A Combination Model for Displacement Interval Prediction of Concrete Dams Based on Residual Estimation

Xin Yang 1, Yan Xiang 1,2,*, Guangze Shen 1,2 and Meng Sun 3

1 Nanjing Hydraulic Research Institute, Nanjing 210029, China
2 State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering, Nanjing 210098, China
3 Jiangsu Estuary Waterway for Huaihe River Project Management Office, Huai’an 223000, China
* Correspondence: yxiang@nhri.cn

Abstract: Accurate prediction and reasonable warning for dam displacement are important contents of dam safety monitoring. However, it is difficult to identify abnormal displacement based on deterministic point prediction results. In response, this paper proposes a model that integrates several strategies to achieve high-precision point prediction and interval prediction of dam displacement. Specifically, the interval prediction of dam displacement is realized in three stages. In the first stage, a displacement prediction model based on Extreme gradient boosting (XGBoost) is constructed. In the second stage, the prediction error sequence of XGBoost model is generated by the residual estimation method proposed in this paper, and the residual prediction model based on artificial neural network (ANN) is constructed through the maximum likelihood estimation method. In the third stage, the interval estimation of the noise sequence composed of the training error of the ANN model is carried out. Finally, the results obtained above are combined to realize the interval prediction of the dam displacement. The performance of the proposed model is verified by the monitoring data of an actual concrete dam. The results show that the hybrid model can not only achieve better point prediction accuracy than the single model, but also provide high quality interval prediction results.

Keywords: displacement prediction model; interval prediction; residual estimation; Extreme gradient boosting; artificial neural network

1. Introduction

As an important engineering measure to impound water flow, once a dam fails, uncontrolled floods will bring devastating disasters to the downstream area [1–3]. The safety of dams has always been a concern for engineers [4,5]. In order to achieve the most reliable operating state, a large number of sensors are installed to monitor structural response in real time, including fracture aperture, displacement, uplift pressure, and so on [6]. Among these, displacement is the most intuitive and reliable manifestation of the safety state of concrete dams [7,8]. Therefore, it is necessary to establish an effective displacement prediction model based on the prototype observation data and detect abnormal deformation to prevent the occurrence of hidden dangers and accidents.

Some scholars have recognized this and carried out research on the displacement prediction model of concrete dams. In general, this can be divided into three categories: the deterministic model, hybrid model and statistical model [9–11]. The deterministic model describes and predicts the deformation behavior of the dam through numerical simulation. It is mostly used in the design, construction and initial impoundment stages of dams that lack sufficient long-term monitoring data. However, the establishment of numerical models is cumbersome and time-consuming [12,13]. Many mechanical parameters involved in the model are difficult to be accurately determined [14,15]. In addition, the assumptions of boundary and geometric conditions also limit the prediction accuracy.
of deterministic models [16,17]. The hybrid model is a combination of the other two, which uses numerical methods to calculate the hydraulic component which is easy to simulate, while using statistical models to construct more complex temperature and aging components [18,19].

Compared with the previous two kinds of model, it is more common to construct a displacement prediction model for concrete dams based on a data-driven statistical model, usually including the hydrostatic-season-time (HST) model [20] and the hydrostatic-temperature-time (HTT) model [21]. Early practice is to assume that the factors are independent of each other, using least squares or partial least squares to solve the coefficients of the factors. The statistical model is favored by scholars and engineers for its simple structure, clear physical meaning and easy operation. The displacement prediction model based on a statistical regression method represents the dam displacement as a linear combination of a series of factors. However, the dam is a complex and nonlinear dynamic system. With the development of artificial intelligence algorithms, machine learning methods characterized by strong nonlinear mapping ability are gradually replacing statistical regression methods in developing displacement prediction models for concrete dams [4,6]. These modeling approaches include support vector machine (SVM) [22–24], random forests (RF) [25,26], extreme learning machine (ELM) [27,28], adaptive network-based fuzzy inference system [29,30] and Gaussian process regression [31].

It can be concluded from the literature survey that scholars are mostly limited to improving the prediction accuracy in the study of concrete dam displacement models. These high-precision models can give specific predicted values for dam displacement, but it is difficult to intuitively distinguish anomalies from a comparison with measured values. It should be noted that any work done with the prototype observation data is to analyze the dam behavior to ensure its safe operation [32]. Therefore, a high-precision displacement prediction model should not be the end of the task, but rather a powerful tool to achieve it. Compared with the simple point prediction model, the interval prediction method will be more practicable [33,34].

A common practice is to determine the upper and lower bounds of dam displacement based on the confidence interval method [35]. To be specific, it is assumed that the residual of the prediction model satisfies the normal distribution with mean 0 and variance $\sigma$, and the confidence interval of the residual is determined according to the significance level. Then, the confidence interval of the dam displacement is obtained by combining the predicted value with the confidence interval of the residual. Thus, when the concrete dam encounters a certain load combination, measured values are compared with the predicted interval to verify that they are within a predefined range. The measured values beyond the interval have low confidence, indicating that abnormal deformation of the dam may occur, and early warning should immediately be given to the site personnel.

