumstances, the direction in which the current objective function approaches the minimum is the global optimal direction. But for the general form, it is difficult to solve. In terms of this issue, the XGBoost model uses the gradient boosting method to optimize the objective function. The second–order Taylor expansion of the objective function under the tth iteration can be expressed as:

$$\begin{aligned} L^{(t)} \cong \sum_{i=1}^{N} \left[l\left(\hat{y}_i^{(t-1)}, y_i\right) + g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] \\ + \Omega(f_t) \end{aligned} \quad (18)$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(\hat{y}_i^{(t-1)}, y_i\right)$; $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(\hat{y}_i^{(t-1)}, y_i\right)$. When iterating to the tth round, $l\left(\hat{y}_i^{(t-1)}, y_i\right)$ is explicit. Assuming that I_j is an instance of the jth leaf node, w_j is the value on the jth leaf node. Hence equation (18) can be simplified as:

$$\begin{aligned} \hat{L}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i=I_j} g_i \right) \cdot w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \cdot w_j^2 \right] \\ + \gamma T \end{aligned} \quad (19)$$

Set $\partial \hat{L}^t / \partial w_j$ to be equal to 0, the optimal vector of scores on leaf nodes can be obtained as:

$$w^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (20)$$

By substituting equation (20) into equation (19), equation (19) can be written as:

$$\hat{L}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i=I_j} g_i\right)^2}{\sum_{i=I_j} h_i + \lambda} + \gamma T \tag{21}$$

Finally, the optimization of the objective function is transformed into the optimization of equation (21). Equation (21) can also be used to evaluate the structure quality of decision trees under the tth round iteration in XGBoost model. When the structure of trees is determined, the quality is only related to the first and second derivatives of the loss function.

C. FACTORS IMPORTANCE COMPUTATION

There are several alternative ways to calculate feature importance. The importance of a certain feature can be described by the total number of times that it is selected during the growth of decision trees. Beyond that, the feature importance can also be defined as the proportion of the model improvement to the total improvement across all splits, when a certain feature is selected as the segmentation variable [35]. The improvement indexes can be reduced error, information gain or reduction of Gini index and so on.

The variation amplitude of these indexes decrease gradually during the generation of gradient boosting decision trees. Therefore, the improvement during node splits of the first few decision trees plays a decisive role in the computation result of factor importance. However, XGBoost takes a subsample approach of samples and features in each node split. That is, when we take the magnitude of improvement in metrics as the definition of feature importance in the XGBoost model, the results may depend on the generation of the first few decision trees, which is usually contingent and not sufficient to reflect the actual situation. Given those, we describe the feature importance through the number of times that a certain feature is selected as segmentation variable in this study.

For the ensemble learning algorithms which build on decision trees, Breiman *et al*. [36] proposed a method to define the importance of input factors:

$$I_k(F) = \frac{1}{M}\sum_{m=1}^{M} I_k\left(T_m\right) \quad = \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{J-1} 1_j\left(X^k\right) \tag{22}$$

where T_m is the mth decision tree in model F; M is the total number of trees; $I_k^2\left(T_m\right)$ denotes the importance of factor X^k in T_m; J is the number of internal nodes which include $J-1$ non–terminal nodes in T_m; $1_j\left(X^k\right)$ is the indicator function of whether to select variable X^k as the segmentation variable at node jth node.

D. THE METHODOLOGY OF HGWO

1) GWO

GWO is a new swarm intelligence optimization algorithm proposed in 2014 by Mirjalili *et al*. [37], which simulates the cooperative mechanism of gray wolf group during predation. The intelligent feedback mechanism and adaptive convergence factors enable it to achieve a balance between global search and local optimization. For this reason, GWO has good performance in solving precision and convergence speed with the characteristics of simple structure and easy implementation [38].

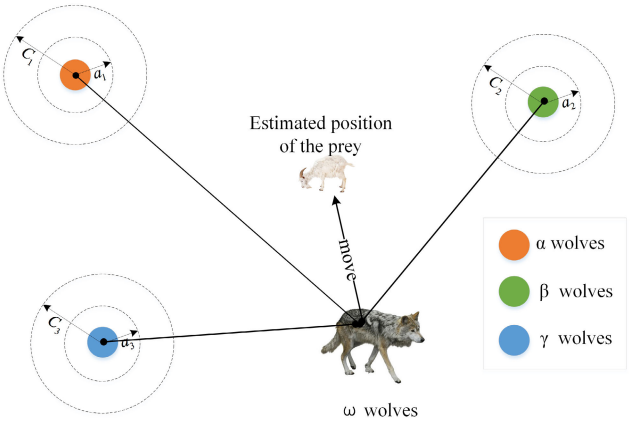


FIGURE 2. The diagram of GWO.

Gray wolf groups consist of α, β, δ and ω wolves. α wolves which usually denote the best solutions are the highest leaders of the pack; β wolves which assist α wolves are the sub–optimal individuals of wolves; δ wolves which represent the third fitness solutions need to carry out the orders of α wolves and β wolves, while managing ω wolves; ω wolves is the search individuals. α wolves, β wolves and δ wolves jointly assess the prey position and guide ω wolves to move, so as to realize all–round encircle of the prey, finally attack and capture the prey (see Figure 2).

GWO simulates gray wolf groups with the whole process of the social class, division of labor, and hunting prey, specifically including the following three steps:

1. Surround the prey: The gray wolves search and approach the prey gradually. The mathematical model can be written as:

$$\begin{cases} D = \left|C \cdot X_p(t) - X(t)\right| \\ X(t+1) = X_p(t) - A \cdot D \\ A = 2a \cdot r_1 - a \\ C = 2r_2 \end{cases} \tag{23}$$

where t is the number of iterations; X_p represents the position vector of the current prey; X denotes the position vector of grey wolf population; A and C are coefficient vectors; r_1 and r_2 are random vectors with the value interval of [0, 1], a decreases linearly from 2 to 0 with t.

Mathematically, the behavior of encircling prey can be simulated by reducing the value of a, the updated formula of a is as follows:

$$a(t) = 2 - \frac{2t}{M} \tag{24}$$

where M is the maximum number of iterations.

2. Hunting: Keep three of the fittest wolves in each iteration. According to their positions, the position vector of

grey wolves can be updated. The mathematical model can be expressed as:

$$\begin{cases} D_\alpha = |C_1 X_\alpha - X| \\ D_\beta = \left|C_2 X_\beta - X\right| \\ D_\delta = |C_3 X_\delta - X| \\ X_1 = X_\alpha - A_1 D_\alpha \\ X_2 = X_\beta - A_2 D_\beta \\ X_3 = X_\delta - A_2 D_\delta \\ X(t+1) = (X_1 + X_2 + X_3)/3 \end{cases} \tag{25}$$

where X_α, X_β and X_δ respectively represent the positions of α, β and δ wolves in the current population; X is the position vector of gray wolf population as before; D_α, D_β and D_δ respectively represent the distance between the candidate wolves and three fittest wolves.

3. Attack and search for prey: A is a random number on $[-a, a]$. The range of which will fluctuate accordingly with the linear decrease of a. When $|A| > 1$, gray wolves try to disperse in their respective areas to search for prey. On the contrary, gray wolves will launch attacks to hunt prey when $|A| < 1$.

2) DE

DE is a global optimization algorithm proposed by Storn and Price according to the mechanism of biological evolution which mainly includes mutation, crossover and selection operations [39], [40].

