


Article

Seepage Prediction Model for Roller-Compacted Concrete Dam Using Support Vector Regression and Hybrid Parameter Optimization

Mei-Yan Zhuo ¹, Jinn-Chyi Chen ^{1,2,*} , Ren-Ling Zhang ³, Yan-Kun Zhan ³ and Wen-Sun Huang ^{1,4}

¹ School of Hydraulic Engineering, Fujian College of Water Conservancy and Electric Power, Yong'an 366000, China; myzhuo_fcwcep@163.com (M.-Y.Z.); n8894103@gmail.com (W.-S.H.)

² Previously at Department of Environmental and Hazards-Resistant Design, Huaan University, New Taipei 223011, Taiwan

³ Fujian Shuikou Power Generation Group, Youxi Basin Power Generation Co., Ltd., Sanming 365100, China; zrl02120@163.com (R.-L.Z.); zj980311@163.com (Y.-K.Z.)

⁴ Previously at Ecological Soil and Water Conservation Research Center, National Cheng Kung University, Tainan 70101, Taiwan

* Correspondence: jinnchychen@gmail.com or chenjinnchyi@163.com

Abstract: In this study, a seepage prediction model was established for roller-compacted concrete dams using support vector regression (SVR) with hybrid parameter optimization (HPO). The model includes data processing via HPO and machine learning through SVR. HPO benefits from the correlation extraction capability of grey relational analysis and the dimensionality reduction technique of principal component analysis. The proposed model was trained, validated, and tested using 22 years of monitoring data regarding the Shuidong Dam in China. We compared the performance of HPO with other popular methods, while the SVR method was compared with the traditional time-series prediction method of long short-term memory (LSTM). Our findings reveal that the HPO method proves valuable real-time dam safety monitoring during data processing. Meanwhile, the SVR method demonstrates superior robustness in predicting seepage flowrate post-dam reinforcement, compared with LSTM. Thus, the developed model effectively identifies the factors related to seepage and exhibits high accuracy in predicting fluctuation trends regarding the Shuidong Dam, achieving a determination coefficient $R^2 > 0.9$. Further, the model can provide valuable guidance for dam safety monitoring, including diagnosing the efficacy of monitoring parameters or equipment, evaluating equipment monitoring frequency, identifying locations sensitive to dam seepage, and predicting seepage.

Keywords: support vector regression; hybrid parameter optimization; grey relational analysis; principal component analysis; roller-compacted concrete dam; seepage



Citation: Zhuo, M.-Y.; Chen, J.-C.; Zhang, R.-L.; Zhan, Y.-K.; Huang, W.-S. Seepage Prediction Model for Roller-Compacted Concrete Dam Using Support Vector Regression and Hybrid Parameter Optimization.

Water **2023**, *15*, 3511. <https://doi.org/10.3390/w15193511>

Academic Editor: Jianjun Ni

Received: 29 August 2023

Revised: 2 October 2023

Accepted: 5 October 2023

Published: 8 October 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Reservoir dams are large-scale transnational projects constructed for water supply, agriculture, industry, power generation, flood control, tourism, and cultural entertainment, among other purposes. Various countries consider them as major projects [1]. Among these projects, roller-compacted concrete (RCC) dams have gained popularity in China over the last century as they combine the safety of concrete dams with the construction convenience of roller dams. Accordingly, thorough assessments of dam safety are becoming increasingly crucial to mitigate the potential risk of dam breakage. Statistics indicate that dam breaks are predominantly caused by seepage, accounting for 30–40% of all dam failures [2]. However, accurate placement of necessary seepage sensors, particularly for older RCC dams with outdated monitoring systems, remains challenging due to the random nature of seepage occurrences. Therefore, collecting data on seepage often requires manual observations.

There are four common modeling methods associated with dam safety and dam monitoring issues: deterministic method, statistical method, machine learning algorithm, and hybrid method. The deterministic method is a model developed based on the dam's structural form, mechanical parameters, and foundation and environmental characteristics. Using these, the seepage of the dam is described and predicted through theoretical or numerical analysis, such as a one-dimensional consolidation theory for dam seepage in saturated soils [3], a mathematical model between the seepage field and the stable temperature field of a dam body [4,5], and a finite element simulation model for the temperature field of a dam body on seepage fields [6]. However, this numerical model has several limitations, such as the requirement for large amounts of field data, complex calibration procedures using rigorous optimization techniques, and a comprehensive understanding of the underlying physical processes [7,8]. Some researchers use statistical methods, such as those widely used in multiple linear regression, stepwise regression, and partial least squares regression, to solve unknown model coefficients for studying the thermal displacement of concrete dams [9] and monitoring dam structural health [10]. The advantages of statistical methods are clear physical explanations, simple model structure, and fast execution speed [11]. However, due to the fact that most statistical methods use linear regression, this often limits the accuracy and reliability of fitting when dealing with complex nonlinear problems such as dam seepage and deformation. Recently, machine learning algorithms (MLAs) have demonstrated great potential in predicting and modeling nonlinear characteristics in many fields, such as analysis on air quality [12,13], hydrology [14–16], and dam seepage [17,18]. MLAs provide an effective method to handle vast amounts of dynamic, nonlinear, and noisy data, particularly in cases where essential physical relationships cannot be accurately understood, making MLAs particularly suitable for interpreting dam behavior [19]. In many cases, MLAs can provide more accurate predictions than statistical models [11]. There are also hybrid methods that combine MLAs with deterministic or statistical method to improve the accuracy of the prediction model for dam deformation [20] and seepage [21,22].

The most commonly used MLAs in dam safety and monitoring issues include Gaussian process regression [19,23], support vector machine (SVM) [11,24–28], artificial neural networks (ANNs) [29–32], extreme learning machine [33], adaptive network-based fuzzy inference system [34], radial basis function networks [35], random forest [36], boosted regression trees [37], and extreme gradient boosting [38]. Looking back at previous studies, methods used by researchers may not be consistent due to differences in dam types, environmental factors, and monitoring and simulation parameters. Overall, SVM and ANN are the two most widely used MLAs in dam safety monitoring [11,35]. The back propagation neural network (BPNN) and long short-term memory (LSTM) in ANNs were used for predicting dam seepage and uplift pressure [31,32]. The accuracy of seepage prediction can be further improved by using ANNs. However, dam safety monitoring systems lack the support of high-performance clusters, where deep neural networks are rarely applied, urgently requiring a simple but efficient seepage prediction model. Compared with ANN, SVM has advantages in solving small sample, high-dimensional, and nonlinear problems [11,19,28]. Accordingly, in this study, we used a regression algorithm based on SVM, namely, support vector regression (SVR), which has excellent nonlinear processing capability for solving the nonlinear problem of dam seepage prediction and is edge-end device-friendly owing to its calculation simplicity [39].

In addition, different forms of dams, such as roller compacted concrete (RCC) dams and traditional concrete dams, have different seepage mechanisms and characteristics [40]. However, there have been few previous studies focusing on the prediction of seepage in RCC dams. In the study of RCC dams, Wei et al. [21] used statistical regression methods combined with numerical analysis to establish a seepage prediction model for RCC dams. Their research focuses on the variation characteristics of monitoring parameters over time and the delayed effects of water level and rainfall. The input factors for SVR are usually selected based on experience or manual engineering [39,41,42]. However, this study focuses

on the effectiveness of screening all monitoring parameters without preset conditions as well as establishing a seepage prediction model for RCC dams using simple and quick-acting SVR combined with hybrid parameter optimization (HPO).

This study takes the Shuidong RCC dam in China as an example for seepage prediction. The dam has 22 years of monitoring data, during which time it has experienced reinforcement engineering. The seepage before and after dam reinforcement has undergone significant changes; thus, the dam provides a good example for the development of a seepage prediction model of an RCC dam and to examine the reasonability of the model. However, previous studies lacked long-term monitoring data to develop and examine the related model. In addition, multiple physical parameters were monitored in the case of this study, with monitoring points distributed around the dam. There are a total of 60 physical monitoring parameters covering both temporal and spatial factors. Some of the monitoring factors may be ineffective, but, currently, there is no effective data processing method in this study area to verify the applicability of monitoring points, and there is a lack of intelligent seepage prediction models. Therefore, in order to develop an optimal input factor set for seepage prediction in this study, grey relational analysis (GRA) was employed to identify input factors showing a high correlation with seepage prediction, followed by principal component analysis (PCA) to eliminate input factors with duplicate contributions. The effectiveness and precision of the SVR model with hybrid parameter optimization (HPO) for predicting seepage in RCC dams were demonstrated. The primary contributions of this study are specified below.

- (1) A novel SVR-based prediction model for RCC dam seepage is proposed and evaluated using 22 years of monitoring data with two distinct seepage patterns (before and after dam reinforcements), demonstrating good prediction accuracy and robustness.
- (2) An HPO approach is introduced to screen the input factors of the SVR model, which combines the correlation analysis ability of GRA and the data dimensionality reduction capability of PCA.
- (3) The proposed SVR model incorporating HPO provides new insights for seepage research and safety monitoring of RCC dams, including the placement of uplift pressure orifices as well as the selection of the type and frequency of dam-monitoring data.
- (4) The methodology employed by HPO for the selection and screening of input parameters in seepage prediction models exhibits broader application potential in advanced prediction modeling work.

2. Study Area

2.1. Background on Shuidong Dam

Shuidong Dam is located on the Youxi River, a tributary of the middle reaches of the Minjiang River in Fujian Province, China (Figure 1). The watershed area is 3784.5 km² above the dam site. The Shuidong Dam is an RCC dam constituting two water-retaining dam sections on the left and right banks and one overflow section, as shown in Figure 2. The maximum height of the dam is 63 m, and the top length is 197 m. The dam was completed in 1994 and is mainly used for hydroelectric power generation and partially to supply public water. Owing to the poor quality of the incompletely compacted concrete sections and severe seepage in the dam body, the safety of the dam was endangered. Therefore, reinforcement of the dam was conducted from 14 October 2002 to 1 July 2003.

2.2. Dataset Description

The physical parameters considered for monitoring the dam include seepage flows, upstream and downstream water levels, uplift pressures, vertical and horizontal displacements, and air temperatures. The data recording period included pre-reinforcement monitoring data from 1 September 1999 to 31 December 2002 and post-reinforcement monitoring data from 1 January 2004 to 30 June 2021. Table 1 lists the recording frequency, sample size, and monitoring methods for all physical parameters. These physical parameters were measured at different locations and recorded at different scales: day, month, quarter, and

year. Thus, a total of 60 factors (labeled 1–60) were considered in this study, as listed in Table A1 in Appendix A.1, including seepage flow q_s ; dates in year, quarter, month, and day (denoted as t_y , t_q , t_m , and t_d , respectively); upstream water level h_u ; downstream water level h_d ; water level difference h_D between h_u and h_d (i.e., $h_D = h_u - h_d$); uplift pressure coefficient C_{li} ; vertical and horizontal displacements with interval and cumulative (denoted as d_{Vi} , D_{Vi} , d_{Hi} , and D_{Hi} , respectively); elevation E_{li} ; and air temperature T_i . These factors were mostly measured at different positions around the dam, such as T_i and C_{li} from 13 observation sites at the dam base; d_{Vi} , D_{Vi} , and E_{li} from 7 observation points; and d_{Hi} and D_{Hi} from 3 sites. The minimum recording unit for these factors is the day. However, in cases of heavy rainfall or abnormal monitoring values, some factors such as seepage flow may be recorded multiple times a day. In these cases, abnormal data were excluded and the multiple recorded values were averaged as representative data for the day. Owing to inconsistent sampling or recording times for each physical parameter, samples with the same corresponding time for all factors were selected for analysis in this study.