However, there are some drawbacks to this method. The mean of the residuals of the displacement prediction model is often not 0 as we expected and the deviation may not be negligible. If the residual mean is reflected in the prediction interval, it will be contrary to the original intention of the prediction model. Especially when the mean of the residual is large, applying a bias to the predicted result may cause a large error in some originally accurate point predictions, even causing a false warning.

Previous studies have shown that, although the selection of input factors in the statistical model is reasonable, and various machine learning methods are efficient in mining the nonlinear mapping relationship between factors and dam displacement, there is still a small amount of information in the model residuals that can be further mined [36,37]. At the same time, the combined forecasting model with multiple technologies has been proved to have stronger information mining ability [38,39].

Therefore, based on the above considerations, this paper proposes a displacement interval prediction method for concrete dams combining XGBoost strategy and ANN algorithm. Firstly, the possible prediction errors of the XGBoost model are estimated on the training set based on the residual estimation method proposed in this paper. Then an
ANN model is established to fit the above residual sequence. Finally, the XGBoost-based displacement prediction model and the ANN-based residual prediction model are combined to form the hybrid XGBoost-ANN model to realize the prediction of dam displacement. Since the ANN model is based on the maximum likelihood estimation (MLE) method to train the data, its training error can maximally meet the normal distribution with the mean of 0 as the noise sequence. In this way, high-precision point prediction for displacement of concrete dams can be realized while interval prediction of dam displacement is carried out, providing a reliable early warning standard for field personnel.

The remainder of this paper is organized as follows. Several approaches covered in this paper are illustrated in Section 2, including the fundamentals of XGBoost and ANN, as well as the details of the proposed residual estimation method. In Section 3, the feasibility of the proposed method is demonstrated based on the prototype monitoring data of a concrete arch dam. Finally, a summary and suggestions for further work are given in Section 4.

2. Methodology

2.1. Extreme Gradient Boosting (XGBoost)

XGBoost is an efficient, portable and flexible parallel tree boosting system, which has received wide attention since it was proposed by Chen and Guestrin in 2016 [40]. It has been successfully used in many areas [41–43]. XGBoost is an ensemble learning algorithm based on boosting strategy. The XGBoost algorithm consists of a series of base learners (usually decision trees) with weak predictive ability. The residual of the previous weak learner is used as the training data of the next weak learner. Each weak learner is established to reduce the model residual in the gradient direction. Finally, all weak learners are combined to obtain a strong learner. The XGBoost algorithm can be considered as an improvement and extension of the gradient boosting decision tree (GBDT) algorithm. The fundamentals of GBDT can be found in the literature [44,45].

XGBoost can be represented as an additive model that optimizes the current state to enhance the model at each iteration. For a given set of data samples \(D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}(x_i \in R^m, y_i \in R)\), where \(x_i\) and \(y_i\) denote input factors and response variables, respectively. At step \(t\), the model can be expressed as:

\[
F_t(x_i) = F_{t-1}(x_i) + f_t(x_i),
\]

where \(F_{t-1}(x_i)\) is the model at step \(t-1\) and \(f_t(x_i)\) is the sub-model, which is a single decision tree in this study, at the current step. The optimization objectives of XGBoost algorithm are:

\[
Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{j=1}^{t} \Omega(f_t)
\]

\[
= \sum_{i=1}^{n} l(\hat{y}_i^{(t)} + f_i(x_i)) + \sum_{j=1}^{t} \Omega(f_t) + \text{constant},
\]

where \(\hat{y}_i\) is the estimate value of the \(i\)th response variable by the model, \(l\) represents the loss function and \(\Omega\) denotes the structural risk term representing the complexity of the model. The introduction of \(\Omega\) can effectively prevent overfitting. The more complex the construction of the decision tree is, the larger the value of \(\Omega(f)\) is. It is defined as follows:

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2,
\]

where \(\gamma\) and \(\lambda\) are hyperparameters, \(T\) is the number of leaf nodes in decision tree \(f\), and \(\omega\) is the vector composed of output values of all leaf nodes.