1.mutation operation:three different individual vectors X_{r_1}, X_{r_2} and X_{r_3} ($r_1 \neq r_2 \neq r_3 \neq i$) are randomly selected from the population in each generation. A new individual is constructed by one of them as the basis vector and the difference of the other two as the difference vector, which can be expressed as:

$$V_i(P+1) = X_{r_1}(P) + F_r \cdot \left(X_{r_2}(P) - X_{r_3}(P)\right) \tag{26}$$

where $V_i(P+1)$ refers to the position vector of the ith new individuals after mutation; P is the number of iterations; F_r denotes the scaling factor in the range of [0,2].

2.crossover operations: the test vector is generated by exchanging information between the new and former individuals, which can be expressed as:

$$U_{i,j}(P+1) = \begin{cases} V_{i,j}(P+1), & \text{if rand } (0,1) \leq C_r \text{ or } j = j_{\text{rand}} \\ X_{i,j}(P), & otherwise \end{cases} \tag{27}$$

where $U_{i,j}(P+1)$ represents the jth gene of the ith test vector; C_r is the crossover factor with the range of [0,1]; j_{rand} is the random integer within $[1, d]$ (d is the total number of genes). It ensures that at least one gene of the offspring comes from the mutant individual to avoids being caught in the local optimal solution.

3. selection operation: DE adopts greedy selection strategy which retains fitter individuals by one–to–one comparisons between the parent generation and the child generation. The selection rule is as follows:

$$X_i(P+1) = \begin{cases} U_i(P+1), & \text{if } f\left(U_i(P+1)\right) \leq f\left(X_i(P)\right) \\ X_i(P), & otherwise \end{cases} \tag{28}$$

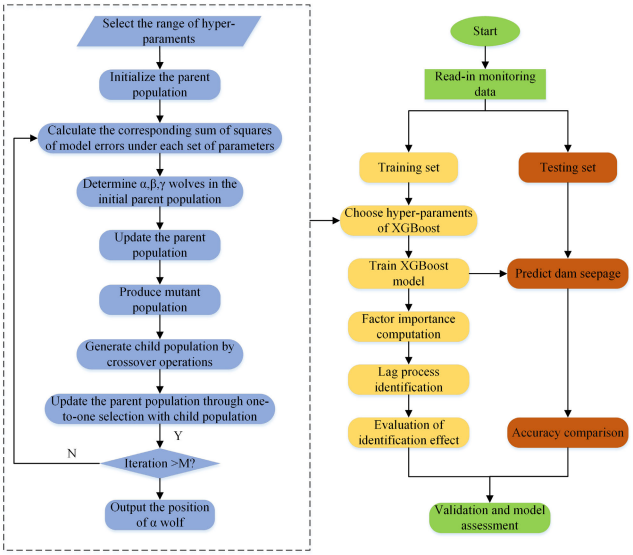


FIGURE 3. Flowchart of model implementation.

3) HGWO

GWO has a strong local optimization ability, whereas differential evolution can effectively help the algorithm jump out of the local optimal solution to achieve global search. Therefore, Zhu *et al*. [41] proposed a HGWO combining DE and GWO.

Step1: The parent population is randomly generated through equation (29), which can be expressed as:

$$X_i^j = X^j(\text{low}) + \left(X^j(up) - X^j(\text{low})\right) \cdot \text{ rand } (0,1) \tag{29}$$

where X^j(low) and X^j(up) are the lower bound and upper bound of the jth gene respectively; $i \in [1, k]$ and k is the population size.

Step2: Calculate the fitness function value of the individuals in the parent population and arrange in order to find out the positions of α, β and δ wolves.

Step3: According to the positions of three fittest wolves in the parent generation population, update the positions of individuals with equation (25).

Step4: Use equation (26) to get the mutant population of the parent population.

Step5: The child population can be obtained by crossing the parent population and mutant population in the form of equation (27).

Step 6: Make one–to–one selection in the parent population and child population through equation (28) to retain fitter individuals.

Step7: If the number of iterations is less than the given maximum, return Step2 to continue the iteration. Otherwise,

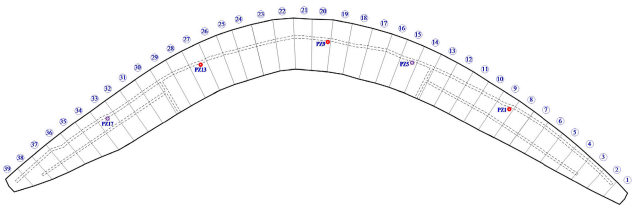


FIGURE 4. Layout of the selected piezometer tubes.

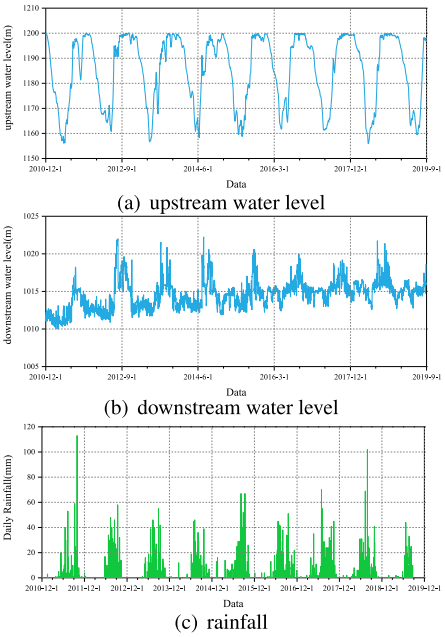


FIGURE 5. Collected data of environmental factors.

the position of α wolves in the current population is the optimal solution of the algorithm.

E. THE PROPOSED SEEPAGE PREDICTION AND IDENTIFICATION METHOD

According to the above–mentioned theories, a novel seepage behavior prediction model for concrete dams based on HGWO–XGBoost is proposed. In the light of this model, an effective identification method of lag process is proposed. Figure 3 shows the flowchart of model implementation.

III. CASE STUDY

A. ENGINEERING INTRODUCTION

A concrete double–curvature arch dam is located in Panzhihua City, Sichuan Province, China. The maximum dam height is 240m and the elevation of dam crest is 1205m. In order to effectively monitor the uplift pressure, 22 piezometric tubes are arranged along the dam foundation, numbered PZ01∼PZ22. The piezometric tube water level is measured by osmometers.

There are 5 piezometer tubes located in the first row behind the impervious curtain, among which PZ05 and PZ17 have been damaged. In this study, the remaining intact piezometer

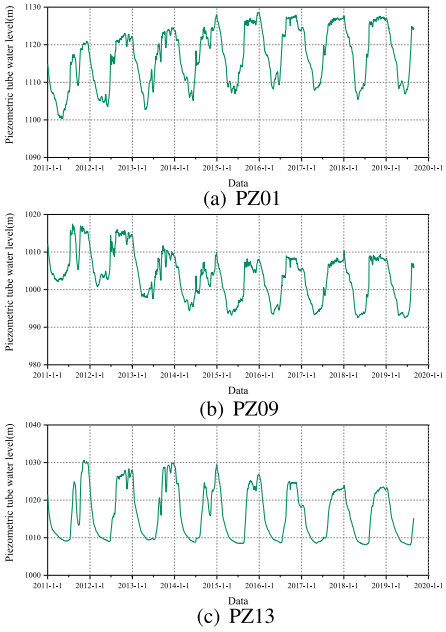


FIGURE 6. Process lines of piezometric tube water level.

tubes PZ01, PZ09 and PZ13 are selected. Figure 4 shows the layout of the piezometer tubes.

The complete monitoring sequence from 2011 to 2017 was taken as the training set, and the measured data from 2018 to August 2019 were taken as the testing set. The environmental variables and monitoring data are measured almost once a day, thus there are more than 3200 sets of sample data at each of the three piezometer tubes. The process lines of environmental factors and piezometric tube water level are shown in the figure 5 and figure 6 respectively.