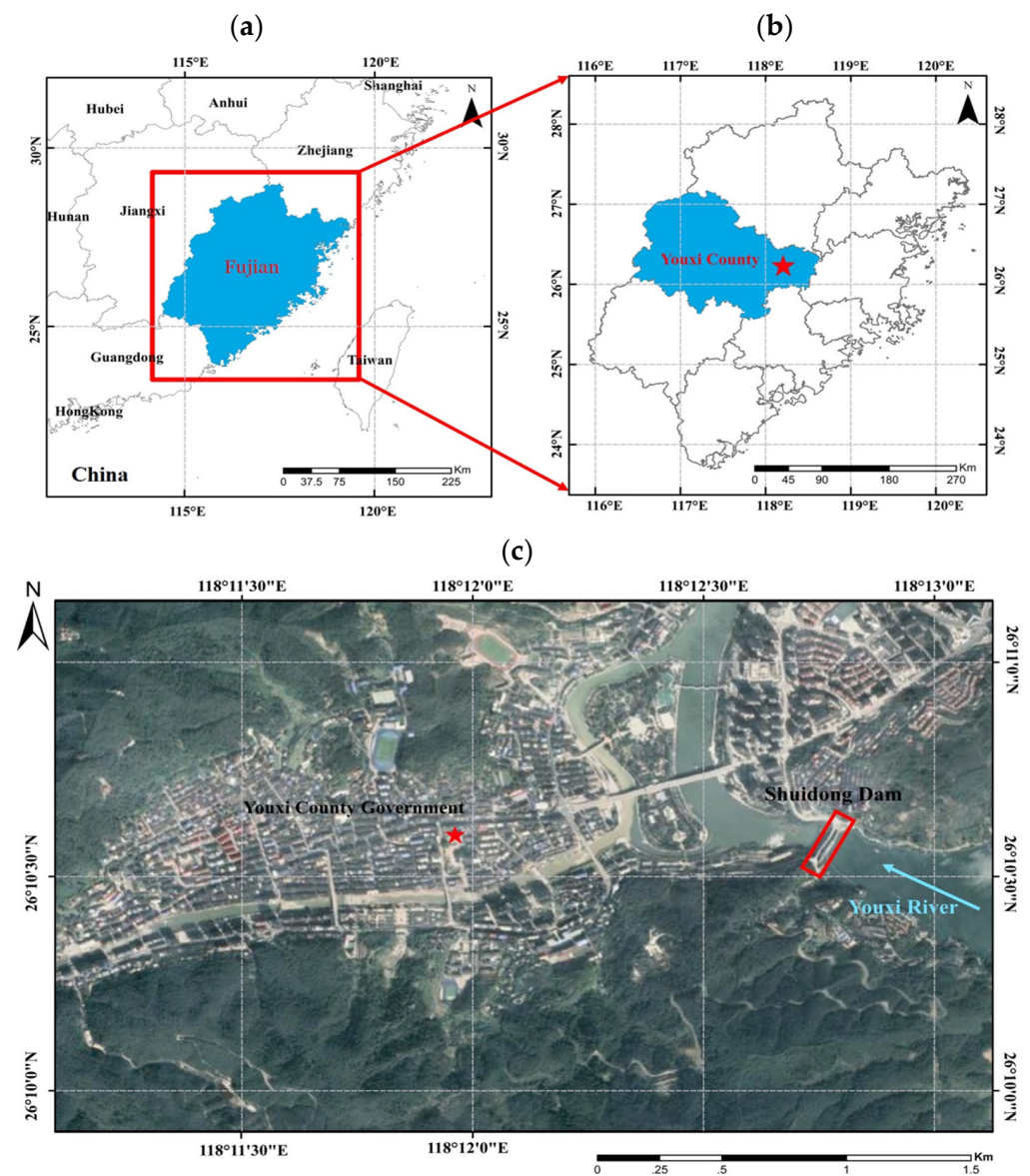


Figure 1. Location of Shuidong Dam on the Youxi River, a tributary of the middle reaches of the Minjiang River in Fujian Province, China. (a) Fujian Province, (b) Youxi county and county government (red star symbol) in Fujian Province, and (c) Surroundings of Shuidong Dam.

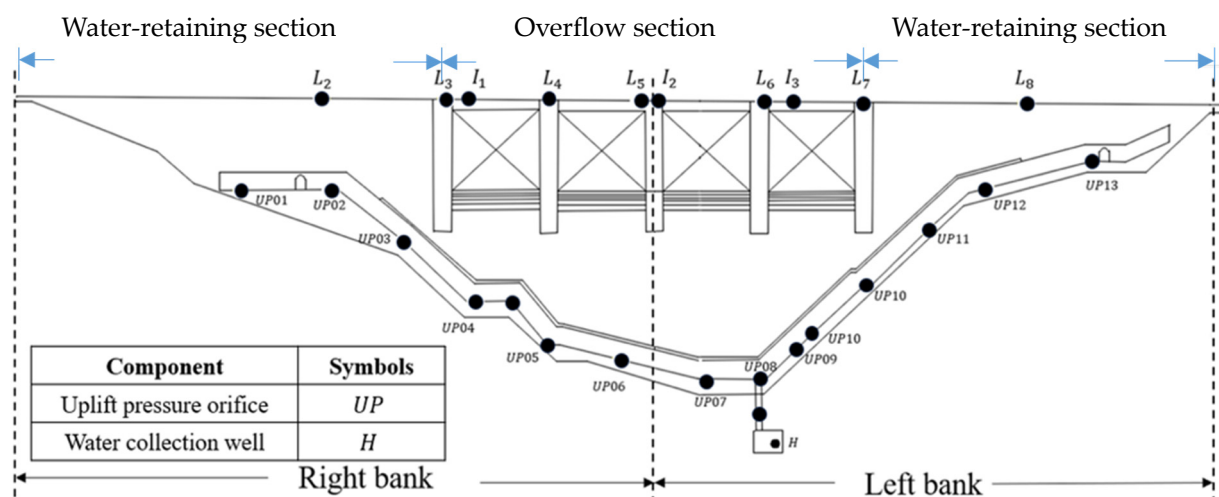


Figure 2. Cross section and compositions of Shuidong Dam, including two retaining water sections on the left and right banks and one overflow section. The changes in temperature, water level, and uplift pressure were observed using 13 uplift pressure orifices spanning the dam foundation, namely, UP01–UP13. A collection well, indicated as H, was used to measure seepage. Horizontal and vertical displacements were measured at three points, I₁–I₃, and seven points, L₂–L₈, on the dam crest, respectively.

Table 1. Physical parameters monitored from 1 January 1999 to 30 June 2021, in this study.

Monitoring Parameter	Monitoring or Recording Method	Recoding Frequency	Size of Sample
Seepage flow q_s (l/h)	A collection well at H (Figure 2) collected the total seepage flow rate q_s from the dam body and the dam base at the right and left banks of the dam; q_s was determined by water level changes at the well, measured using an electromagnetic water level gauge with an accuracy of ≤ 0.02 mm.	Data are generally recorded once a day. In special circumstances, such as heavy rainfall or abnormal seepage, some parameters will be measured multiple times a day.	9097
Upstream water level h_u	The upstream water level was monitored at the water inlet of the dam. A pontoon water level gauge was used to measure the water level with an accuracy of ≤ 0.01 m.		8253
Downstream water level h_d	The downstream water level was monitored at the tailwater of the dam. A pontoon water level gauge was used to measure the water level with an accuracy of ≤ 0.01 m.		8253
Uplift pressure coefficients C_{li}	Piezometers were installed at 13 locations, UP01–UP13 (Figure 2), to determine C_{li} through water levels or rock bed level *. The piezometer was manufactured by Geokon (GK4500AL-70KPa model) with a measurement range of 170 KPa and an accuracy of 0.025% FS.		8339–9061
Air temperature T_i	The temperature was measured using a thermistor at the same 13 positions, UP01–UP13 (Figure 2), with an accuracy of ± 0.02 °C.		9195–9559

Table 1. Cont.

Monitoring Parameter	Monitoring or Recording Method	Recoding Frequency	Size of Sample
Elevation E_{li}	Elevation E_{li} was observed at 7 points, L ₂ –L ₈ (Figure 2), at the dam top, using 1st-class digital levels (Leica DNA03) with an accuracy of 0.2".		288
Vertical displacement, d_{Vi} and D_{Vi} (mm)	Changes in elevation, including interval vertical displacement d_{Vi} (mm) and accumulated vertical displacement D_{Vi} (mm), were calculated at 7 observation points, L ₂ –L ₈ (Figure 2), at the dam top. D_{Vi} is the summation of d_{Vi} calculated from 1 January 1999. d_{Vi} was determined by 1st-class digital levels (Leica DNA03) with an accuracy of 0.2".	Data are generally recorded once a month. In special circumstances, such as heavy rainfall or abnormal monitoring values, some parameters will be measured multiple times a month.	288
Horizontal displacement, d_{Hi} and D_{Hi} (mm)	Interval horizontal displacement d_{Hi} (mm) and accumulated horizontal displacement D_{Hi} (mm) were measured at 3 observation points, I ₁ –I ₃ (Figure 2), at the dam top. D_{Hi} is the summation of d_{Hi} calculated from 1 January 1999. d_{Hi} was determined using a total station (Leica TS60i) with an accuracy of 0.5".		308

Note(s): * Uplift pressure coefficient C_{li} at measuring point i ($= 1-13$) with water level h_i can be calculated using bottom water level h_d , top water level h_u , or bedrock elevation at measuring point h_{bi} ; i.e., $C_{li} = (h_i - h_d) / (h_u - h_d)$ when $h_d > h_{bi}$, and $C_{li} = (h_i - h_{bi}) / (h_u - h_d)$ when $h_d < h_{bi}$.

This work aimed to develop a model of seepage associated with the factors listed in Table A1 for the Shuidong Dam. All the factors monitored locally were considered for two main reasons: firstly, to avoid artificially filtering out important input parameters; secondly, to find parameters with high correlation through system optimization and check the rationality of the proposed model. That is to say, the input parameters in this study assume that all monitoring factors (including time and space) are valid and then filter out invalid factors through data processing methods. In addition, random selection was performed to select 10% of the pre- and post-reinforcement data for the training and validation of the model while the remaining 90% were used as the test set.

3. Methodology: Dam Seepage Model with Hybrid Parameter Optimization (HPO)

In this study, a dam seepage model with HPO based on SVR was proposed considering 22 years of monitoring data on the Shuidong Dam; a corresponding flowchart of the model is presented in Figure 3. The HPO approach, which combines the correlation analysis ability of GRA and the data dimensionality reduction capability of PCA, was introduced to the input factors of SVR. The principles of the GRA, PCA, and SVR are briefly described in the following sections.

3.1. HPO

Data processing, which includes cleaning, transforming, and organizing data, is crucial for MLA because it helps prepare data optimization for analysis and modeling. Two methods are commonly used for data optimization in MLA. One method screens for the relevance of predictors, such as the Pearson correlation coefficient method and point-biserial correlation, which can screen out input factors with higher relevance to the prediction variables. However, these methods cannot distinguish between the impact factors with repeated contributions [43]. The other method reduces the spatial dimensionality of feature variables, such as PCA, discriminant analysis, and multidimensional scaling. These methods have more restrictions on the type of data and can only retain the primary data information, which does not guarantee the validation and relevance of the information [44,45]. Thus, we introduced an HPO technique that selects input factors with high

correlation using GRA. The input factors with repeated contributions were subsequently removed via PCA for the dam seepage prediction model. The principles and methods of the GRA and PCA are briefly presented in the following two sections.