Different from GBDT, which only uses the first-order derivative of the loss function in model training, XGBoost carries out the second-order Taylor expansion of the loss function:

\[
\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) = \sum_{i=1}^{n} \left( l(y_i, F_{t-1}(x_i)) + \frac{\partial l}{\partial F_{t-1}(x_i)} f_t(x_i) + \frac{1}{2} \frac{\partial^2 l}{\partial^2 F_{t-1}(x_i)} f_t^2(x_i) \right).
\]
Since the former \( t - 1 \) sub-models have been determined, all the other parts in the formula are constant except the part related to \( f_t(x) \), which does not affect the optimization solution of \( f_t(x) \). Then the objective function can be simplified as:

\[
Obj^{(t)} = \sum_{i=1}^{n} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t),
\]

(5)

where \( g_i = \frac{\partial}{\partial F_{t-1}(x_i)} \) and \( h_i = \frac{\partial^2}{\partial^2 F_{t-1}(x_i)} \). With \( F_{t-1}(x) \) determined, \( g_i \) and \( h_i \) can be easily computed. Define the sample set on the \( j \)th leaf node as \( I(j) = \{x_i | q(x_i) = j\} \), where \( q(x_i) \) is the index function that maps samples to leaf nodes. Given that the regression value on the \( j \)th leaf node is \( \omega_j \), where \( \omega_j = f_t(x_i) \) and \( i \in I(j) \), then the Formula (5) can be transformed into:

\[
Obj^{(t)} = \sum_{j=1}^{J} \left[ \left( \sum_{i \in I(j)} g_i \right) \omega_j + \frac{1}{2} \left( \sum_{i \in I(j)} h_i + \lambda \right) \omega_j^2 \right] + \gamma T.
\]

(6)

Given \( \sum_{i \in I(j)} g_i = G_j \) and \( \sum_{i \in I(j)} h_i = H_j \), the Formula (6) can be further simplified as:

\[
Obj^{(t)} = \sum_{j=1}^{J} \left[ G_j \omega_j + \frac{1}{2} \left( H_j + \lambda \right) \omega_j^2 \right] + \gamma T,
\]

(7)

At each iteration, a new decision tree is generated in the direction that minimizes the Formula (7) to update the model, so that the loss function can gradually approach the minimum value. Compared with the first derivative, the second derivative helps the gradient to descend more quickly and accurately. When used, select the loss function on demand, and XGBoost can be used for both classification and regression.

2.2. Artificial Neural Network (ANN)

The ANN model, which is composed of neurons that simulate the basic unit of information storage and processing in the human brain, shows intelligent behaviors such as self-learning and self-organization [46,47]. The computational structure and learning rules of ANN follow the design of biological neural networks. The process by which nerve cells receive stimulation from surrounding cells and generate corresponding output signals can be simulated in a model using linear weighting and function mapping, while the adjustment process of network structure and weight is implemented using an optimization learning algorithm [48]. Since it was proposed in the 1940s, it has received extensive attention from scholars.

In this paper, the multi-layer perceptron (MLP) algorithm, which is commonly used in the field of neural networks, is used to realize the information mining of the residual sequence. Its topology diagram is shown in Figure 1. The MLP model should contain at least one hidden layer in addition to the input layer and the output layer. The output signal of neurons in any hidden layer or output layer can be obtained by:

\[
y_l^b = f \left( \sum_a w_{l-1}^{ab} y_{l-1}^a + \theta_l^b \right),
\]

(8)

where \( w_{l-1}^{ab} \) represents the weight vector between the \( a \)th neuron in the previous layer and the \( b \)th neuron in the \( l \)th layer, \( y_{l-1}^a \) is the output of the \( a \)th neuron in layer \( l - 1 \), \( \theta_l^b \) represents the bias vector, and \( f \) represents the activation function. The weight vector and bias vector are usually given an initial value and then the gradient descent method is used to reverse adjust the value of which according to the cost function.
After further data mining on the residuals of the XGBoost model, a new residual sequence \( r(x_i) \) with variance \( \sigma_r^2 \) is formed. Theoretically, the new residual should be mainly derived from data noise, which satisfies the normal distribution with mean zero. Therefore, the occurrence probability of residual \( r(x_i) \) can be expressed as:

\[
P(r(x_i)) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp \left( -\frac{r^2(x_i)}{2\sigma_r^2} \right).
\] (9)

In order to meet the above conditions as much as possible, we use MLE to train the model. The log form of the likelihood function can be expressed as:

\[
L(D_r) = \sum_{i=1}^{N} \ln \left[ \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp \left( -\frac{r^2(x_i)}{2\sigma_r^2} \right) \right] = -\frac{1}{2} \sum_{i=1}^{N} \left( \ln(2\pi\sigma_r^2) + \frac{r^2(x_i)}{\sigma_r^2} \right).
\] (10)

The larger the value of \( L(D_r) \), the closer the sequence \( D_r \) is to a normal distribution with mean zero. The training goal of ANN is to minimize the cost function, thus the cost function in this study is defined as:

\[
C = \sum_{i=1}^{N} \left( \ln(2\pi\sigma_r^2) + \frac{r^2(x_i)}{\sigma_r^2} \right)
\] (11)

### 2.3. Residual Estimate

Displacement monitoring equipment generally records values once a day. The number of samples in the training set is generally not more than 10,000 when building displacement prediction models. The order of magnitude of the training set is usually less than the complexity of the machine learning algorithm. Therefore, even if a series of measures such as sub-sampling and setting structural risk terms are taken, the over-fitting phenomenon cannot be avoided when constructing displacement prediction models. From the perspective of model performance, the accuracy of machine learning algorithms on training sets is sometimes significantly better than that on testing sets, while this phenomenon is not obvious in linear regression models with low model complexity. Because the residual distribution difference between training set and testing set is large, the application of the interval prediction method based on the confidence interval method in prediction models established by machine learning algorithm is limited.