B. EVALUATION INDICATORS

In order to understand the accuracy of the model more objectively, four evaluation indicators, including determination coefficient (R^2), mean absolute percent error ($MAPE$), root mean square error ($RMSE$) and mean absolute error (MAE), are used to evaluate the performance of the models in the training set and the testing set. The calculation formulas are as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n} \left(y_t - \frac{1}{n}\sum_{t=1}^{n} y_t\right)^2} \tag{30}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{\left|y_t - \hat{y}_t\right|}{\left|y_t\right|} \tag{31}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2} \tag{32}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \left|y_t - \hat{y}_t\right| \tag{33}$$

TABLE 1. Input factors considered for the XGBoost model.

Components	Input forms	Days before the monitoring day
Upstream water level	Original measured value	0,1,2,...,30
Downstream water level		0
Rainfall		0,1,2,...,30
Temperature	Simple harmonic	–
Aging	Polynomial	–

where y_t is the measured value and $\hat{y}_t$ is the output value of the model. The better model performance is, the larger R^2 or the smaller error index is.

C. PARAMETER OPTIMIZATION FOR XGBOOST

For machine learning algorithms, the selection of hyper–parameters has a crucial impact on the model performance. As well as adding a regularization term to the objective function and using the shrinkage method, XGboost also adopts the subsample method to avoid overfitting. In each iteration, a certain proportion of samples will be randomly selected for the learning of a single decision tree, and a certain proportion of features for the growth of a decision tree. This strategy makes full use of the existing data sets, but prevents strong correlation features from being applied overmuch to cause overfitting meanwhile, when processing high–dimensional data.

On the choice of lag factors, scholars usually adopt the previous segmental average values or equivalent values of the upstream water level and rainfall during the first 30 days before the monitoring day to consider the influence of these two kinds of factors on the dam seepage, and some good results have been obtained [2], [12], [19], [24].

In view of the above characteristics, in order to effectively distinguish the lag effect of upstream water level component and rainfall component, the original measured upstream water level and rainfall of the monitoring day and the first 30 days were selected as the model inputs in this paper. The forms of downstream water level component, temperature component and aging component were the same as equation (10). The factors considered for the XGBoost model are listed in table 1.

In XGboost, eight parameters need to be determined so as to control generation of decision trees and avoid overfitting. The meaning and search range of these parameters are shown in the table 2. GWO, HGWO, DE and GA were used to optimize the model respectively. The parameters of HGWO and DE are set to: population size $SearchAgents\_no = 10$; maximum number of iterations $Max\_iter = 200$; scaling factor $F_r = 0.5$; crossover probability $C_r = 0.2$. GWO has two parameters, $SearchAgents\_no$ and Max_iter, which are set the same as HGWO. In GA, $SearchAgents\_no$ and Max_iter are set the same as GWO. In addition, GA has two parameters: crossover fraction F_c and migration fraction F_m, which are set to 0.9 and 0.01 respectively.

In order to improve the generalization ability of the model, avoid overfitting and eliminate the influence of current data on algorithm performance as far as possible, we adopted five–fold cross validation during parameter optimization to evaluate the performance of the model under the certain parameter combination. For any piezometer tube, the training set was randomly divided into five parts with approximately the same size, four parts of which were taken as the training set and the remaining part as the testing set in turn. The average of the five results was taken as the final output of the model. The parameter optimization results are shown in table 3, where the best metric values of the models are marked in boldface.

Optimized by any of the four algorithms, the XGBoost model all shows high accuracy with little error. It is noticeable that under the condition of the same population size and iterations, HGWO has better global optimization ability than other algorithms. The performance of HGWO–XGBoost model is better than that of other models regardless of piezometer tubes and evaluation indicators, except having the same R^2 as some of the models. The other three algorithms perform nearly. By combining GWO with DE, HGWO outperforms either of them. Especially at PZ09, the *MAPE*, *RMSE* and *MAE* of HGWO–XGBoost model are less than 40% of those of GWO–XGBoost model.

D. MODELS PERFORMANCE COMPARISON

As can be seen from the previous analysis, compared with the other three algorithms, HGWO obviously has better global optimization capability. To demonstrate the feasibility of the proposed model, six state–of–art methods were introduced as baseline models. They include SVR, multilayer perceptron (MLP), extreme learning machine (ELM), multiple linear regression (MLR), random forest(RF) and gradient boosting decision tree (GBDT), which have been widely used in dam safety monitoring. Among them, RF and GBDT belong to the ensemble learning method based on decision tree as XGBoost.

So as to avoid ill–posed problems in traditional statistical models, equivalent values of upstream water level and rainfall were used to replace the influence of these two components. We utilized a trial–and–error method on the basis of HGWO to calculate the training set, and then the number of lag days and influence days were obtained. The selection and computation of input variables are the same as equation (10), and the computation results are shown in table 4.

For the remaining five baseline models based on machine learning, in order to objectively evaluate their ability to identify and predict the dam seepage behavior, the model inputs are the same as that of XGboost, meanwhile adopting HGWO to optimize hyper–parameters.

The key hyper–parameters of comparison models are shown in table 5. The performance of SVR models is mainly controlled by regularization parameter C and kernel coefficient *gamma*, the range of which is [0,1000] and [0.001,1] respectively. The key hyper–parameters of MLP are mainly including the number of neurons in the hidden layers (N_e) and learning rate (l_e), the range of which is [10,500] and [0.01,1]

TABLE 2. Explanation and search space of the control parameters in XGboost.

Parameters	Explanation	Search domain
eta	The learning rate, which can be reduced to improve the robustness of the model	[0.01,0.3]
min_child_weight	The minimum sum of sample weights in leaf nodes	[1,5]
max_depth	Maximum depth of a single regression tree	[3,10]
$subsample$	The random sampling rate of each tree	[0.5,1]
$colsample\_bytree$	The random sampling rate of the features per tree	[0.5,1]
$alpha$	The weight of the L1 regularization term	[1,5]
$lambda$	The weight of the L2 regularization term	[1,5]
$n\_estimators$	The number of regression trees, that is, the maximum number of iterations of the weak learner	[500,1000]

TABLE 3. The optimal parameters obtained by optimization algorithms.

Piezometer tubes	Algorithm	Parameters of XGBoost	R^2	$MAPE$	$RMSE$	MAE
PZ01	GWO	[6, 5, 0.5, 0.816, 1, 0.067, 761, 3.677]	**9.998E-01**	8.708E-05	1.319E-01	9.721E-02
	HGWO	[6, 3.708, 0.611, 0.732, 1, 0.063, 731, 4.523]	**9.998E-01**	**7.209E-05**	**1.104E-01**	**8.046E-02**
	DE	[5, 4.940, 0.541, 0.763, 1, 0.106, 923, 4.521]	9.997E-01	8.271E-05	1.257E-01	9.229E-02
	GA	[5, 4.521, 0.628, 0.975, 2, 0.079, 789, 4.696]	9.997E-01	8.082E-05	1.216E-01	9.020E-02
PZ09	GWO	[6, 4.384, 0.602, 0.605, 5, 0.191, 833, 3.132]	9.997E-01	7.550E-05	1.061E-01	7.515E-02
	HGWO	[9, 1.189, 0.636, 0.559, 4, 0.156, 795, 3.872]	**9.999E-01**	**2.659E-05**	**4.023E-02**	**2.645E-02**
	DE	[5, 1.198, 0.734, 0.514, 2, 0.151, 771, 4.742]	**9.999E-01**	4.021E-05	5.682E-02	4.037E-02
	GA	[9, 2.123, 0.776, 0.626, 4, 0.246, 851, 4.538]	**9.999E-01**	3.571E-05	5.441E-02	3.585E-02
PZ13	GWO	[6, 3.190, 0.602, 0.599, 4, 0.073, 622, 1.171]	9.998E-01	6.810E-05	9.613E-02	6.931E-02
	HGWO	[5, 1.705, 0.566, 0.5, 1, 0.072, 822, 1]	**9.999E-01**	**5.124E-05**	**6.994E-02**	**4.918E-02**
	DE	[4, 2.828, 0.732, 0.714, 2, 0.166, 954, 4.954]	**9.999E-01**	5.945E-05	8.437E-02	6.052E-02
	GA	[3, 3.612, 0.563, 0.798, 2, 0.237, 974, 2.296]	9.998E-01	7.543E-05	1.045E-01	7.679E-02

TABLE 4. The computation results of lag effect by trial–and–error method.