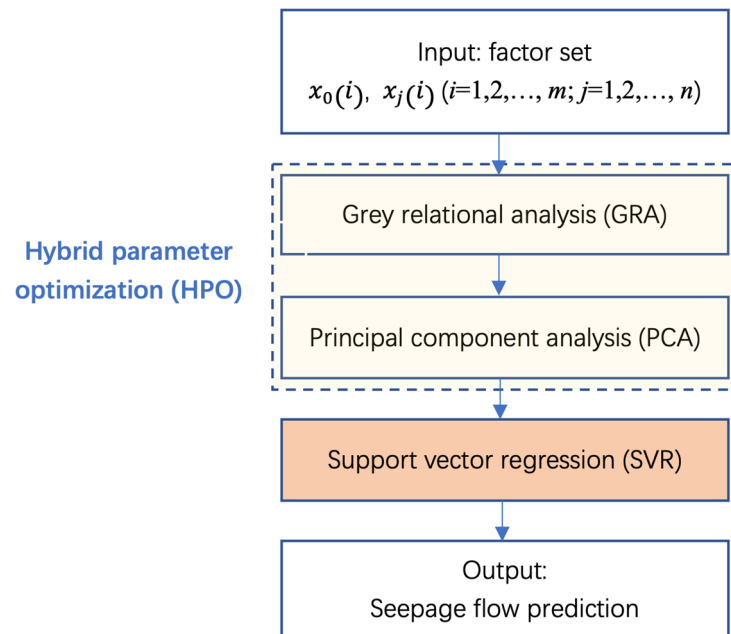


Figure 3. Flowchart of proposed dam seepage model with hybrid parameter optimization (HPO) based on support vector regression (SVR).

3.2. Grey Relational Analysis (GRA)

GRA is used in grey system theory [46] to determine whether the elements in two systems are homogeneous or heterogeneous. If the development trends of the two elements are the same, then there is a strong correlation between them [47,48]. The basic concept of GRA is to determine the degree of geometric similarity between reference and comparison sequences and compare how closely they are related: the more similar they are, the greater the correlation. GRA involves the following steps: determining the input factors (reference and comparison sequences), normalizing the factors, calculating the grey relational coefficient, and sorting the grey relational grades.

The input factors in this study include the seepage flow rate (reference sequence) and other monitoring factors (comparison sequence). All sequence data should be normalized to the same order of magnitude, which is suitable for the rational analysis of the GRA. If $x'_j(i)$ is a monitoring value of an input factor x'_j at sequence sampling point i , in which $i = 1, 2, \dots, m$, and m is the number of sequence samples for the factor, the normalization of the factor $x_j(i)$ can be calculated as:

$$x_j(i) = \left(x'_j(i) - \min(x'_j) \right) / \left(\max(x'_j) - \min(x'_j) \right) \quad (1)$$

where j is the number of factors $j = 0, \dots, n$: $j=0$ for the reference sequence ($x_0(i)$), and $j = 1, \dots, n$ for the comparison sequences. Here, $x_0(i)$ is represented by the seepage flow rate $q_s(i)$. The grey rational coefficient between the sequences of $x_j(i)$ and $q_s(i)$, i.e., $\gamma(q_s(i), x_j(i))$, is defined as:

$$\gamma(q_s(i), x_j(i)) = \frac{\min_j \min_i |q_s(i) - x_j(i)| + \zeta \max_j \max_i |q_s(i) - x_j(i)|}{|q_s(i) - x_j(i)| + \zeta \max_j \max_i |q_s(i) - x_j(i)|} \quad (2)$$

where ζ is the distinguished coefficient, which is typically set to 0.5. By averaging the grey relational coefficients γ obtained from all sampling points i ($= 1, 2, \dots, m$) in factor j , the grey relational grade at j , $r_G(j)$, between $x_j(i)$ and $q_s(i)$ can be calculated as follows:

$$r_G(j) = \frac{1}{m} \sum_{i=1}^m \gamma(q_s(i), x_j(i)) \quad (3)$$

Based on the long-term monitoring data of seepage flow $q_s(i)$ and multiple factors $x_j(i)$, in which $j = 1$ to 60 (as indicated in Table A1), from 1999 to 2021 in the Shuidong Dam, this study calculated the grey relational coefficient using Equation (2), and the grey relational grade using Equation (3); the grey relational grades for all factors were sorted to determine the factors of high relational grade ($r_G(j) > 0.4$) for the subsequent PCA.

3.3. Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that uses a linear transformation to transform original high-dimensional data into several variables in a lower mutually exclusive dimension [49,50]. Compared with other dimension-reduction techniques, the greatest possible amount of original information can be obtained using PCA [51,52]. Therefore, the PCA method was adopted in this study to reduce the dimensionality of the factors after GRA as well as to shorten the calculation time of the models developed subsequently. The basic principle of PCA involves diagonalizing a matrix and then calculating its eigenvectors and eigenvalues. The PCA involves the following processes for data dimensionality reduction: first, the sample matrix must be standardized; second, the covariance matrix is calculated; third, the eigenvalue of the covariance matrix and the eigenvector are determined; finally, the cumulative contribution rate of the eigenvector is obtained. The first k eigenvectors with a cumulative contribution rate greater than 99% are extracted, and the dimensionality reduction eigenvector matrix is realized. The cumulative contribution rate L of the first k eigenvectors was calculated as:

$$L = \sum_{i=1}^k \lambda_i / \sum_{i=1}^m \lambda_i \quad (4)$$

where λ_i represents the eigenvalues of an eigenvector E_i within the number of eigenvectors m . Dimensionality reduction eigenvector matrix μ_T was derived from the transpose of the matrix of eigenvector μ_i as:

$$E_T = (E_1, E_2, \dots, E_k)^T \quad (5)$$

The optimal number of principal components is an important part of HPO. The optimal number of main components is determined by the following steps. Firstly, we calculate the contribution value or eigenvalue λ_i at each principal component E_i and sort them from maximum ($i = 1$) to minimum ($i = m$). Secondly, we calculate the cumulative contribution rate L and select components that meet $L > 99\%$. Thirdly, we calculate the error and accuracy (as described in the following Equations (6)–(8)) under different principal components as well as the time spent (such as training and validation time and file I/O time). Finally, we choose feature vectors that have a shorter computation time and higher accuracy. Note that step 3 must be combined with SVR analysis.

3.4. Support Vector Regression (SVR)

SVR is an MLA used for regression analysis. The basic idea behind SVR is to determine a function that can predict the values of a target variable based on the values of one or more input variables [53]. Primarily, in SVR, input variables are mapped into a high-dimensional feature space in which a linear regression model can be used to make predictions. The algorithm then uses a subset of training data, known as support vectors, to define a hyperplane that separates the data into different classes. The goal of the SVR method is to determine a hyperplane that minimizes the error between the predicted and actual values.

One of the main advantages of SVR is its ability to handle nonlinear relationships between the input and target variables. This is achieved using nonlinear kernel functions to map the data into a higher-dimensional feature space where a linear regression model can be used. This has a wide range of applications, including time-series forecasting, stock price prediction, and image analysis [54–56].

The process of developing a seepage model based on SVR in this study is detailed as follows: first, we selected the first k principal components of the PCA as the input variables for the SVR. Second, the parameter set must be established, specifically, the optimal values of penalty coefficient c and kernel function parameter g , which provide the model training. Then, we compared the output of the model, i.e., the predicted seepage flowrate, with the on-site monitored seepage flowrate and determined the coefficient of determination and mean square error. During this process, the times required for model training, testing, and validation were measured.

During the training and validation processes of our SVR model, penalty coefficient c and kernel function parameter g were determined using the K-fold cross-validation (K-CV) method to identify the optimal combination [57–59]. The K-value was equivalent to the number of groups in the partitioned training set. There is no specific limit on how much is chosen from K; however, the larger the K value, the more groups, and the smaller the size of each group, which may not be sufficient to train the model. Meanwhile, the smaller the K value, the fewer groups, which may lead to overfitting of the model: $K = 5$ is a relatively moderate parameter and is widely used. Thus, in this study, we employed fivefold validation with a grid search to obtain the best parameters for the SVR model. The group of penalty coefficients c and kernel function parameters g with the highest R^2 value and smallest value of c were chosen as the training parameters.

The calculation of the SVR model was divided into training and test subsets. After training, model performance was evaluated using the test subset, including the accuracy, error, and relevant execution times of the model, such as testing, training, and validation times. The errors were assessed using the common indicators of mean squared error (E_{ms}) and mean absolute percentage error (E_{ma}). The accuracy was determined by coefficient of determination (R^2). E_{ms} , E_{ma} , and R^2 were calculated as follows:

$$E_{ms} = \frac{1}{n} \sum_{i=1}^n (q_{si} - q'_{si})^2 \quad (6)$$

$$E_{ma} = \frac{1}{n} \sum_{i=1}^n \left| \frac{q_{si} - q'_{si}}{q_{si}} \right| \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (q_{si} - q'_{si})^2}{\sum_{i=1}^n (q_{si} - \bar{q}_{si})^2} \quad (8)$$

where q'_{si} , q_{si} represent the predicted seepage flow and monitored seepage flow, respectively, at the i -th sample and \bar{q}_{si} is mean value of q_{si} . The range value of R^2 is from 0 to 1. The higher the value of R^2 , the higher the accuracy of the model.

In this study, the Sklearn package in Python was used for the SVR calculation. The radial basis function (RBF) is employed as the kernel function [60]. The K-fold cross-validation (K-CV) method for parameter optimization was also obtained from the Sklearn package. In this study, a random selection process was utilized to select 10% of the pre- and post-reinforcement data for training and validation of the model, while the remaining 90% were used as the test set.

For assessing the performance of our model, the Python time package was employed to record the time, and the hardware configuration included an Intel i5-12500 2.5 GHz processor and 16 GB of RAM. The analysis was conducted using Python version 3.10.5.

4. Results and Discussions

4.1. HPO

4.1.1. GRA

The seepage flow of a dam is affected by several factors [27,61] including the uplift pressure, temperature, and displacement. This study considered all the factors monitored locally for two main reasons: first, to avoid artificially filtering out important input parameters, and second, to find parameters with high correlation through system optimization and check the reasonableness of the model. A total of 60 input factors, as indicated in Table A1, were considered to analyze the correlation with seepage flow q_s . The grey relational grades for all factors were calculated using Equation (3). Figure 4 shows a grey relational grade r_G between the input factors and seepage flow q_s . A clear cutoff value is observed at $r_G = 0.4$, indicating that the 14 factors with $r_G > 0.4$ (shaded in light green in Figure 4) have a higher correlation with seepage flow q_s . These factors include the temperature at point UP04 (T_4); uplift pressure coefficients at points UP01, UP02, UP03, UP04, UP06, UP11, and UP12 (i.e., C_{I1} , C_{I2} , C_{I3} , C_{I4} , C_{I6} , C_{I11} , and C_{I12}); interval vertical displacement at points L2, L3, and L7; and cumulative horizontal displacement at points I1, I2, and I3, as shown in Figure 2. As indicated by the GRA results, the temperature, uplift pressure, and vertical displacement exhibited a strong correlation with seepage flow. This finding aligns with common hydrological knowledge [62,63]. Notably, 13 uplift pressure orifices (Figure 2) in the dam base were considered to survey the uplift pressure coefficients C_{Ii} ($i = 1-13$), but only 7 (C_{I1} , C_{I2} , C_{I3} , C_{I4} , C_{I6} , C_{I11} , and C_{I12}) showed a high correlation with seepage flow. This result may help guide the layout and monitoring of uplift pressure device points by increasing or decreasing the frequency of monitoring points related to seepage. However, among the factors of higher r_G , the datasets of d_{Vi} and D_{Vi} are small, with less than 235 samples, which could potentially impact the accuracy of the subsequent prediction model; consequently, these factors were excluded from the analysis. Thus, the final impact factors determined via the PCA and SVM analyses were air temperature T_4 and uplift pressure coefficients C_{I1} , C_{I2} , C_{I3} , C_{I4} , C_{I6} , C_{I11} , and C_{I12} . All factors had the sample number of 6585 on the same date scale in days. The input impact factors had dimensions of 8×6585 . The advantage of the GRA is that it can quickly check input factors that are highly related to seepage; however, some input factors may contain duplicate information in subsequent SVR seepage prediction models. This redundancy may prolong the training time of the model and consume more resources to store the monitoring data, which is not ideal for real-time energy-saving monitoring of dam safety. Therefore, PCA was performed after the GRA to eliminate redundant information from the input factors.