In order to accurately estimate the prediction error of the model to ensure that the interval prediction is appropriate and meaningful, we propose a residual estimation method referring to the K-fold cross validation method. The residual estimation process is shown in Figure 2 and the specific implementation process is as follows:
Step 1: Divide the training set into k mutually exclusive sample subsets by the same time interval.

Step 2: The k subsets are successively traversed. The current subset is used as the validation set, and the remaining K-1 subsets are used as the training set to establish the prediction model based on XGBoost algorithm.

Step 3: The output results of the XGBoost model on total K validation sets are integrated as the regression values of the model on the training set, and then the residual sequence corresponding to the regression values is obtained.

3. Case Study

3.1. Project Overview and Data Introduction

This study takes a concrete double-curvature arch dam in southwest China as an example. The dam is located in the main stream of the Yalong River at the junction of Muli County and Yanyuan County in Liangshan Yi Autonomous Prefecture, Sichuan Province. The aerial view, layout plan and location information of the dam are shown in Figure 3. The basin area above the dam site covers 103,000 square kilometers, with a total storage capacity of 7.76 billion cubic meters. The normal water level of the reservoir is 1880 m and the dead water level is 1800 m. The maximum dam height is 305 m. A number of plumb lines are laid at No. 5, No. 9, No. 11, No. 13, No. 16 and No. 19 dam sections to measure the radial displacement of the dam, as shown in Figure 4, where PL represents the plumb line system and IP represents the inverse plumb line system.

![Figure 3. Case-study dam project: (a) aerial view and (b) layout chart.](image-url)
The process lines of PL13–1–PL13–5 are shown in Figure 5. It can be observed that the dam displacement at the five monitoring points on the same plumb line has similar variation rules but different amplitude of variation. With the decrease of the elevation of monitoring points, the variation range of dam displacement at monitoring points decreases gradually. The monitoring point PL13–1 located at the crest of the middle dam section has almost the largest annual variation among all the monitoring points. Therefore, we selected PL13–1 as the study object to establish a prediction model.

The dataset extracted from the monitoring database for this study spans August 2014 to August 2018 and contained 1163 sets of data. The monitoring data from August 2014 to December 2017 which contained 921 sets of samples was used to train the model, and the monitoring data from January 2018 to August 2018 which contained 242 sets of samples was taken as the testing set. The measured process line of the monitoring data is drawn in Figure 6, in which the positive values represent the radial displacement towards downstream. The process line of the upstream water level in the corresponding period is also drawn in Figure 6. It can be seen that the radial displacement of the dam is closely related to the reservoir water level.
Figure 6. Prototype monitoring data recorded by the automated monitoring system: (a) radial displacement at PL13–1; (b) recorded upstream water level.

3.2. The Selection of Input Factors

The displacement of a concrete dam is regarded as a comprehensive reflection of hydraulic effect ($\delta_H$), temperature effect ($\delta_T$) and aging effect ($\delta_\theta$). For $\delta_H$, taking the horizontal displacement along the river as an example, the displacement of any point of the dam body occurs mainly because of the following three reasons: (1) displacement caused by direct action of hydraulic load on the dam body (see Figure 7a), (2) displacement caused by foundation deformation due to hydraulic load (see Figure 7b) and (3) displacement caused by the rotation of foundation surface due to water weight (see Figure 7c). According to previous research, it is suggested that $\delta_H$ is written as a quartic polynomial of water depth $H$ for arch dams [49].

Thermal expansion and contraction of concrete and bedrock have a significant impact on dam displacement. $\delta_T$ can be characterized by the previous mean or equivalent temperature of the measured temperature. However, due to technical conditions, the measured value sequence of thermometers buried in concrete dams is often insufficient or incomplete. When the dam is in a quasi-steady temperature field, $\delta_T$ can be simplified as
the form of simple harmonic of multi-period which is also used in this paper. The causes of aging displacement in concrete dams is very complex, which reflects the influence of creep and plastic deformation of concrete and bedrock. \( \delta \) is usually expressed as a combination of a linear and a logarithmic function of time \( t \).

Therefore, the input factor set used to construct the displacement prediction model of the arch dam can be expressed as \([50]\):

\[
S_{\text{input}} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} = \{H, H^2, H^3, H^4, \sin \frac{2\pi t}{365}, \cos \frac{2\pi t}{365}, \sin \frac{4\pi t}{365}, \cos \frac{4\pi t}{365}, \theta, \ln t\},
\]

where \( t \) is cumulative days since the completion date and \( \theta \) is equal to \( t/100 \).