Piezometer tubes	Upstream water level		Rainfall	
	Lag days	Influence days	Lag days	Influence days
PZ01	0	1	4	7
PZ09	0	1	10	8
PZ13	1	9	14	11

TABLE 5. The hyper–parameters of comparison models.

Baseline models	Hyper-parameters	Domain	PZ01	PZ09	PZ13
SVR	C	[0,1000]	63	12	8
	$gamma$	[0.001,1]	0.0011	0.001	0.0015
MLP	N_e	[10,500]	97	165	10
	l_e	[0.01,1]	0.011	0.0038	0.051
ELM	L	[10,500]	373	324	293
RF	max_depth	[3,10]	9	8	8
	$n\_estimators$	[500,1000]	684	500	853
	l_m	[1,5]	1	1	2
	$colsample$	[0.5,1]	0.510	0.5	0.519
GBDT	max_depth	[3,10]	3	4	3
	$n\_estimators$	[500,1000]	730	687	777
	l_m	[1,5]	1	3	1
	$colsample$	[0.5,1]	0.711	0.5	0.557
	$subsample$	[0.5,1]	0.5	0.5	0.514
	eta	[0.01,0.3]	0.118	0.027	0.256

respectively. For ELM, there is only one control parameter needed to be optimized: L, the number of hidden nodes with the range of [10,500]. The RF model and the XGBoost model have four and six key hyper–parameters, respectively. Among them l_m is the minimum number of samples required to be at a leaf node, with the range of [1,5]. The other parameters have the same meaning as their counterparts in XGBoost (see table 2). The performance of the models on the training set is also evaluated by using five–fold cross validation during parameter optimization, and the computation results are shown in table 6 where the best metric values of the models are marked in boldface. The histograms of evaluation indexes are shown in figure 7.

We can see from the table 6 and figure 7 that in terms of the fitting results of the training sets, the models based on machine learning algorithms have excellent performance in accuracy. The fitting accuracy of the traditional statistical model is slightly lower at PZ09 and PZ13, but the error is acceptable, indicating that the input factors in section II-A can well summarize the causes of seepage behavior. In other words, the input factors selected in this study are reasonable. The XGBoost model and the GBDT model are the most outstanding overall, following by the RF model and the SVR model. The XGBoost model performs better at PZ09, while the GBDT model performs slightly better at PZ01 and PZ13.

As far as testing sets are concerned, models perform differently than they do on the training sets. The accuracy of the three ensemble learning models is particularly outstanding. It can be seen that the performance of the XGBoost model is the best among comparison models in all respects at all

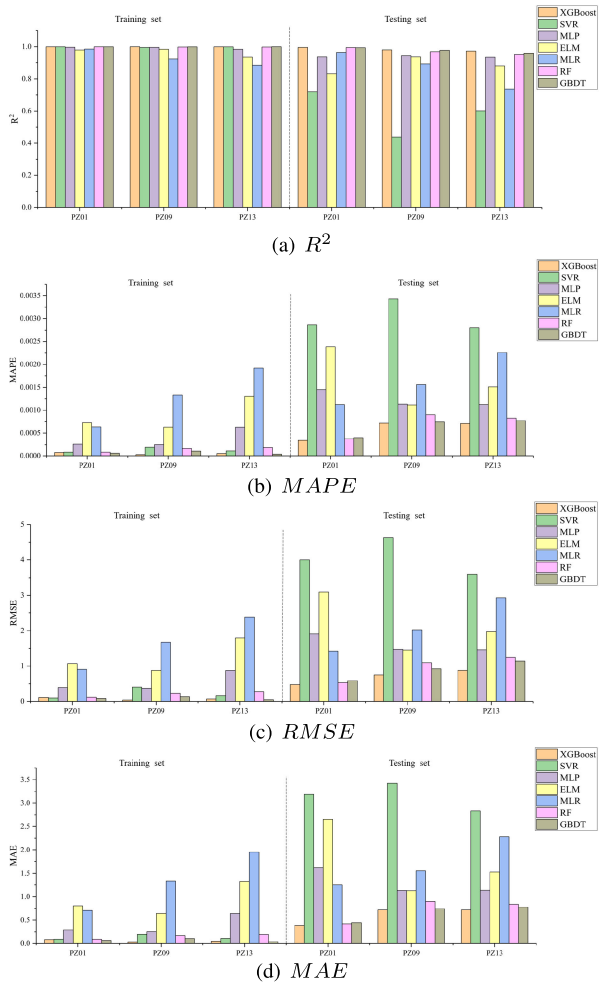


FIGURE 7. The histograms of evaluation indexes.

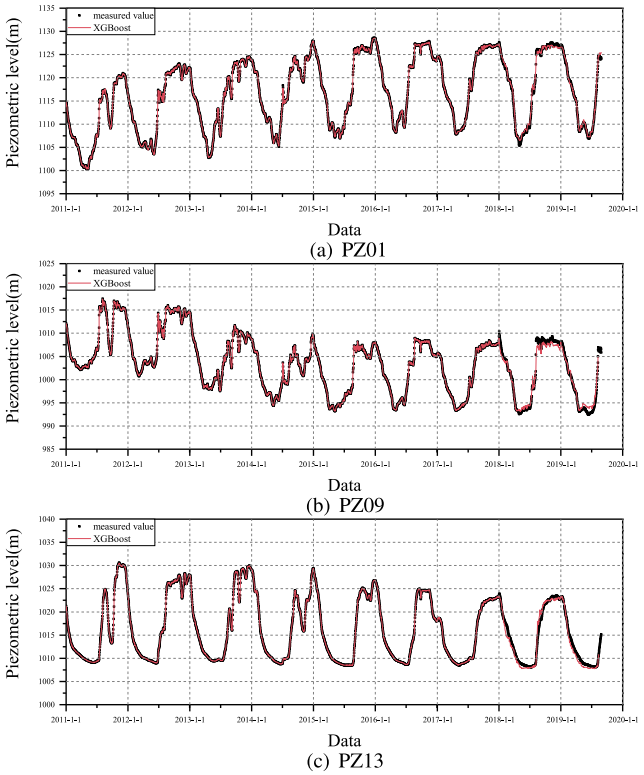


FIGURE 8. Performance of the best model.

piezometer tubes. In the other four models, the MLP model have the best and most stable performance with little prediction error at all tubes. The ELM model and MLR model follow, but perform poorly at some monitoring points.

It could not be neglected that the performance of the XGBoost model on the testing sets is slightly inferior to the SVR model on the training sets, while the prediction accuracy of the SVR model is greatly reduced compared to the fitting results. In addition, we can find that the prediction accuracy of the XGBoost model is slightly better than that of the GBDT model, which is contrary to the performance they show on the training sets. It indicates that the introduction of regularization term makes the XGBoost model avoid falling into over–fitting more effectively. The optimal model results are shown in figure 8.