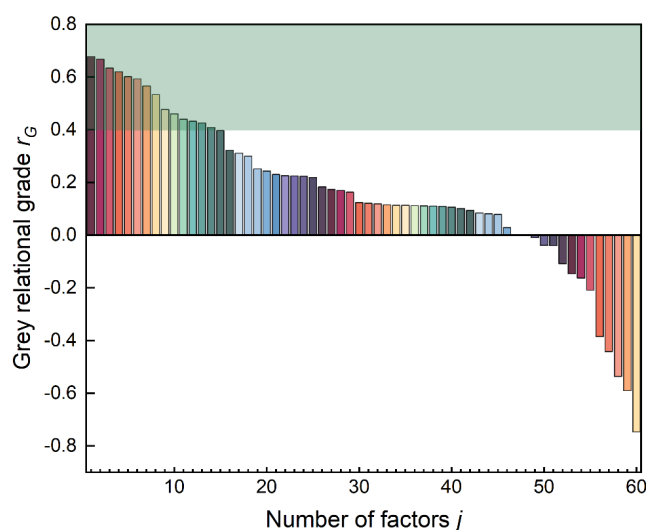


Figure 4. Grey relational grade r_G between input factors x_j (number of factors $j = 1$ to 60) and seepage flow q_s . The 60 factors used in the GRA are listed in Appendix A, Table A1.

4.1.2. Comparison with On-Site Investigation

From the on-site investigation of Shuidong Dam, it was found that there is a large amount of calcium carrier (white substance) on the inner wall of the dam. This is because the seepage of the dam carried away the calcium in the concrete, and it remained on the surface of the wall, which is particularly evident on the right bank of the dam compared to the left bank. From the data of calcium sampling collected from multiple points on the corridor of the dam between 2011 and 2017, we found that the average proportion of calcium content in the water on the right bank is 12.91%, and the proportion on the left bank is 10.53%. The percentage of calcium in the seepage water on the right bank was higher than that on the left bank. Moreover, from the on-site investigation results, monitoring of water drainage from the dam foundation on the right bank was more obvious than that on the left bank.

The GRA results indicate that 14 out of the 60 factors have a high grey relational grade with seepage, including seven uplift pressures at points UP01, UP02, UP03, UP04, UP06, UP11, and UP12, horizontal and vertical displacements at six monitoring points (L2, L3, L7, I1, I2, and I3), and a thermometer at UP4 point (T_4), with most of them distributed on the right bank of the dam, as shown in Figure 2. In this case, seepage on the right bank of the dam may be more pronounced than that on the left bank, which is roughly consistent with the on-site investigation results. This can inform the management unit that it may be necessary to strengthen the monitoring work on the right bank, such as increasing the sampling frequency of the current monitoring points or adding new monitoring points. For the daily recorded uplift pressure data, it is possible to consider increasing the number of measurements, especially during heavy rainfall or when the upstream water level of the dam increases. We are also considering adding new uplift pressure measurement points on the right bank to provide more favorable data. At present, the dam only monitors the total seepage flow from the left and right banks, and the monitoring of individual seepage flows from the right and left banks also needs to be considered.

Date-related factors, including t_y , t_q , t_m , and t_d , and most thermometers installed at the uplift observation points, excluding thermometer T_4 , exhibit little correlation with seepage. Temperature changes often affect the seepage flow rate of a dam. The results of the GRA in this study show that among the 13 long-term monitoring thermometers, only thermometer T_4 presents a high correlation with the seepage flowrate. These thermometers are of the thermistor type, the operation of which relies on the inside of the desiccant and requires frequent inspection and replacement. When the desiccant becomes wet, the temperature cannot be measured accurately. In the internal corridor of a dam, humidity often reaches 90%, and these thermometers need to be replaced every 2–3 days for accuracy. This may be because most of the thermometers installed at the uplift pressure observation points were not maintained, and their measurement values were not accurate. Currently, they have been replaced with observation points outside the corridor. Point T_4 may be one of the best-maintained thermometers in the internal corridor. Based on the detection results of the thermometer, the GRA can effectively detect useful data as well as filter out invalid monitoring points or factors. This will serve as an important reference for investigators or monitors to diagnose whether the monitoring device, such as a thermometer, in this example, is functioning properly.

4.1.3. PCA

Figure 5 displays the principal component contribution and cumulative principal component contribution associated with various eigenvectors or principal components (E_1 – E_8) based on PCA. The cumulative principal component contribution of the first one eigenvector (E_1), or the first principal component, exceeded 95%, and the top three eigenvectors (E_1 , E_2 , and E_3) contributed more than 99%. When the top five principal components were selected, the cumulative principal component contribution rate L (by Equation (4) with $m = 8$ and $k = 5$) was approximately 99.9%, indicating that characteristics of factors can be retained via the projection of the original eight-dimensional data into

three dimensions using PCA. Figure 6 depicts the changes in monitoring seepage q_s and various principal components from 1999 to 2021. The values of the principal component or eigenvalues extracted from the input data are also presented. The fluctuations of the first principal component are sharp, while the remaining seven components exhibit smaller values and fluctuations. Therefore, we introduced SVR to analyze the nonlinear relationship between the eight eigenvectors and the seepage flowrate.

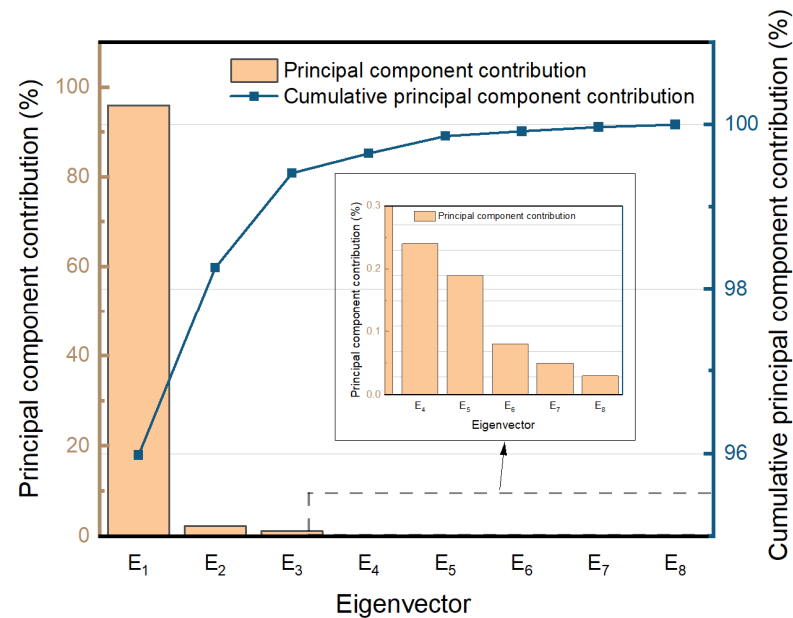
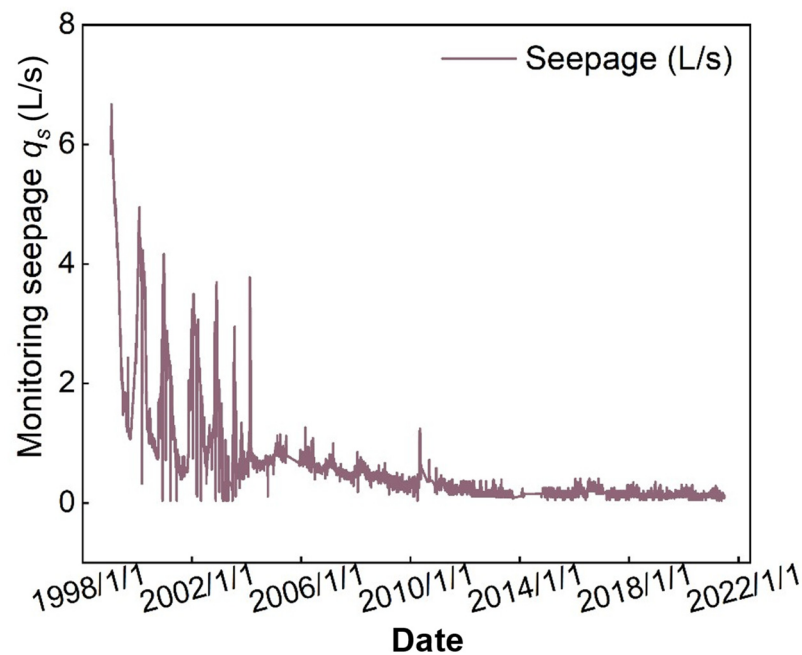


Figure 5. Principal component contribution (eigenvalue λ_i) and cumulative principal component contribution rate L associated with various eigenvectors or principal components (E_1 – E_8) based on PCA.



(a) Monitoring seepage q_s

Figure 6. Cont.

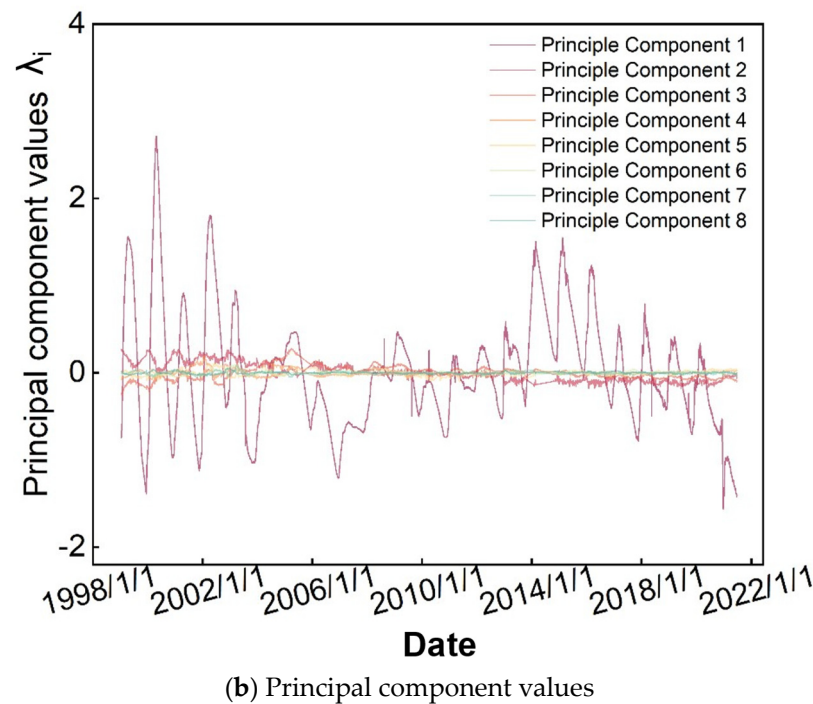


Figure 6. Changes in monitoring seepage q_s and various principal component values or eigenvalues λ_i from 1999 to 2021.

4.2. Dam Seepage Model Based on SVR

The group of penalty coefficients c and kernel function parameters g with the highest R^2 value and smallest value of c were chosen as the training parameters for SVR by using the grid search technique. Figure 7 displays the calculation results of the K-fold cross-validation (K-CV) with $K = 5$ and five eigenvectors as inputs. Detailed information on selecting K through hyperparameter optimization can be found in Appendix A.2. The input data comprised $n = 658$ samples between 1999 and 2021. As R^2 is maximum when $c = 10$ and $g = 0.1$, these were selected as the optimal parameters to evaluate the SVR model. Using the same approach, the optimal parameter sets (c and g) obtained using different numbers of eigenvectors (between three and eight) are listed in Table 2. Kernel function parameter g decreased as the number of inputted eigenvectors increased. Thus, by increasing the dimensions of the input data or the number of eigenvectors, smaller g values are required to fit the training results. Additionally, the various times required to conduct the SVR model with different numbers of input eigenvectors were measured, as indicated in Table 2. These times include the time of file I/O, i.e., the time required to read and write files (T_{io}), the time of training and validation (T_t), and the prediction time (T_p).

Table 2. Training, validation, and testing of SVR model using different numbers of inputted eigenvectors.