### 3.3. Construction and Comparison of Displacement Prediction Models

#### 3.3.1. Model Parameters

We control the growth of decision trees in the XGBoost model by setting the maximum depth and the number of decision trees in advance. At the same time, in order to improve the generalization ability of the model, each tree does not take all samples as the training set, but randomly selects a certain proportion of samples. When the decision tree grows downward, it does not consider all features, but randomly selects a certain proportion from which to find the optimal feature for bifurcation. The random sampling rates of samples and features are represented by \( R_s \) and \( R_f \), respectively. In addition, the XGBoost model also prevents overfitting by setting the learning rate and regularization terms. As shown in Section 2.2, the two coefficients of regularization terms are \( \gamma \) and \( \lambda \), respectively. Particle swarm optimization algorithm is used to search for the optimal hyper-parameter combination. According to the suggestions of Zhang et al. \([51]\), the search domain of these parameters and optimal parameter combination are shown in Table 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Interpretation</th>
<th>Domain</th>
<th>Optimal Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>max_depth</td>
<td>The maximum depth of a single tree</td>
<td>[3, 10]</td>
<td>7</td>
</tr>
<tr>
<td>n_estimators</td>
<td>The total number of trees</td>
<td>[500, 1000]</td>
<td>953</td>
</tr>
<tr>
<td>( R_s )</td>
<td>Sampling rates of samples</td>
<td>[0.5, 1]</td>
<td>0.858</td>
</tr>
<tr>
<td>( R_f )</td>
<td>Sampling rates of features</td>
<td>[0.5, 1]</td>
<td>0.750</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Learning rate</td>
<td>[0.01, 0.3]</td>
<td>0.219</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>The coefficient of structural risk terms</td>
<td>[1, 3]</td>
<td>1.150</td>
</tr>
<tr>
<td>( \lambda )</td>
<td></td>
<td>[1, 3]</td>
<td>1.005</td>
</tr>
</tbody>
</table>

In order to investigate the accuracy of the XGBoost model itself, three popular machine learning algorithms were selected to establish the dam displacement prediction models respectively, including ANN, random forest (RF) and support vector machine (SVM).

In this paper, because the dimension of input factors is not high, we adopt the network structure with a single hidden layer. The number of neurons in the hidden layer (\( N_e \)) and the learning rate (\( \alpha \)) are set as hyperparameters. The search fields for \( N_e \) and \( \alpha \) are set to \([5, 200]\) and \([0.01, 1]\) respectively. We adopted the ReLU activation function and set the maximum number of iterations during model training as 5000. In the process of model iteration, the adaptive moment estimation algorithm is used to update the weight and bias of the network.

Similar to XGBoost algorithm, RF is also an ensemble learning algorithm based on decision trees. Different from XGBoost algorithm which adopts boosting strategy to integrate the results of multiple decision trees, RF uses bagging strategy to integrate the decision trees. The key parameters of the RF model include the number of decision trees, the
maximum depth of trees and feature sampling rate, the parameter symbols and search field setting of which are the same as those shown in Table 1.

SVM uses the strategy of mapping samples from low-dimensional space to high-dimensional space to make them as linearly separable as possible in order to learn the non-linear mapping relationship between inputs and outputs. The SVM model has two key parameters, kernel coefficient (gamma) and regularization parameter ($C_r$). Their search fields are set to $[0.001, 0.5]$ and $[1, 100]$ respectively.

The input factors for all models are as described in Section 2. To be fair, all machine learning algorithms were optimized by particle swarm optimization (PSO) in advance when constructing displacement prediction models. The size of particles and the number of iterations were set to 20 and 100, respectively. Besides, the inertia weight of particles was set at 0.8, and both learning factors were set at 0.5. Five-fold cross validation was applied to the optimization process. The objective function of the optimization algorithm is to minimize the mean square error on the five validation sets. The parameter optimization results are shown in Table 2.

### Table 2. Parameter optimization results of ANN, RF and SVM models.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>ANN</th>
<th>Parameters</th>
<th>RF</th>
<th>Parameters</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_e$</td>
<td>84</td>
<td>$n_{\text{estimators}}$</td>
<td>638</td>
<td>gamma</td>
<td>0.003</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.018</td>
<td>max_depth</td>
<td>9</td>
<td>$C_r$</td>
<td>16.049</td>
</tr>
<tr>
<td>$R_f$</td>
<td>0.524</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It should be noted that in the ANN model, when the magnitude difference between the weight matrix and the inputs is large, the multiplication of the two items may cause some unnecessary numerical problems. Similar problems will also appear in the SVM model. Therefore, before the training of these two models, both the training set and testing set need to be normalized, and the output results of models should also be inversely normalized. The decision tree considers the relative size of the difference between samples when node splitting and has nothing to do with the magnitude of data. Therefore, there is no need to preprocess the data when establishing the XGBoost model and the RF model, which is also one of their advantages.