The methods mentioned above can be divided into two major categories. One is single model, including SVR, MLP, ELM, and MLR. The other is the integration model with more advances, including XGBoost, RF, and GBDT. The basic strategy of these three ensemble learning algorithms is to integrate multiple decision trees to complete data mining work, through bagging or boosting method. Compared with single models, the combination strategy makes the integrated models have potential to explore the deeper nonlinear relationship between factors and response variables, as well as more stable performance on testing sets.

At the same time, the bigger size of the algorithm structure makes the integration models have more hyper-parameters. Especially in the XGBoost model, the introduction of subsampling and regularization term strategy makes it have the most hyper-parameters, but the comparison results also show that it has the optimal performance. Although the integration models generate more computational costs, it is negligible in terms of the scale of data in dam safety monitoring.

E. RELATIVE IMPORTANCE OF INPUT COMPONENTS

Considering that the HGWO-XGBoost model has the highest accuracy, we only focus on this model to interpret the piezometric tube water level. We computed the importance of all input factors according to the method introduced in section II-C. Then the importance of factors was summed up according to the components they belong to. We tested the validity of the proposed method by analyzing whether the relative importance of each component is consistent with common engineering perceptions.

Figure 9 shows the relative importance of the five input components at all three of monitoring points. It can be seen from the obtained results that the upstream water level component is the most important to uplift pressure, with more than 40% relative importance in the whole of three points,

TABLE 6. Performance assessment of different models.

Piezometer tubes	Models	Training set				Testing set			
		R^2	$MAPE$	$RMSE$	MAE	R^2	$MAPE$	$RMSE$	MAE
PZ01	XGBoost	9.998E-01	7.209E-05	1.104E-01	8.046E-02	**9.960E-01**	**3.451E-04**	**4.761E-01**	**3.857E-01**
	SVR	9.998E-01	8.062E-05	9.631E-02	9.006E-02	7.192E-01	2.863E-03	3.998E+00	3.193E+00
	MLP	9.973E-01	2.632E-04	3.917E-01	2.940E-01	9.358E-01	1.453E-03	1.912E+00	1.623E+00
	ELM	9.798E-01	7.240E-04	1.069E+00	8.084E-01	8.320E-01	2.384E-03	3.092E+00	2.657E+00
	MLR	9.854E-01	6.340E-04	9.066E-01	7.080E-01	9.644E-01	1.121E-03	1.424E+00	1.250E+00
	RF	9.998E-01	8.040E-05	1.184E-01	8.983E-02	9.950E-01	3.749E-04	5.314E-01	4.187E-01
	GBDT	**9.999E-01**	**5.773E-05**	**8.191E-02**	**6.445E-02**	9.939E-01	3.974E-04	5.874E-01	4.433E-01
PZ09	XGBoost	**1.000E+00**	**2.659E-05**	**4.023E-02**	**2.645E-02**	**9.803E-01**	**7.167E-04**	**7.505E-01**	**7.180E-01**
	SVR	9.955E-01	1.982E-04	4.049E-01	1.996E-01	4.386E-01	3.432E-03	4.629E+00	3.426E+00
	MLP	9.962E-01	2.524E-04	3.729E-01	2.539E-01	9.431E-01	1.132E-03	1.474E+00	1.133E+00
	ELM	9.843E-01	6.301E-04	8.752E-01	6.418E-01	9.359E-01	1.113E-03	1.454E+00	1.128E+00
	MLR	9.229E-01	1.328E-03	1.670E+00	1.334E+00	8.925E-01	1.562E-03	2.026E+00	1.560E+00
	RF	9.985E-01	1.723E-04	2.345E-01	1.729E-01	9.687E-01	9.030E-04	1.093E+00	9.028E-01
	GBDT	9.995E-01	1.023E-04	1.315E-01	1.027E-01	9.778E-01	7.415E-04	9.209E-01	7.404E-01
PZ13	XGBoost	9.999E-01	5.124E-05	6.994E-02	4.918E-02	**9.730E-01**	**7.085E-04**	**8.768E-01**	**7.185E-01**
	SVR	9.995E-01	1.056E-04	1.598E-01	1.076E-01	6.014E-01	2.802E-03	3.598E+00	2.837E+00
	MLP	9.844E-01	6.291E-04	8.743E-01	6.408E-01	9.343E-01	1.122E-03	1.461E+00	1.138E+00
	ELM	9.345E-01	1.302E-03	1.789E+00	1.326E+00	8.799E-01	1.512E-03	1.975E+00	1.534E+00
	MLR	8.840E-01	1.920E-03	2.382E+00	1.954E+00	7.355E-01	2.259E-03	2.931E+00	2.286E+00
	RF	9.984E-01	1.923E-04	2.781E-01	1.958E-01	9.522E-01	8.300E-04	1.246E+00	8.413E-01
	GBDT	**1.000E+00**	**3.606E-05**	**4.761E-02**	**3.668E-02**	9.598E-01	7.699E-04	1.142E+00	7.802E-01

which is 43.37%, 41.70% and 50.29% respectively. It is followed by the rainfall component of which the relative importance at the three points is 27.20%, 32.19% and 22.73%, respectively. The time component has a minor influence on seepage. The relative importance of the time component at PZ01 and PZ13 is close (10.71% and 11.06%, respectively), but at PZ09 which is slightly lower (6.93%). The relative importance of the temperature component is slightly greater than that of the time component. The downstream water level component has minimum influence on the uplift pressure with around 4% at each of the three points.

From the perspective of engineering experience, the magnitude of upstream water level plays a crucial role in dam seepage, following by the effect of groundwater that is closely related to rainfall. In this case study, upstream water level is 140∼180 m higher than downstream water level which has far less variation meanwhile. The correlation between the downstream water level component and the dam seepage should be far less in comparison with the upstream water level component. Therefore, the calculated results in figure 9 is accordance with practical engineering knowledge.

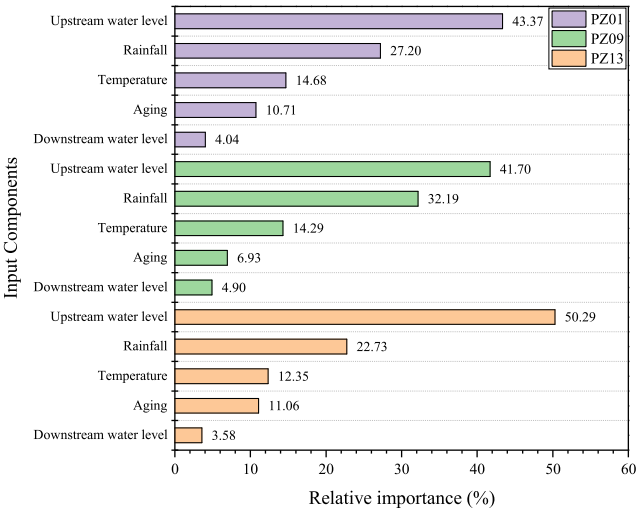


FIGURE 9. Relative importance of input components.

F. LAG PROCESS IDENTIFICATION

Whether at PZ01, PZ09 or PZ13, the upstream water level and rainfall have the greatest influence on uplift pressure. However, both of them show a strong time lag effect. In order to control the seepage behavior of the dam effectively, it is necessary to accurately evaluate the lag process of the upstream water level and rainfall.