Number of Inputted Eigenvectors k	c	g	Training and Validation Time T_t (s)	File I/O Time T_{io} (s)	R^2 (Testing)	E_{ms} (L/s) ² (Testing)	E_{ma} (Testing)
3	10	0.1	9.843	0.910	0.945	0.045	17.19%
4	10	0.1	10.024	1.005	0.933	0.096	25.11%
5	10	0.1	11.031	1.103	0.953	0.041	16.41%
6	100	0.01	15.123	1.399	0.941	0.056	19.18%
7	100	0.01	16.255	1.309	0.948	0.035	15.16%
8	100	0.01	20.518	1.456	0.986	0.041	16.41%

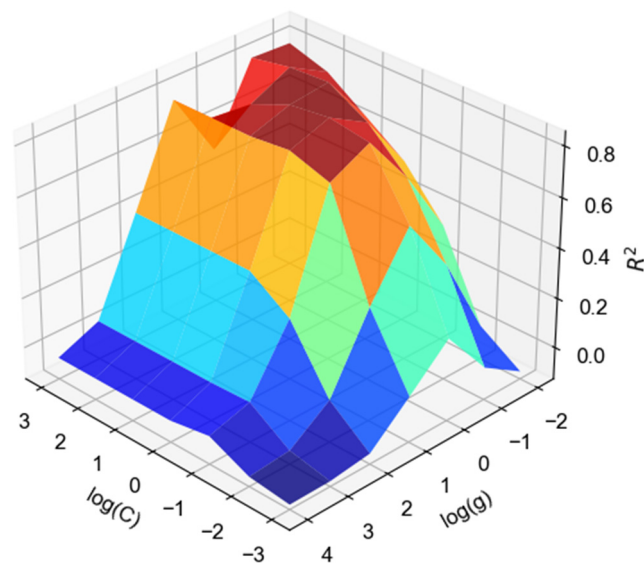


Figure 7. Training of SVR model via K-fold cross-validation (K-CV) when $K = 5$ and five eigenvectors were selected. The R^2 value is maximum when $\log(c) = 1$ and $\log(g) = -1$; i.e., $c = 10$ and $g = 0.1$.

As shown in Figure 8, T_{io} and T_t have a gradually increasing trend as the number of inputted eigenvectors k increases. For example, T_t with three eigenvectors required 9.843 s, which was only half the time required for eight eigenvectors (20.518 s). However, T_p does not increase with the increase in k . T_p is the shortest between $k = 4$ and 5. Figure 9 shows the relationship between the determination coefficient R^2 and the number of inputted eigenvectors k . R^2 exhibits high fluctuations within the range of $k = 3$ to 7. When seven eigenvectors were inputted, the accuracy of the model did not increase significantly, probably because inputting more eigenvectors did not provide more effective information.

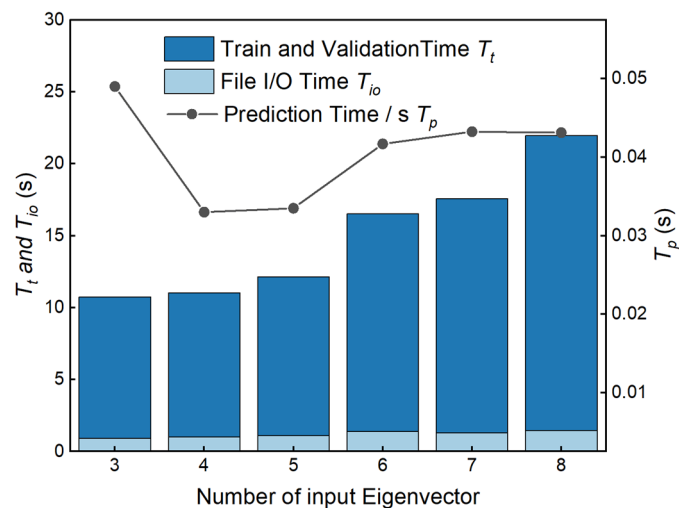


Figure 8. Time of training and validation (T_t), time of file I/O (T_{io}), and prediction time (T_p) with different numbers of eigenvectors in the SVR model.

Based on the process described in Section 3.3 and the analysis of the relevant figures and tables mentioned above (Figures 8 and 9 and Table 2), the results showed that inputting five eigenvectors can achieve high accuracy while ensuring a shorter calculation time. Therefore, we selected five as the optimal number for the main components and used this as the basis for subsequent analysis.

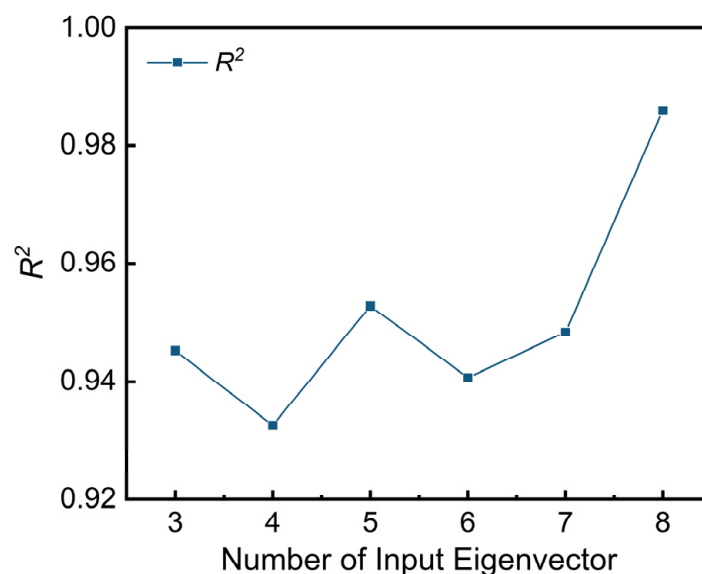


Figure 9. Determination coefficient R^2 with different numbers of eigenvectors in the SVR model.

4.3. Comparison with Other Data Processing Methods

Maximal information coefficient (MIC) is a statistical method that quantifies the degree of association or correlation between two sets of data points without making any assumptions about the underlying relationship. MIC is particularly useful for identifying relationships in complex and high-dimensional datasets where traditional linear correlation measures may fail to capture the underlying structure. Therefore, three commonly used data processing techniques, i.e., GRA, PCA, and MIC, were introduced for comparison with the HPO, specifically, for the combination of GRA and PCA proposed in this study. As shown in Figure 10a–c, the GRA and HPO methods present higher R^2 values and lower E_{ms} and E_{ma} values. The HPO and GRA methods produce good results compared with PCA and MIC. Thus, GRA can extract the nonlinear relationship with the seepage flow better than PCA or MIC. However, the GRA method cannot effectively reduce input factors that have the same contribution to seepage prediction. Redundant input factors increase the training time of the model and waste computing resources, such as storage and energy consumption, which are unfriendly to edge devices. As shown in Figure 10d, training and validation time T_t of the proposed model (HPO method) in this study is 11.03 s, which is only one-third of that of the GRA method, with $T_t = 33.60$ s. Further, compared with the GRA model, the HPO method can perform well in the real-time monitoring of dam safety. Overall, the above results indicate that the proposed HPO method can effectively reduce the dimensionality of the input factors without losing important information from the SVR prediction model.

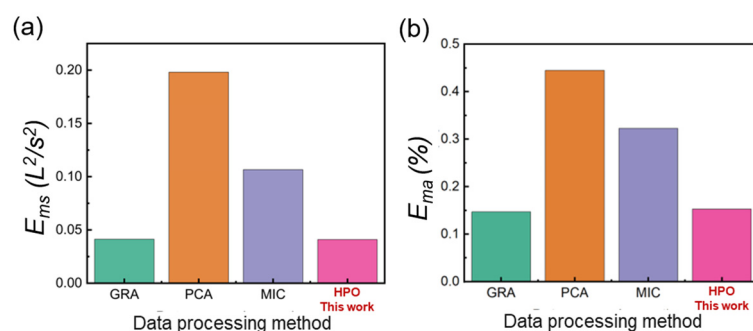


Figure 10. Cont.

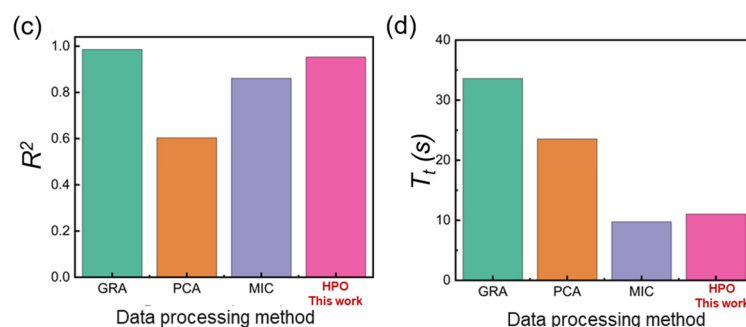


Figure 10. (a) Mean squared error E_{ms} , (b) mean absolute percentage error E_{ma} , (c) determination coefficient R^2 , and (d) time for training and validation T_t ; evaluation among GRA, PCA, MIC, and HPO (the model proposed in this study) methods.

4.4. Comparison of Models

The LSTM model is a type of artificial neural network model widely used in previous studies [32]. We considered the seepage monitoring data collected at the Shuidong Dam from 1999 to 2021 to compare the seepage prediction values of the proposed model in this study (HPO-SVR), a single LSTM model, and the LSTM with HPO (HPO-LSTM) model, as shown in Figures 11 and 12. Comparing the three models is to understand the execution performance between SVM and LSTM, as well as to examine the effectiveness of HPO in data processing. As shown in Figure 11, the light-green area in the figure represents the period of dam reinforcement, and no monitoring data are available. A large seepage flow was found from December to January each year before dam reinforcement; however, the seepage flow was less than 1 L/s throughout the year after dam reinforcement. The burst point of the seepage flow occurred randomly, which is different from that of the dam before reinforcement. Both models could predict the sudden increase in seepage in 2010. Clearly, the seepage prediction after dam reinforcement by LSTM is significantly higher than that of seepage monitoring. Figure 13 depicts the prediction of seepage flow against the monitoring of seepage flow for the proposed HPO-SVR and LSTM models. The values predicted using our model are more consistent with the monitoring values (closer to the 45° line) for both the larger and smaller seepage (pre- and post-dam reinforcements) cases, whereas the predicted seepage values of the LSTM model are high when the monitoring seepage was smaller than approximately 0.5 L/s and low when the monitoring seepage was larger than approximately 1 L/s.

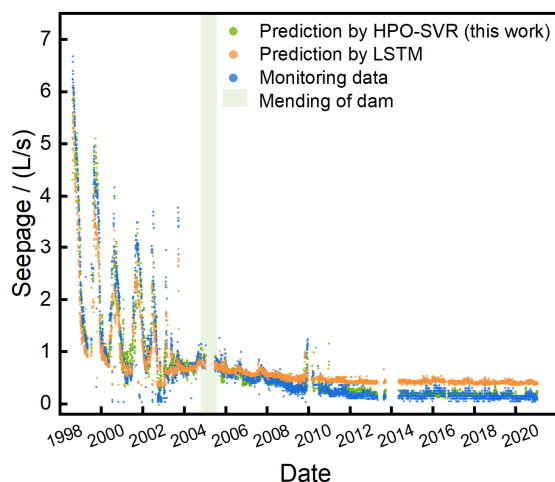


Figure 11. Comparison of seepage monitoring of the Shuidong Dam between the model proposed in this study (SVR model with mixed parameter optimization, shown with the orange circles) and the LSTM model (blue circles) for seepage prediction. The light-green colored area in the figure represents the period of dam repair, for which no monitoring data are available.