#### 3.3.2. Evaluation Indices

It can be seen from Figure 6 that the absolute values of the measured values at PL13–1 monitoring point are mostly above 10 mm. In comparison, the gap between the output values of different prediction models, which may be less than 1 mm or even 0.1 mm, seems very slight. When the outputs of four prediction models are plotted together with the measured values on a single graph, it may not be intuitive to see which model performs better with the naked eye alone. Therefore, in order to objectively reflect the accuracy of models, three indices, mean absolute percentage error (MAPE), mean square error (MSE) and coefficient of determination ($R^2$), are introduced in this paper. The formulas of these indices are as follows:

\[ \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \]

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \]

\[ R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \]

where $y_i$ is the measured value, $\hat{y}_i$ is the predicted value of the model and $\bar{y}$ is the mean value of the measured value. MAPE and MSE are close to 0 and $R^2$ is close to 1, indicating that the model has high accuracy.
3.3.3. Performance Comparison

Four displacement prediction models were established respectively according to the optimal parameter combination given in Table 2. Their performance on the training set and testing set was evaluated and the results are shown in Table 3. It can be observed that the XGBoost model fits the data in the training set almost perfectly, significantly better than the other three models. The RF model, which is also based on ensemble learning, has the second best fitting accuracy, higher than the ANN model. Although the SVM model has the worst performance on the training set, it actually shows satisfactory accuracy.

Table 3. Quantitative evaluation of fitting and prediction performance of displacement prediction models.

<table>
<thead>
<tr>
<th>Set</th>
<th>Indices</th>
<th>XGBoost</th>
<th>SVM</th>
<th>ANN</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAPE</td>
<td>$1.083 \times 10^{-2}$</td>
<td>$1.220 \times 10^{-1}$</td>
<td>$4.221 \times 10^{-2}$</td>
<td>$2.661 \times 10^{-2}$</td>
</tr>
<tr>
<td>Training Set</td>
<td>MSE</td>
<td>$9.018 \times 10^{-2}$</td>
<td>$7.467 \times 10^{-1}$</td>
<td>$3.175 \times 10^{-1}$</td>
<td>$2.264 \times 10^{-1}$</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>$1.000$</td>
<td>$9.969 \times 10^{-1}$</td>
<td>$9.994 \times 10^{-1}$</td>
<td>$9.997 \times 10^{-1}$</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>$1.632 \times 10^{-1}$</td>
<td>$4.623 \times 10^{-1}$</td>
<td>$1.996 \times 10^{-1}$</td>
<td>$2.501 \times 10^{-1}$</td>
</tr>
<tr>
<td>Testing Set</td>
<td>MSE</td>
<td>$9.845 \times 10^{-1}$</td>
<td>$1.839$</td>
<td>$9.269 \times 10^{-1}$</td>
<td>$1.085$</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>$9.930 \times 10^{-1}$</td>
<td>$9.784 \times 10^{-1}$</td>
<td>$9.945 \times 10^{-1}$</td>
<td>$9.925 \times 10^{-1}$</td>
</tr>
</tbody>
</table>

However, the performance of these models on the testing set is not consistent with that on the training set. Firstly, the performance of all four models on the testing set is significantly worse than their performance on the training set. Especially for the XGBoost model and RF model that perform best on the training set, their accuracy attenuation on the testing set is the largest. This means that, even if we adopt a cross-validation strategy to obtain the optimal parameters of the models, these four displacement prediction models based on machine learning still have different degrees of overfitting. It may be that, compared with the other two models, the XGBoost model and RF model based on ensemble learning have higher complexity, which leads to more obvious overfitting. Secondly, the prediction accuracy of the ANN model is comparable to that of the XGBoost model, although it significantly lags behind the XGBoost model on the training set. Thirdly, the prediction accuracy of SVM model and RF model is the worst. Given the excellent performance of the RF model on the training set, it shows surprising error on the testing set, indicating that it has the most serious overfitting.

3.4. Construction of Residual Prediction Models

3.4.1. Data Preparation

From the previous calculation results, it can be seen that the XGBoost algorithm and the ANN algorithm have good accuracy. Therefore, we have reason to expect that the hybrid prediction model combining the two can obtain better model accuracy and reliable interval prediction results based on the residual estimation idea proposed in this paper can be obtained. In order to estimate the prediction error of the XGBoost model more accurately, we obtain the residual sequence based on the method proposed in Section 2.3. The time interval was taken as one month and thus the original training set was divided into 41 months.