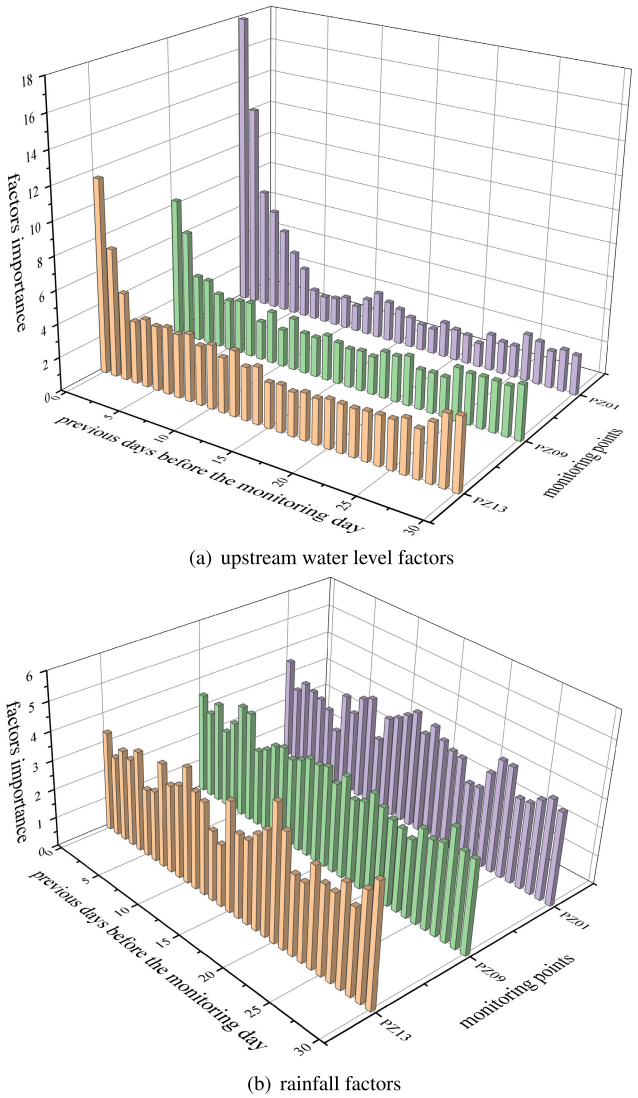
According to the computation results of factor importance obtained in section III-E, we analyzed the upstream water level factors and rainfall factors separately. Take the upstream water level component as an example. With the passage of time, the water level factor of the monitoring day will become previous water level factor, and the previous water level factors will turn into more previous ones simultaneously. Therefore, the importance distribution of water level factors to uplift pressure on the monitoring day and the first 30 days can be equivalently regarded as the lag process of water level on uplift pressure. The distribution of factors importance is shown in figure 10.

According to figure 10, it shows some consistent lag laws at the three piezometer tubes:

• The water level on the monitoring day has the most significant influence on uplift pressure. The effect of this water level doesn't disappear immediately, but gradually declined

TABLE 7. Validity comparison of lag process identification via different equivalent values.

Piezometer tubes	Models	Training set				Testing set			
		R^2	$MAPE$	$RMSE$	MAE	R^2	$MAPE$	$RMSE$	MAE
PZ01	Model I	9.8507E-01	6.4831E-04	9.1805E-01	7.2395E-01	**9.6653E-01**	**1.0801E-03**	**1.3804E+00**	**1.2045E+00**
	Model II	9.8506E-01	6.4793E-04	9.1846E-01	7.2352E-01	9.6592E-01	1.0933E-03	1.3929E+00	1.2193E+00
	Model III	9.8538E-01	6.3673E-04	9.0841E-01	7.1104E-01	9.6557E-01	1.0977E-03	1.4000E+00	1.2242E+00
	Model IV	**9.8544E-01**	**6.3399E-04**	**9.0662E-01**	**7.0800E-01**	9.6438E-01	1.1211E-03	1.4239E+00	1.2503E+00
PZ09	Model I	9.1782E-01	1.3711E-03	1.7237E+00	1.3773E+00	**9.0306E-01**	**1.5098E-03**	**1.9234E+00**	**1.5079E+00**
	Model II	9.1820E-01	1.3688E-03	1.7197E+00	1.3750E+00	8.9991E-01	1.5273E-03	1.9544E+00	1.5254E+00
	Model III	9.2189E-01	1.3329E-03	1.6805E+00	1.3389E+00	8.9761E-01	1.5307E-03	1.9767E+00	1.5289E+00
	Model IV	**9.2291E-01**	**1.3284E-03**	**1.6695E+00**	**1.3343E+00**	8.9248E-01	1.5617E-03	2.0257E+00	1.5599E+00
PZ13	Model I	**8.8450E-01**	1.9161E-03	**2.3762E+00**	1.9504E+00	**7.3717E-01**	**2.2443E-03**	**2.9219E+00**	**2.2713E+00**
	Model II	8.8449E-01	**1.9160E-03**	2.3763E+00	**1.9503E+00**	7.3633E-01	2.2474E-03	2.9265E+00	2.2743E+00
	Model III	8.8395E-01	1.9203E-03	2.3819E+00	1.9547E+00	7.3639E-01	2.2560E-03	2.9262E+00	2.2831E+00
	Model IV	8.8396E-01	1.9200E-03	2.3818E+00	1.9543E+00	7.3554E-01	2.2585E-03	2.9309E+00	2.2856E+00



(a) upstream water level factors

(b) rainfall factors

FIGURE 10. The lag process of factors at three piezometer tubes.

over time. In the first 4 days or so, the effect is relatively distinct and drops rapidly, and then slowly declines to be stable.

- The lag process of upstream water level on uplift pressure accords with the expectation that it is approximately normal distribution, but obviously it is not a strict normal distribution.
- As far as rainfall is concerned, the most interesting observation is that its lag process does not obey the normal distribution at all. The lag process of rainfall shows the characteristic of multi–peak and fluctuation, which is completely different from anticipation.

In this study, the trial–and–error method based on HGWO–MLR and factor importance computation method which was on the basis of HGWO–XGBoost were utilized respectively to identify the lag process of upstream water level and rainfall. Comparing the results obtained by these two methods, it can be drawn that there is some difference. In order to explore the rationality of the results, we used these two methods to construct the equivalent values of upstream water level and rainfall firstly.

It can be observed from figure 10 that there is a clear influence period on the upstream water level. The water level factors within the period have an obvious effect on uplift pressure of the monitoring day, whereas ones outside the period have a comparatively small and average influence. The influence period of upstream water level at PZ01 and PZ09 is roughly the first 7 days, whereas which at PZ13 is 14 days. In terms of the effect of rainfall on uplift pressure, the lag process is more complicated. Thus, the 30 days is taken as the influence period of rainfall.

In view of the relative importance computation result, the weight of factors during influence period were defined, of which the sum was 1. The factors in the influence period were weighted and summed to obtain the equivalent values (Y_{Hu1} and Y_{p1}). The equivalent values (Y_{Hu2} and Y_{p2}) by the trial–and–error method were obtained through equation (5) and equation (7) respectively in the case of table 4. Then stepwise regression method was utilized to establish traditional HPTT statistical model. Four models were built at each monitoring point, include model I ($Y_{Hu1} + Y_{p1}$), model II ($Y_{Hu1} + Y_{p2}$), model III ($Y_{Hu2} + Y_{p1}$) and

model IV ($Y_{Hu2}+Y_{p2}$). In those models, the equivalent values of water level and rainfall were matched in pairs, other input factors were the same as equation (10). The computation results are exhibited in the table 6, where the best metric values of the models are marked in boldface.