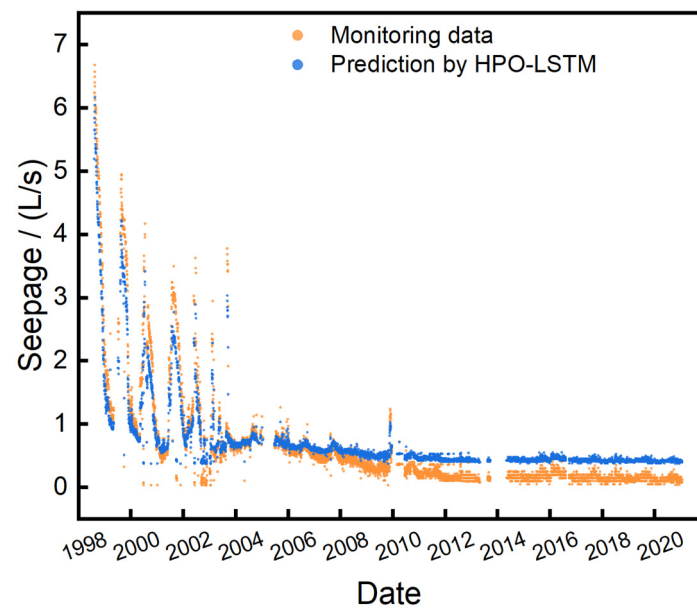


Figure 12. Comparison of seepage value of the Shuidong Dam between the monitoring data and the LSTM model with HPO methods when inputting one eigenvector.

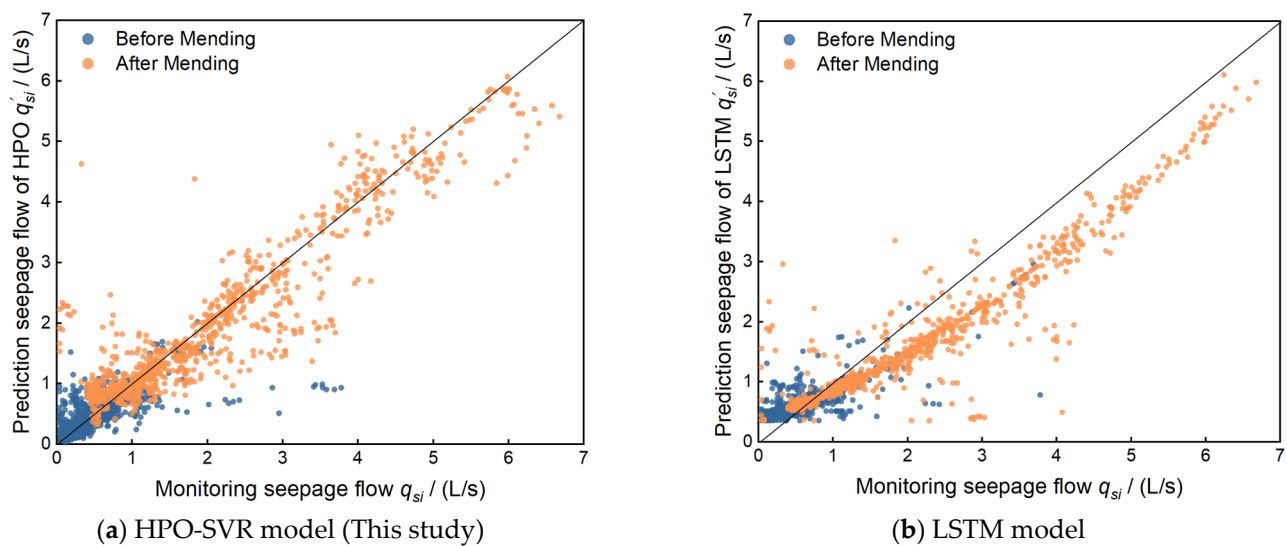


Figure 13. Prediction seepage flow q'_{si} against monitoring seepage flow q_{si} for (a) proposed HPO-SVR model and (b) LSTM model. In the overall process, covering the period before and after dam reinforcements, the proposed HPO model in this study provides a better prediction of the seepage from the Shuidong Dam than the LSTM model.

The R^2 , E_{ms} , and E_{ma} values of our proposed model (HPO-SVR), the LSTM model, and the HPO-LSTM model were calculated and listed in Table 3. The accuracy and error of the LSTM model in predicting the seepage before dam reinforcement are slightly better (with slightly higher R^2 values and slightly lower E_{ms} and E_{ma} values) than those of the model proposed in this study. However, the model proposed in this study exhibits a higher accuracy, with $R^2 = 0.9558$, and a lower E_{ma} , of 36.49%, for the prediction of seepage flow after dam reinforcement. Further, in terms of the overall process, covering the period before and after dam reinforcements, the model proposed in this study, with $R^2 = 0.9407$ and $E_{ma} = 22.21\%$, provides a better prediction of the seepage from the Shuidong Dam than the LSTM model, with $R^2 = 0.9173$ and $E_{ma} = 27.24\%$. This means that HPO-SVR can predict well under both pre and post reinforcement seepage models, while LSTM only has slightly higher accuracy under the pre-reinforcement seepage model, indicating that

HPO-SVR has high generalizability. As indicated in Table 3, the HPO-LSTM model has a higher R^2 and lower E_{ms} and E_{ma} values compared with the LSTM model. HPO was able to bring higher accuracy to LSTM. However, the HPO-LSTM model, with $R^2 = 0.9256$ and $E_{ma} = 22.57\%$, is still not as accurate as HPO-SVR, with $R^2 = 0.9407$ and $E_{ma} = 22.21\%$. Based on the results of Table 3, SVR is a more effective MLA for solving dam seepage problems, and HPO can be applied to other MLAs to improve their prediction accuracy. We also examined the training and validation time T_t for various models (HPO-SVR, LSTM, and HPO-LSTM), as indicated in Table 4. The training and validation time of the HPO-SVR model, with $T_t = 11$ s, is the shortest among the three models. The T_t of the HPO-LSTM model is 87 s, while the T_t of the LSTM model without importing HPO for data processing is 146 s. The time ratio of the two models (LSTM vs. HPO-LSTM) is 1.7 times. This shows that combining HPO with machine learning algorithms has high timeliness and indicates that HPO proposed in this study has high generalizability. In this case study, HPO-SVR can be used to predict seepage faster, using fewer computational and storage resources, and is particularly suitable for dam monitoring systems with outdated equipment. For example, among all the monitoring data collected in this study, many from outdated and unmaintained thermometers can also be successfully filtered by HPO. This can save the time of this model in seepage prediction.

Table 3. The comparison of performance indicators (determination coefficient R^2 , mean square error E_{ms} , and mean absolute percentage error E_{ma}) between the model proposed in this study (HPO-SVR), LSTM model, and LSTM model with HPO (HPO-LSTM).

Performance Indicators	Statement	HPO-SVR (This Work)	LSTM	HPO-LSTM
R^2	dam before reinforcement	0.9323	0.9628	0.9581
	dam after reinforcement	0.9558	0.8867	0.8912
	Total period *	0.9407	0.9173	0.9256
E_{ms} (L/s) ²	dam before reinforcement	0.5711	0.4656	0.3422
	dam after reinforcement	0.0421	0.0878	0.0573
	Total period *	0.0751	0.1130	0.0776
E_{ma}	dam before reinforcement	21.52%	19.43%	16.66%
	dam after reinforcement	36.49%	52.70%	42.57%
	Total period *	22.21%	27.24%	22.57%

Note(s): * Period between 1999 and 2021, covering the dam before and after reinforcements.

Table 4. Training and validation time T_t (s) for various models (HPO-SVR, LSTM, and HPO-LSTM).

Model	HPO-SVR (This Work)	LSTM	HPO-LSTM
T_t (s)	11	146	87

In this case study, SVR was performed using uplift pressure and temperature factors related to seepage (i.e., C_{l1} , C_{l2} , C_{l3} , C_{l4} , C_{l6} , C_{l11} , C_{l12} , and T_4). These factors were obtained after data processing of HPO; they are considered important monitoring factors and require appropriate management and maintenance. When these daily monitoring factors are provided, the SVM model proposed in this study can estimate or predict the seepage flowrate of the Shuidong Dam, providing reference for management units.

5. Conclusions and Recommendations

5.1. Conclusions

This study proposed a seepage prediction model (HPO-SVR) based on SVR in machine learning algorithms with data processing of HPO and conducted research focusing on the Shuidong Dam in China as an example. The HPO combines GRA to filter out meaningless monitoring factors and uses PCA to reduce the calculation time of the model.

The data processing results showed that factors highly correlated with the seepage flow from the dam include uplift pressure, vertical and horizontal displacements, and air temperature. Most effects of these monitored factors were apparent on the right bank of the dam, implying that seepage on the right bank is more significant than on the left bank. This finding is consistent with on-site investigations, thus demonstrating the effectiveness of the proposed data processing model. The proposed HPO method was used to compare with other commonly used methods (i.e., GRA, PCA, and MIC) and showed that HPO could reduce the training, validation, and testing time threefold while maintaining high accuracy. We also considered the seepage monitoring data collected at Shuidong Dam from 1999 to 2021 to compare the seepage prediction values of HPO-SVR, LSTM, and HPO-LSTM models. The results show that the HPO-SVR model ($R^2 = 0.9407$ and $E_{ma} = 22.21\%$) provides better seepage prediction performance than the LSTM model ($R^2 = 0.9173$ and $E_{ma} = 27.24\%$) and the HPO-LSTM model ($R^2 = 0.9256$ and $E_{ma} = 22.57\%$). In addition, HPO-SVR has the shortest training and validation time among the three models. More specifically, the proposed SVR model with HPO can effectively extract the nonlinear relationship between input factors and seepage, demonstrating good accuracy in the seepage prediction of the Shuidong Dam.

5.2. Recommendations

This study takes all monitoring data of the Shuidong Dam as an example, and these monitoring data have not been preprocessed, assuming that all monitoring data are valid. After being processed by the HPO, these data can effectively reduce or reduce the input factors of SVR, thereby predicting the seepage flowrate of the dam. However, it should be noted that SVR relies on the selection of input factors; too many input factors will increase training difficulty or cause overfitting, and selecting fewer input factors leads to low model accuracy [64]. Due to the differences in dam body, foundation, and surrounding environment in different case locations, the monitoring parameters used may not be the same. Therefore, the seepage flow prediction model proposed in this study for RCC dams may not be applicable to other dams. However, the proposed HPO-SVR model in this study can serve as an important reference for other RCC dams in data processing (selecting reasonable monitoring parameters or monitoring points) and establishing seepage flow prediction models. After all, compared to other evaluation models, the model proposed in this article has the advantages of simplicity, efficiency, and high accuracy overall.

One of the important contributions of this study is the use of HPO to process all monitoring data in the case. Previous studies have rarely combined mechanical learning with HPO to predict the seepage of RCC dams. Therefore, this study developed a prediction model for seepage using a simple and commonly used machine learning model, SVR combined with HPO, and preliminarily compared it with LSTM in neural networks. However, there are many methods for mechanical learning, and further research is needed to combine HPO with other advanced neural network models such as BPNN. Generally speaking, more advanced mechanical learning models may improve accuracy but typically consume more training and validating time. With the advancement of cloud computing technology, the training and inference of machine learning models will be migrated to the cloud. This shift will reduce reliance on local computing resources. Therefore, future research can consider using more complex MLA combined with the HPO method to improve the accuracy of the model and reduce computational resource consumption. In addition, the impact of the frequency of dam monitoring data on the prediction model can also be explored. Choosing an appropriate monitoring frequency can reduce the burden on local information storage resources and ensure the accuracy of predictions.

Author Contributions: Methodology, M.-Y.Z. and J.-C.C.; resources, R.-L.Z. and Y.-K.Z.; software, M.-Y.Z.; investigation, R.-L.Z., Y.-K.Z., M.-Y.Z., J.-C.C. and W.-S.H.; visualization, M.-Y.Z. and J.-C.C.; writing—original draft, M.-Y.Z. and J.-C.C.; writing—review and editing, J.-C.C.; supervision, J.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Fund of Fujian College of Water Conservancy and Electric Power, Grant Number YJRCKYQD2101.

Data Availability Statement: The data supporting plots within this paper and other findings of this study are available from the first author upon reasonable request.