The XGBoost model was established each time, with the data during one month as the testing set and the other data in the original training set as the new training set. The models followed the parameter setting illustrated in Section 3.2. For any two sets of training set generated during the traversal process, only one month of data did not coincide. The difference data accounts for a small proportion of the entire training set. Therefore, it can be considered that the amount of information they contain is roughly the same. In view of this, the output results of the models on these 41 testing sets can be approximated as the output results by the same model, which provides a prerequisite for the estimation of prediction residuals.
3.4.2. Performance Exploration

The generated residual sequence on the training set is shown in Figure 8. The residual sequence on the testing set is also drawn in Figure 8, which is the difference between the measured values and the predicted values of the XGBoost model on the original testing set. According to the above results, the ANN algorithm is used to establish the residual prediction model. The input factors of the model are consistent with the previous displacement prediction models. Similarly, in order to give full play to the performance of the model, PSO and cross validation are also used to optimize the structure of the neural network. The fitting results and prediction results of the residual prediction model are shown by purple and red lines in Figure 8, respectively. It can be seen from Figure 8 that both the value and trend of residual sequence can be well fitted by the ANN model, which also indicates that the residual sequence of the displacement prediction model does have valuable information that can be further excavated. This shows that the combined displacement prediction model based on residual estimation is effective.

![Figure 8](image)

Figure 8. The fitting and prediction results of the residual prediction model.

We combined the displacement prediction model based on XGBoost with the residual prediction model based on ANN to generate the hybrid XGBoost-ANN model, so that the point prediction results combining the displacement prediction values and the residual prediction values are obtained. In order to further demonstrate the performance improvement effect of the residual prediction model on the single model and the efficiency of the XGBoost-ANN model, we also carried out residual estimation for the other three displacement prediction models proposed in Section 3.3.1 by referring to the establishment process of the XGBoost-ANN model.

According to the method proposed in Section 2.3, ANN models were established for residual sequences of the SVM model, ANN model and RF model, respectively, and finally the displacement prediction model and residual prediction model were combined. The establishment of the displacement prediction model for concrete dams aims to predict the subsequent dam deformation behavior, and the ability of these four machine learning models to mine the information of the training set has been verified in Section 3.3.3. Therefore, we only focus on the performance of the combined model on the testing set. Based on the three evaluation indexes proposed in Section 3.3.2, the evaluation results of the four combined models are also shown in Table 4.

Table 4. Quantitative evaluation of prediction performance of combined models for displacement prediction.
As can be seen from Table 4, compared with the other three algorithms, the displacement prediction model based on XGBoost algorithm still has the best accuracy. However, different from the results shown in Table 3, it is followed by the RF-ANN model, instead of the ANN-ANN model. This shows that the residual prediction model can effectively compensate for the overfitting phenomenon, so that the performance of the machine learning algorithm can be further played. By comparing the results in Table 3 and Table 4, it can be seen that, compared with the single model, the prediction accuracy of the combined prediction model integrated with residual prediction has been significantly improved. In addition, it is worth noting that even the SVM-ANN model with the worst performance in the four combined models has significantly better prediction accuracy than the XGBoost model with the best performance in the single models.

3.5. Interval Prediction of Displacement

Because the residual prediction model proposed in this paper was trained according to the MLE method, the training error of the model could be considered as the noise sequence of the data. Since the noise sequence is approximately normally distributed, it can be estimated based on the confidence interval method. The variances of the noise sequences corresponding to the four hybrid models are listed in Table 5. We took the significance level of 0.01 and 0.05, respectively, which was commonly used in engineering, and obtained the confidence intervals of the noise sequence. Finally, we combined the point prediction results and interval estimation results of the residual prediction values on the testing set to obtain the interval prediction results of the dam displacement at the PL13–1 monitoring point. The displacement observation value falling outside the interval can be considered as an abnormal displacement with a mathematical probability of 95% or 99%.

Table 5. The discrimination results of outliers.

<table>
<thead>
<tr>
<th>Indices</th>
<th>XGBoost-ANN</th>
<th>SVM-ANN</th>
<th>ANN-ANN</th>
<th>RF-ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>The variance of the noise sequence</td>
<td>0.184</td>
<td>0.475</td>
<td>0.412</td>
<td>0.291</td>
</tr>
<tr>
<td>Number of outliers with a probability of 95%</td>
<td>11</td>
<td>20</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Proportion of outliers with a probability of 95%</td>
<td>4.55%</td>
<td>8.26%</td>
<td>8.26%</td>
<td>6.61%</td>
</tr>
<tr>
<td>Number of outliers with a probability of 99%</td>
<td>5</td>
<td>13</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Proportion of outliers with a probability of 99%</td>
<td>2.07%</td>
<td>5.37%</td>
<td>4.13%</td>
<td>1.65%</td>
</tr>
</tbody>
</table>