It can be observed from table 6 that model I and model II perform relatively well on the training set at PZ13. Besides, the training error of model IV is the smallest among models at PZ01 and PZ09, followed by model III, and model I has the lowest accuracy. The most interesting observation is that the rank order of model precision on the testing sets is almost the opposite of which on the training sets. At all of the three piezometer tubes, the model I has the highest prediction accuracy whereas the model IV has the lowest. The prediction accuracy of model III or model IV which contains one equivalent value obtained by the trial–and–error method and the other by the proposed method is somewhere in between. It indicates that compared with the trial–and–error method, the equivalent value of the upstream water level and rainfall obtained by the proposed method contains more information that is effective. In other words, the lag process obtained through equation (22) can better reflect the actual situation, although the lag effect of rainfall component completely fails to meet the expected assumption of normal distribution.

IV. CONCLUDING REMARKS AND FUTURE WORK

A. CONCLUSION

After the above comprehensive research, the main conclusions of this paper are as follows:

1. The DE method can effectively help GWO jump out of the local optimal solution. The HGWO has a stronger global search capability compared with the original GWO, DE and GA method.

2. After a comprehensive comparison with the other six state–of–art methods, results indicated the HGWO–XGBoost performed better in terms of stability, accuracy and resistance to overfitting for uplift pressure modeling.

3. The computation method of factor importance proposed by Breiman *et al*. can reasonably explain the relative importance of each component to uplift pressure. The obtained results show that the upstream water level component is the most important to uplift pressure, being followed by the rainfall component, whereas the downstream water level component matters least.

4. Based on the computation results of factor importance by the HGWO–XGBoost model, the lag process of upstream water level and rainfall can be availably identified. The comparison with the trial–and–error method shows that the result obtained by the proposed method is more in line with the actual situation.

B. FUTURE WORK

Taking three piezometer tubes of a concrete double–arch dam as an example, we verified the application value of the proposed method to predict uplift pressure of concrete dams and identify lag process of factors. In addition to uplift pressure, seepage behavior of concrete dams also includes leakage flow and dam–around seepage. However, the influencing factors of these seepage behaviors are almost the same, so the method proposed in this paper is also applicable to other seepage characteristics.

Based on data mining, we evaluated the lag process of upstream water level and rainfall on seepage of concrete dams. The first limitation is that the results is obtained by mathematical method on the basis of the monitoring data. Therefore, they should be verified from the perspective of numerical simulation and experiment in future studies. Moreover, the selection of upstream water level and rainfall factors in the previous one month in this paper is based on the experience of predecessors. For specific projects, it is probably necessary to conduct a certain discussion on the selection of previous factors.

REFERENCES

[1] Y. Li, T. Bao, X. Shu, Z. Chen, Z. Gao, and K. Zhang, "A hybrid model integrating principal component analysis, fuzzy C-means, and Gaussian process regression for dam deformation prediction," *Arabian J. Sci. Eng.*, pp. 1–14, Sep. 2020, doi: 10.1007/s13369-020-04923-7.

[2] C. Gu and Z. Wu, *Safety Monitoring of Dams and Dam Foundations-Theories & Methods and Their Application*. Nanjing, China: Hohai Univ. Press, 2006, pp. 64–70.

[3] C. Gu, L. Hu, and Q. Zhang, "An analytic model for base flow of dam seepage," *Rock Soil Mech.*, vol. 26, no. 7, pp. 1033–1037, 2005, doi: 10.16285/j.rsm.2005.07.006.

[4] A. De Sortis and P. Paoliani, "Statistical analysis and structural identification in concrete dam monitoring," *Eng. Struct.*, vol. 29, no. 1, pp. 110–120, Jan. 2007, doi: 10.1016/j.engstruct.2006.04.022.

[5] Q. Zhang, Z. Wu, and C. Gu, "Seepage flow monitoring model for rockfill-earth dams based on lag effect," *J. Hydraulic Eng.*, vol. 1, no. 2, pp. 85–89, 2001, doi: 10.3321/j.issn:0559-9350.2001.02.016.

[6] H. Huang and B. Chen, "Dam seepage monitoring model based on dynamic effect weight of reservoir water level," *Energy Procedia*, vol. 16, pp. 159–165, Jan. 2012, doi: 10.1016/j.egypro.2012.01.027.

[7] B. Wei, M. Gu, H. Li, W. Xiong, and Z. Xu, "Modeling method for predicting seepage of RCC dams considering time-varying and lag effect," *Struct. Control Health*, vol. 25, no. e20812, pp. 1–14, 2018, doi: 10.1002/stc.2081.

[8] X. Cheng, Q. Li, Z. Zhou, Z. Luo, M. Liu, and L. Liu, "Research on a seepage monitoring model of a high core rockfill dam based on machine learning," *Sensors*, vol. 18, no. 27499, pp. 1–14, 2018, doi: 10.3390/s18092749.

[9] J. Yang, D. Hu, and Z. Wu, "Multiple co-linearity and uncertainty of factors in dam safety monitoring model," *J. Hydraulic Eng.*, vol. 35, no. 12, pp. 99–105, 2004, doi: 10.13243/j.cnki.slxb.2004.12.016.

[10] D. Li, Q. Qiangqiang, Z. Jun, and W. Jianye, "A comparative study on the processing methods of multicollinearity in dam monitoring data," *Urban Geotechnical Invest. Surveying*, no. 6, pp. 139–142, 2017, doi: CNKI:SUN:CSKC.0.2017-06-036.

[11] C. Gu, Y. Wang, Y. Peng, and B. Xu, "Ill-conditioned problems of dam safety monitoring models and their processing methods," *Sci. China Technol. Sci.*, vol. 54, no. 12, pp. 3275–3280, Dec. 2011, doi: 10.1007/s11431-011-4573-z.

[12] Z. Wu, *Safety Monitoring Theory & Its Application of Hydraulic Structures*. Nanjing, China: Hohai Univ. Press, 1990, pp. 130–142.

[13] C. Lin, T. Li, S. Chen, X. Liu, C. Lin, and S. Liang, "Gaussian process regression-based forecasting model of dam deformation," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8503–8518, Dec. 2019, doi: 10.1007/s00521-019-04375-7.