Acknowledgments: We appreciate assistance of J. Zheng in software and graphics processing. We also thank the editor and three anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Appendix A.1. Input Factors x_j of GRA

Table A1. Factors x_j corresponding to each number j in Figure 4.

Number j	Factors x_j
1	C_{111} : Uplift pressure coefficient at location UP11
2	C_{11} : Uplift pressure coefficient at location UP01
3	D_{V2} : Accumulated vertical displacement at location L2 (mm)
4	C_{14} : Uplift pressure coefficient at location UP04
5	C_{16} : Uplift pressure coefficient at location UP06
6	C_{112} : Uplift pressure coefficient at location UP12
7	D_{H3} : Accumulated horizontal displacement at location I3 (mm)
8	C_{12} : Uplift pressure coefficient at location UP02
9	D_{V7} : Accumulated vertical displacement at location L7 (mm)
10	D_{H1} : Accumulated horizontal displacement at location I1 (mm)
11	C_{13} : Uplift pressure coefficient at location UP03
12	D_{V5} : Accumulated vertical displacement at location L5 (mm)
13	T_4 : Air temperature at location UP04 (°C)
14	D_{H2} : Accumulated horizontal displacement at location I2 (mm)
15	D_{V3} : Accumulated vertical displacement at location L3 (mm)
16	C_{110} : Uplift pressure coefficient at location UP10
17	h_D : Water level difference between upstream and downstream water levels, $h_D = h_u - h_d$ (m)
18	T_8 : Air temperature at location UP08 (°C)
19	h_u : Upstream water level (m)
20	E_{l8} : Elevation at location L8 (m)
21	E_{l4} : Elevation at location L4 (m)
22	t_d : Date in day
23	D_{V4} : Accumulated vertical displacement at location L4 (mm)
24	E_{l6} : Elevation at location L6 (m)
25	t_m : Date in month
26	t_q : Date in quarter
27	d_{H2} : Interval horizontal displacement at location I2 (mm)
28	C_{15} : Uplift pressure coefficient at location UP05
29	D_{V6} : Accumulated vertical displacement at location L6 (mm)

Table A1. Cont.

Number j	Factors x_j
30	d_{H1} : Interval horizontal displacement at location I1 (mm)
31	d_{H3} : Interval horizontal displacement at location I3 (mm)
32	d_{V3} : Interval vertical displacement at location L3 (mm)
33	d_{V2} : Interval vertical displacement at location L2 (mm)
34	d_{V7} : Interval vertical displacement at location L7 (mm)
35	C_{I9} : Uplift pressure coefficient at location UP09
36	d_{V6} : Interval vertical displacement at location L6 (mm)
37	C_{I8} : Uplift pressure coefficient at location UP08
38	d_{V5} : Interval vertical displacement at location L5 (mm)
39	D_{V8} : Accumulated vertical displacement at location L8 (mm)
40	T_2 : Air temperature at location UP02 (°C)
41	d_{V8} : Interval vertical displacement at location L8 (mm)
42	T_5 : Air temperature at location UP05 (°C)
43	T_3 : Air temperature at location UP03 (°C)
44	T_{10} : Air temperature at location UP10 (°C)
45	d_{V4} : Interval vertical displacement at location L4 (mm)
46	T_7 : Air temperature at location UP07 (°C)
47	T_{13} : Air temperature at location UP13 (°C)
48	C_{I13} : Uplift pressure coefficient at location UP13
49	T_1 : Air temperature at location UP01 (°C)
50	T_{11} : Air temperature at location UP11 (°C)
51	T_{12} : Air temperature at location UP12 (°C)
52	T_6 : Air temperature at location UP06 (°C)
53	T_9 : Air temperature at location UP09 (°C)
54	C_{I7} : Uplift pressure coefficient at location UP07
55	h_d : Downstream water level (m)
56	E_{I3} : Elevation at location L3 (m)
57	E_{I7} : Elevation at location L7 (m)
58	E_{I2} : Elevation at location L2 (m)
59	E_{I5} : Elevation at location L5 (m)
60	t_y : Date in year

Appendix A.2. Hyperparameter Optimization

K-fold cross-validation is a rigorous model evaluation technique where the dataset is divided into K equal-sized subsets or folds. The model has undergone frequent training and testing, with each fold serving as the validation set once and the remaining K-1 folds used for training. This process helps to evaluate the performance of the model on different subsets of data, reducing the risk of bias or overfitting. We tested the performance of K-CV at $K = 2$ to 10. For each case of K, we repeated the test 10 times to obtain the optimal score of R^2 (training R^2) and the average predicted R^2 (prediction R^2), as shown in Table A2 and Figure A1. The selection of K needs to strike a balance between the accuracy of model evaluation and computational costs. A smaller K may be faster, but the evaluation may not be stable enough, while a larger K may be more stable but require more computing

resources. If the total number of datasets is not large enough, an excessively large K will result in a small amount of data in each group, once again leading to unstable results. We found that K values of two to three resulted in higher training R^2 , but the prediction R^2 seemed to decrease; this indicates the instability of the results caused by the K value being too small. When the value of K is greater than five, the larger number of subsets, i.e., fewer samples in each subset, which does not include all features in each subset, resulted in a decrease in accuracy. Thus, $K = 5$ is an appropriate value and was adopted in this study.

Table A2. Training R^2 and prediction R^2 at various values of K.

K	Training R^2	Prediction R^2
2	0.9187	0.8931
3	0.9088	0.9077
4	0.9065	0.9211
5	0.9078	0.9378
6	0.8993	0.9122
7	0.8913	0.8999
8	0.9021	0.9080
10	0.8773	0.8957

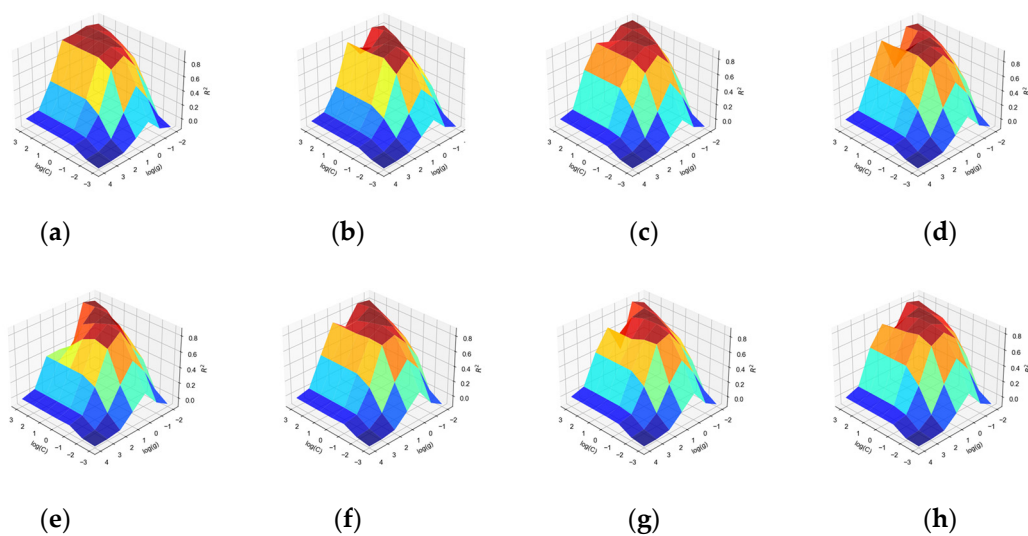


Figure A1. Training of SVR model via K-fold cross-validation (K-CV) when $K = 2, 3, 4, 5, 6, 7, 8$, and 10 (shown in order in (a–h)) and five eigenvectors were selected.

References

1. Brigandì, G.; Candela, A.; Aronica, G.T. Analysis of the Effects of Reservoir Operating Scenarios on Downstream Flood Damage Risk Using an Integrated Monte Carlo Modelling Approach. *Water* **2023**, *15*, 550.
2. Fang, C.; Duan, Y. Statistical Analysis of Dam-Break Incidents and Its Cautions. *Yangtze River* **2010**, *41*, 96–101. (in Chinese) [[CrossRef](#)]
3. Terzaghi, K. *Theoretical Soil Mechanics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1943. [[CrossRef](#)]
4. Chai, J. On Mathematical Model for Coupled Seepage and Temperature Field in Concrete Dam. *Chin. J. Hydroelectr. Power* **2000**, *1*, 27–35.
5. Wu, Z.; Song, H. Study on Shallow Geothermal Field and Seepage Field Coupling Based on Lu Model. *J. Hydraul. Eng.* **2015**, *46*, 326–333.
6. Cao, B.; Yang, J.; Chen, L.; Zhang, A.; Mao, H. Finite Element Simulation of Seepage Thermal Monitoring of Earth-Rock Dam Based on COMSOL. In Proceedings of the Third International Conference on Optoelectronic Science and Materials (ICOSM 2021), Hefei, China, 10–12 September 2021; Chen, S., Wang, P., Eds.; SPIE: Bellingham, WA, USA, 2021; p. 38. [[CrossRef](#)]