It should be noted that the measured displacement falling outside the interval is generally regarded as an outlier. The outliers are suspected of abnormal displacement but they are not necessarily. The causes of outliers should be investigated to further determine the operation state of the dam. Mathematically speaking, even if the prediction error of the model strictly follows the normal distribution, when the significance level is 0.05 or 0.01, 5% or 1% of the measured values will still be judged as outliers, although this is due to random error. In fact, the dam in the case study has been operating safely for the duration of our study, so all outliers are not abnormal displacement and should be considered misjudgments.
A qualified prediction interval should have both excellent anomaly recognition ability and low misjudgment rate. In order to explore the discrimination ability of our proposed interval prediction method on abnormal displacement, we provide statistics on the measured values falling outside the prediction interval, and the results are also listed in Table 5. The proportion of outliers in Table 5 refers to the proportion of outliers in all 242 measured displacements in the testing set. The results show that the hybrid XGBoost-ANN model and RF-ANN model not only have higher sensitivity to outliers, which can be inferred from their smaller noise variance, but also have lower misjudgment rate. At either a significance level of 0.05 or 0.01, their misjudgment rates were only slightly over 5% or 1%. Therefore, the interval prediction method proposed in this paper is reasonable.

Relative to the measured value, the interval width of less than 2 mm is imperceptible on the figure. Therefore, we took the measured value as the reference value to draw the relative value between the point prediction and interval prediction results of the XGBoost-ANN model and the measured values, as shown in Figure 9. In Figure 9, a value of zero means equal to the measured value, while a positive value means greater than the measured value. The solid line in green represents the point prediction results. The dashed line in blue represents the interval boundary when the significance level is 0.05, and the dashed line in red represents the interval boundary when the significance level is 0.01. If the upper dashed line breaks zero, it means that the measured value exceeds the upper bound of the interval; otherwise, if the lower dashed line breaks zero, it means that the measured value exceeds the lower bound of the interval. When the significance level is 0.01, the calculated results show that the measured values at PL13–1 exceed the upper bound of the interval twice on June 14 and July 12, 2018, and exceed the lower bound of the interval on 6 July, 22 July and 23 July 2018. The outliers all appeared in the flood season of this region with the rising of upstream water level, which may be due to the disturbance of the sensor caused by the regulation and storage behavior of the reservoir. In addition, this may also be due to the fact that the time period is far from the period of the training set, which leads to the decline of the prediction accuracy of the model.

Figure 9. The point prediction and interval prediction results of hybrid XGBoost-ANN model.

4. Conclusions and Future Work

In this paper, a hybrid method combining XGBoost and ANN is proposed. The dam displacement is predicted based on the XGBoost model and the residual error is estimated based on the ANN model. Finally, the interval prediction of concrete dam displacement is realized by combining the two with the interval estimation of noise sequence. The proposed method is used to verify the monitoring data of a concrete arch dam in Southwest China. The superiority of the proposed method is proved by comprehensive analysis and calculation. The main conclusions of this paper are as follows:
Both XGBoost algorithm and ANN algorithm have satisfactory accuracy in predicting the displacement of concrete dam. Compared with other algorithms, such as RF and SVM, they have better performance in constructing displacement prediction models. However, the displacement prediction models based on machine learning methods have an obvious overfitting phenomenon.

The residual prediction model based on ANN can mine the residual information in the noise sequence well. Combining various techniques can help to improve the prediction performance of the model. The introduction of a residual prediction model can alleviate the inevitable overfitting phenomenon in machine learning models. After the introduction of a residual prediction model, the performance of the four displacement prediction models mentioned in this paper has been significantly improved.

Similar to the accuracy evaluation results of the single model, the accuracy of the hybrid XGBoost-ANN model is significantly better than that of the other three hybrid models. Therefore, the displacement prediction model and residual prediction method proposed in this paper are effective and feasible. Due to the improvement of model accuracy brought by the alleviation of overfitting phenomenon, the RF model with obvious overfitting phenomenon combined with the residual prediction model performed significantly better than the other two hybrid models.

Based on the results of residual estimation, the XGBoost-ANN model can give high-precision point prediction results. Besides, it also provides reasonable interval prediction of dam displacement. The displacement interval prediction results provided by the XGBoost-ANN model not only have high sensitivity to outliers, but also have a small misjudgment rate.

For future work, interval prediction models suitable for other dam behaviors such as seepage, fractures, stress and strain should be studied so that field personnel can monitor the safety state of dams from comprehensive angles. In addition, the dam displacement point estimation and interval estimation methods proposed in this paper are suitable for a single measurement point, and such approaches for multiple measurement points are also worth developing.

**Author Contributions:** X.Y.: investigation, visualization and writing—original draft. Y.X.: funding acquisition, conceptualization, data curation and methodology. G.S.: formal analysis and software. M.S.: supervision and visualization. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by National Key R&D Program of China (2022YFC3005405), National Natural Science Foundation of China (51979176), Science and Technology Project of Yunnan Province (202203AAA080009), Reservoir Dam Safety and Management Innovation Team Fund Project of Ministry of Water Resources (Y722003).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