[14] F. Kang, J. Liu, J. Li, and S. Li, "Concrete dam deformation prediction model for health monitoring based on extreme learning machine," *Struct. Control Health Monitor.*, vol. 24, no. 10, p. e1997, Oct. 2017, doi: 10.1002/stc.1997.
[15] S. Chen, C. Gu, C. Lin, K. Zhang, and Y. Zhu, "Multi-kernel optimized relevance vector machine for probabilistic prediction of concrete dam displacement," *Eng. Comput.*, pp. 1–17, Jan. 2020, doi: 10.1007/s00366-019-00924-9.
[16] V. Ranković, A. Novaković, N. Grujović, D. Divac, and N. Milivojević, "Predicting piezometric water level in dams via artificial neural networks," *Neural Comput. Appl.*, vol. 24, no. 5, pp. 1115–1121, Apr. 2014, doi: 10.1007/s00521-012-1334-2.
[17] Y. Li, T. Bao, J. Gong, X. Shu, and K. Zhang, "The prediction of dam displacement time series using STL, extra-trees, and stacked LSTM neural network," *IEEE Access*, vol. 8, pp. 94440–94452, 2020, doi: 10.1109/ACCESS.2020.2995592.
[18] E. Sharghi, V. Nourani, and N. Behfar, "Earthfill dam seepage analysis using ensemble artificial intelligence based modeling," *J. Hydroinformatics*, vol. 20, no. 5, pp. 1071–1084, Sep. 2018, doi: 10.2166/hydro.2018.151.
[19] Z. Shi, C. Gu, E. Zhao, and B. Xu, "A novel seepage safety monitoring model of CFRD with slab cracks using monitoring data," *Math. Problems Eng.*, vol. 2020, pp. 1–13, May 2020, doi: 10.1155/2020/1641747.
[20] J. Hu and F. Ma, "Zoned safety monitoring model for uplift pressures of concrete dams," *Trans. Inst. Meas. Control*, vol. 41, no. 14, pp. 3952–3969, Oct. 2019, doi: 10.1177/0142331219842281.
[21] V. Nouran, E. Sharghi, and M. H. Aminfar, "Integrated ANN model for earthfill dams seepage analysis: Sattarkhan Dam in Iran," *Artif. Intell. Res.*, vol. 1, no. 2, pp. 22–37, 2012, doi: 10.5430/air.v1n2p22.
[22] S. Chen, C. Gu, C. Lin, Y. Wang, and M. A. Hariri-Ardebili, "Prediction, monitoring, and interpretation of dam leakage flow via adaptative kernel extreme learning machine," *Measurement*, vol. 166, Dec. 2020, Art. no. 108161, doi: 10.1016/j.measurement.2020.108161.
[23] F. Salazar, M. Á. Toledo, E. Oñate, and B. Suárez, "Interpretation of dam deformation and leakage with boosted regression trees," *Eng. Struct.*, vol. 119, pp. 230–251, Jul. 2016, doi: 10.1016/j.engstruct.2016.04.012.
[24] S.-W. Wang, Y.-L. Xu, C.-S. Gu, and T.-F. Bao, "Monitoring models for base flow effect and daily variation of dam seepage elements considering time lag effect," *Water Sci. Eng.*, vol. 11, no. 4, pp. 344–354, Oct. 2018, doi: 10.1016/j.wse.2018.12.004.
[25] J. Duan, P. G. Asteris, H. Nguyen, X.-N. Bui, and H. Moayedi, "A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ICA-XGBoost model," *Eng. Comput.*, pp. 1–8, Mar. 2020, doi: 10.1007/s00366-020-01003-0.
[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, San Francisco, CA, USA, 2016, pp. 785–794.
[27] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-xgboost model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.
[28] S.-E. Ryu, D.-H. Shin, and K. Chung, "Prediction model of dementia risk based on XGBoost using derived variable extraction and hyper parameter optimization," *IEEE Access*, vol. 8, pp. 177708–177720, 2020, doi: 10.1109/ACCESS.2020.3025553.
[29] J. Cheng, G. Li, and X. Chen, "Research on travel time prediction model of freeway based on gradient boosting decision tree," *IEEE Access*, vol. 7, pp. 7466–7480, 2019, doi: 10.1109/ACCESS.2018.2886549.
[30] B. Dai, C. Gu, E. Zhao, and X. Qin, "Statistical model optimized random forest regression model for concrete dam deformation monitoring," *Struct. Control Health Monitor.*, vol. 25, no. 6, p. e2170, Jun. 2018, doi: 10.1002/stc.2170.
[31] Y. Xiang, S.-Y. Fu, K. Zhu, H. Yuan, and Z.-Y. Fang, "Seepage safety monitoring model for an Earth rock dam under influence of high-impact typhoons based on particle swarm optimization algorithm," *Water Sci. Eng.*, vol. 10, no. 1, pp. 70–77, Jan. 2017, doi: 10.1016/j.wse.2017.03.005.
[32] K. Zhu, C. Gu, J. Qiu, and H. Li, "The analysis of the concrete gravity Dam's foundation uplift pressure under the function of typhoon," *Math. Problems Eng.*, vol. 2016, pp. 1–9, Jan. 2016, doi: 10.1155/2016/2834192.
[33] H. Li, *Statistical Learning Methods*. Beijing, China: Tsinghua Univ. Press, 2012, pp. 137–153.
[34] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.2307/2699986.
[35] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, Jul. 2017, doi: 10.1016/j.eswa.2017.02.017.
[36] L. Breiman, J. H. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.
[37] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.
[38] S. Mohanty, B. Subudhi, and P. K. Ray, "A new MPPT design using grey wolf optimization technique for photovoltaic system under partial shading conditions," *IEEE Trans. Sustain. Energy*, vol. 7, no. 1, pp. 181–188, Jan. 2016, doi: 10.1109/TSTE.2015.2482120.
[39] K. Balasubramanian and N. P. Ananthamoorthy, "Improved adaptive neuro-fuzzy inference system based on modified glowworm swarm and differential evolution optimization algorithm for medical diagnosis," *Neural Comput. Appl.*, Nov. 2020, doi: 10.1007/s00521-020-05507-0.
[40] C. E. D. S. Santos, R. C. Sampaio, L. D. S. Coelho, G. A. Bestard, and C. H. Llanos, "Multi-objective adaptive differential evolution for SVM/SVR hyperparameters selection," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107649, doi: 10.1016/j.patcog.2020.107649.
[41] A. Zhu, C. Xu, Z. Li, J. Wu, and Z. Liu, "Hybridizing grey wolf optimization with differential evolution for global optimization and test scheduling for 3D stacked SoC," *J. Syst. Eng. Electron.*, vol. 26, no. 2, pp. 317–328, Apr. 2015, doi: 10.1109/JSEE.2015.00037.

KANG ZHANG received the B.S. degree in water resources and hydropower engineering from the Hefei University of Technology, Xuancheng, China, in 2014. He is currently pursuing the Ph.D. degree in hydraulic structural engineering with Hohai University. His research interests include dam safety monitoring theory and dam safety status assessment.

CHONGSHI GU was born in Qidong, Jiangsu, China, in 1962. He received the B.Eng. and Ph.D. degrees in water conservancy and hydropower engineering from Hohai University, Nanjing, China, in 1985 and 1997, respectively. Since 2004, he has been a Professor with the Water Conservancy and Hydropower Engineering College, Hohai University. From 2010 to 2020, he was the Dean of the College. He is the author of six books and more than 150 articles. His research interests include early diagnosis for hydraulic structures, dam safety evaluation, and structural mechanics. He was a recipient of the Award of Scientific and Technological Progress of the State, in 2005, 2007, and 2015, respectively. He is currently the Chief Editor of *Advances in Science and Technology of Water Resources*.

YANTAO ZHU received the B.E. degree in water conservancy and hydropower engineering, from Hohai University, Nanjing, China, where he is currently pursuing the Ph.D. degree with the College of Water Conservancy and Hydropower Engineering. He is a Visiting Scholar with the University of California at Los Angeles, Los Angeles. His research interests include risk assessment of dam and structural health monitoring (SHM).

SIYU CHEN received the B.E. degree from the College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, China, where he is currently pursuing the Ph.D. degree. From 2019 to 2020, he was a Visiting Scholar with the University of Colorado Boulder. His research interests include dam health monitoring, applied machine learning, and uncertainty quantification.

BO DAI received the B.E. degree from the College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, China, where he is currently pursuing the Ph.D. degree. He is a Visiting Scholar with the National University of Singapore. His research interests include machine learning, structural health monitoring, and hydraulic engineering.

YANGTAO LI received the B.S. degree in water resources and hydropower engineering from Shihezi University, Xinjiang, China, in 2014. He is currently pursuing the Ph.D. degree in hydraulic structural engineering with Hohai University. His research interests include data mining and the application of machine learning technology in the field of dam safety monitoring.

XIAOSONG SHU received the B.S. degree in water resources and hydropower engineering from Hohai University, Nanjing, China, in 2014, where he is currently pursuing the Ph.D. degree in hydraulic structural engineering. His research interests include safety assessment and stability analysis of dam structure.