7. Chen, S.; Gu, C.; Lin, C.; Zhao, E.; Song, J. Safety monitoring model of a super-high concrete dam by using RBF neural network coupled with kernel principal component analysis. *Math. Probl. Eng.* **2018**, *2018*, 1712653.
8. Yu, Y.; Liu, X.; Wang, E.; Fang, K.; Huang, L. Dam safety evaluation based on multiple linear regression and numerical simulation. *Rock Mech. Rock Eng.* **2018**, *51*, 2451–2467.
9. Tatin, M.; Briffaut, M.; Dufour, F.; Simon, A.; Fabre, J.P. Thermal displacements of concrete dams: Accounting for water temperature in statistical models. *Eng. Struct.* **2015**, *91*, 26–39. [\[CrossRef\]](#)
10. Milillo, P.; Perissin, D.; Salzer, J.T.; Lundgren, P.; Lacava, G.; Milillo, G.; Serio, C. Monitoring dam structural health from space: Insights from novel InSAR techniques and multi-parametric modeling applied to the Pertusillo dam Basilicata, Italy. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 221–229. [\[CrossRef\]](#)
11. Ren, Q.; Li, M.; Song, L.; Liu, H. An optimized combination prediction model for concrete dam deformation considering quantitative evaluation and hysteresis correction. *Adv. Eng. Inform.* **2020**, *46*, 101154.
12. Jin, X.-B.; Wang, Z.-Y.; Kong, J.-L.; Bai, Y.-T.; Su, T.-L.; Ma, H.-J.; Chakrabarti, P. Deep Spatio-Temporal Graph Network with Self-Optimization for Air Quality Prediction. *Entropy* **2023**, *25*, 247. [\[CrossRef\]](#)
13. Jiang, W.; Zhu, G.; Shen, Y.; Xie, Q.; Ji, M.; Yu, Y. An Empirical Mode Decomposition Fuzzy Forecast Model for Air Quality. *Entropy* **2022**, *24*, 1803. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Lang, Z.; Wen, Q.H.; Yu, B.; Sang, L.; Wang, Y. Forecast of Winter Precipitation Type Based on Machine Learning Method. *Entropy* **2023**, *25*, 138. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Sun, A.Y.; Wang, D.; Xu, X. Monthly Streamflow Forecasting Using Gaussian Process Regression. *J. Hydrol.* **2014**, *511*, 72–81. [\[CrossRef\]](#)
16. Campolo, M.; Soldati, A.; Andreussi, P. Artificial Neural Network Approach to Flood Forecasting in the River Arno. *Hydrol. Sci. J.* **2003**, *48*, 381–398. [\[CrossRef\]](#)
17. El Bilali, A.; Moukhli, M.; Taleb, A.; Nafii, A.; Alabjah, B.; Brouziyne, Y.; Mazigh, N.; Tezine, K.; Mhamed, M. Predicting daily pore water pressure in embankment dam: Empowering machine learning-based modeling. *Environ. Sci. Pollut. Res.* **2022**, *29*, 47382–47398.
18. Xiang, C.; Li, Q.; Zhou, Z.; Luo, Z.; Liu, M.; Liu, L. Research on a seepage monitoring model of a high core rockfill dam based on machine learning. *Sensors* **2018**, *18*, 2749.
19. Lin, C.; Li, T.; Chen, S.; Liu, X.; Lin, C.; Liang, S. Gaussian process regression-based forecasting model of dam deformation. *Neural Comput. Appl.* **2019**, *31*, 8503–8515.
20. Wei, B.; Yuan, D.; Xu, Z.; Lin, L. Modified hybrid forecast model considering chaotic residual errors for dam deformation. *Struct. Control. Health Monit.* **2018**, *25*, e2188. [\[CrossRef\]](#)
21. Wei, B.; Gu, M.; Li, H.; Xiong, W.; Xu, Z. Modeling method for predicting seepage of RCC dams considering time-varying and lag effect. *Struct. Control. Health Monit.* **2018**, *25*, e2081.
22. Wang, S.W.; Xu, Y.L.; Gu, C.S.; Bao, T.F. Monitoring models for base flow effect and daily variation of dam seepage elements considering time lag effect. *Water Sci. Eng.* **2018**, *11*, 344–354.
23. Roushangar, K.; Garekhani, S.; Alizadeh, F. Forecasting Daily Seepage Discharge of an Earth Dam Using Wavelet–Mutual Information–Gaussian Process Regression Approaches. *Geotech. Geol. Eng.* **2016**, *34*, 1313–1326.
24. Ranković, V.; Grujović, N.; Divac, D.; Milivojević, N. Development of support vector regression identification model for prediction of dam structural behavior. *Struct. Saf.* **2014**, *48*, 33–39. [\[CrossRef\]](#)
25. Su, H.; Chen, Z.; Wen, Z. Performance improvement method of support vector machine-based model monitoring dam safety. *Struct. Control Health Monit.* **2016**, *23*, 252–266. [\[CrossRef\]](#)
26. Su, H.; Li, X.; Yang, B.; Wen, Z. Wavelet support vector machine-based prediction model of dam deformation. *Mech. Syst. Sig. Process.* **2018**, *110*, 412–427.
27. Sharghi, E.; Nourani, V.; Behfar, N.; Tayfur, G. Data Pre-Post Processing Methods in AI-Based Modeling of Seepage through Earthen Dams. *Measurement* **2019**, *147*, 106820. [\[CrossRef\]](#)
28. Kang, F.; Li, J.; Dai, J. Prediction of long-term temperature effect in structural health monitoring of concrete dams using support vector machines with Jaya optimizer and salp swarm algorithms. *Adv. Eng. Softw.* **2019**, *131*, 60–76. [\[CrossRef\]](#)
29. Stojanovic, B.; Milivojevic, M.; Milivojevic, N.; Antonijevic, D. A self-tuning system for dam behavior modeling based on evolving artificial neural networks. *Adv. Eng. Softw.* **2016**, *97*, 85–95. [\[CrossRef\]](#)
30. Kao, C.Y.; Loh, C.H. Monitoring of long-term static deformation data of Fei-Tsui arch dam using artificial neural network-based approaches. *Struct. Control Health Monit.* **2013**, *20*, 282–303.
31. Zhang, X.; Chen, X.; Li, J. Improving Dam Seepage Prediction Using Back-Propagation Neural Network and Genetic Algorithm. *Math. Probl. Eng.* **2020**, *2020*, 1404295. [\[CrossRef\]](#)
32. Zhang, J.; Li, W.; Hu, B.; Yang, H.; Wang, H. Design of an LSTM model for dam leakage prediction. In Proceedings of the Fifth International Conference on Mechatronics and Computer Technology Engineering, Chongqing, China, 19–21 August 2022. [\[CrossRef\]](#)
33. Kang, F.; Liu, J.; Li, J.; Li, S. Concrete dam deformation prediction model for health monitoring based on extreme learning machine. *Struct. Control Health Monit.* **2017**, *24*, e1997. [\[CrossRef\]](#)

34. Bui, K.T.T.B.; Bui, D.T.; Zou, J.; Doan, C.; Revhaug, I. A novel hybrid artificial intelligent approach based on neural fuzzy inference model and particle swarm optimization for horizontal displacement modeling of hydropower dam. *Neural Comput.* **2018**, *29*, 1495–1506.
35. Kang, F.; Li, J.; Zhao, S.; Wang, Y. Structural health monitoring of concrete dams using long-term air temperature for thermal effect simulation. *Eng. Struct.* **2019**, *180*, 642–653.
36. Li, X.; Wen, Z.; Su, H. An approach using random forest intelligent algorithm to construct a monitoring model for dam safety. *Eng. Comput.* **2019**, *37*, 39–56.
37. Salazar, F.; Toledo, M.Á.; González, J.M.; Oñate, E. Early detection of anomalies in dam performance: A methodology based on boosted regression trees. *Struct. Control Health Monit.* **2017**, *24*, e2012. [\[CrossRef\]](#)
38. Zhang, K.; Gu, C.; Zhu, Y.; Chen, S.; Dai, B.; Li, Y.; Shu, X. A novel seepage behavior prediction and lag process identification method for concrete dams using HGWO-XGBoost Model. *IEEE Access* **2021**, *9*, 23311–23325. [\[CrossRef\]](#)
39. Song, L.; Hao, L.; Tao, H.; Xu, C.; Guo, R.; Li, Y.; Yao, J. Research on Black-Box Modeling Prediction of USV Maneuvering Based on SSA-WLS-SVM. *J. Mar. Sci. Eng.* **2023**, *11*, 324. [\[CrossRef\]](#)
40. Gu, C.S.; Wei, B.W.; Xu, Z.K.; Liu, D.W. Fluid-solid coupling model based on endochronic damage for roller compacted concrete dam. *J. Cent. S. Univ.* **2013**, *20*, 3247–3255. (in Chinese).
41. Aburomman, A.A.; Reaz, M.B.I. A Novel Weighted Support Vector Machines Multiclass Classifier Based on Differential Evolution for Intrusion Detection Systems. *Inf. Sci.* **2017**, *414*, 225–246. [\[CrossRef\]](#)
42. Yang, X.; Yu, Q.; He, L.; Guo, T. The One-against-All Partition Based Binary Tree Support Vector Machine Algorithms for Multi-Class Classification. *Neurocomputing* **2013**, *113*, 1–7. [\[CrossRef\]](#)
43. Lu, H.S.; Chang, C.K.; Hwang, N.C.; Chung, C.T. Grey Relational Analysis Coupled with Principal Component Analysis for Optimization Design of the Cutting Parameters in High-Speed End Milling. *J. Mater. Process. Technol.* **2009**, *209*, 3808–3817. [\[CrossRef\]](#)
44. Huang, Y.; Shen, L.; Liu, H. Grey Relational Analysis, Principal Component Analysis and Forecasting of Carbon Emissions Based on Long Short-Term Memory in China. *J. Clean. Prod.* **2019**, *209*, 415–423. [\[CrossRef\]](#)
45. Chen, B.P. The Application of the Grey Correlation Method in the Principal Component Analysis. In *Advanced Engineering Forum*; Trans Tech Publications Ltd.: Zurich, Switzerland, 2012; Volume 6, pp. 676–681.
46. Deng, J.L. Introduction to Grey System Theory. *J. Grey Syst.* **1989**, *1*, 1–24.
47. Wang, Z.; Zhang, H.; Wang, Y.; Wang, Y.; Lei, L.; Liang, C.; Wang, Y. Deficit Irrigation Decision-Making of Indigowoad Root Based on a Model Coupling Fuzzy Theory and Grey Relational Analysis. *Agric. Water Manag.* **2023**, *275*, 107983. [\[CrossRef\]](#)
48. Yuan, D.; Jang, G. Coupling Coordination Relationship between Tourism Industry and Ecological Civilization: A Case Study of Guangdong Province in China. *Sustainability* **2023**, *15*, 92. [\[CrossRef\]](#)
49. Cadima, J.F.C.L.; Jolliffe, I.T. Size- and Shape-Related Principal Component Analysis. *Biometrics* **1996**, *52*, 2710–2716. [\[CrossRef\]](#)
50. Ding, K.; Zeng, Y.; Wang, Y.; Lv, D.; Yan, X. AGIM-Net Based Subject-Sensitive Hashing Algorithm for Integrity Authentication of HRRS Images. *Geocarto Int.* **2023**, *38*, 2168071. [\[CrossRef\]](#)
51. Li, J.; Liu, X.; Yao, Q.; Xu, L.; Li, W.; Tan, W.; Wang, Q.; Xing, W.; Liu, D. Tolerance and Adaptation Characteristics of Sugar Beet (*Beta Vulgaris* L.) to Low Nitrogen Supply. *Plant Signal. Behav.* **2022**, *18*, 2159155. [\[CrossRef\]](#)
52. Fantahun, B.; Woldeesemayate, T.; Fadda, C.; Gebrehawaryat, Y.; Pe, E.; Acqua, M.D. Multivariate Analysis in the Dissection of Phenotypic Variation of Ethiopian Cultivated Barley (*Hordeum Vulgare* ssp. *Vulgare* L.) Genotypes. *Cogent Food Agric.* **2023**, *9*, 2157104. [\[CrossRef\]](#)
53. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. [\[CrossRef\]](#)
54. Lin, K.; Lin, Q.; Zhou, C.; Yao, J. Time Series Prediction Based on Linear Regression and SVR. In Proceedings of the Third International Conference on Natural Computation, Haikou, China, 24–27 August 2007; pp. 688–691.
55. Guo, Y.; Han, S.; Shen, C.; Li, Y.; Yin, X.; Bai, Y. An Adaptive SVR for High-Frequency Stock Price Forecasting. *IEEE Access* **2018**, *6*, 11397–11404. [\[CrossRef\]](#)
56. Jiang, S.; Xue, H.; Glover, A.; Rutherford, M.; Rueckert, D.; Hajnal, J.V. MRI of Moving Subjects Using Multislice Snapshot Images with Volume Reconstruction (SVR): Application to Fetal, Neonatal, and Adult Brain Studies. *IEEE Trans. Med. Imaging* **2007**, *26*, 967–980. [\[CrossRef\]](#) [\[PubMed\]](#)
57. He, G.; Cai, G.; Li, Y.; Xia, T.; Li, Z. Weighted Split-Flow Network Auxiliary with Hierarchical Multitasking for Urban Land Use Classification of High-Resolution Remote Sensing Images. *Int. J. Remote Sens.* **2022**, *43*, 6721–6740. [\[CrossRef\]](#)
58. Yao, X. Application of Optimized SVM in Sample. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 540–547.
59. Yigit, E.; Duysak, H. Determination of Flowing Grain Moisture Contents by Machine Learning Algorithms Using Free Space Measurement Data. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2507608. [\[CrossRef\]](#)
60. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
61. Yao, H.; Liu, D. Study on Seepage Monitoring and Analysis of SL Gravity Dam. In Proceedings of the 2021 7th International Conference on Hydraulic and Civil Engineering & Smart Water Conservancy and Intelligent Disaster Reduction Forum (ICHCE & SWIDR), Nanjing, China, 6–8 November 2021; pp. 1475–1478. [\[CrossRef\]](#)
62. Simon, A.; Collison, A.J.C. Pore-Water Pressure Effects on the Detachment of Cohesive Streambeds: Seepage Forces and Matric Suction. *Earth Surf. Process. Landf.* **2001**, *26*, 1421–1442. [\[CrossRef\]](#)

-
63. Huang, Z.; Bai, Y.; Xu, H.; Cao, Y.; Hu, X. A Theoretical Model to Predict the Critical Hydraulic Gradient for Soil Particle Movement under Two-Dimensional Seepage Flow. *Water* **2017**, *9*, 828. [[CrossRef](#)]
 64. Liu, L.; Liang, J.; Ma, L.; Zhang, H.; Li, Z.; Liang, S. Gas Pipeline Flow Prediction Model Based on LSTM with Grid Search Parameter Optimization. *Processes* **2022**, *11*, 63. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.