The Prediction Model of Dam Uplift Pressure Based on Random Forest

Xing Li1,2, Huaizhi Su1,2 and Jiang Hu3

1 College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
2 State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China
3 State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China

370876835@qq.com

Abstract. The prediction of the dam uplift pressure is of great significance in the dam safety monitoring. Based on the comprehensive consideration of various factors, 18 parameters are selected as the main factors affecting the prediction of uplift pressure, use the actual monitoring data of uplift pressure as the evaluation factors for the prediction model, based on the random forest algorithm and support vector machine to build the dam uplift pressure prediction model to predict the uplift pressure of the dam, and the predict performance of the two models were compared and analyzed. At the same time, based on the established random forest prediction model, the significance of each factor is analyzed, and the importance of each factor of the prediction model is calculated by the importance function. Results showed that: (1) RF prediction model can quickly and accurately predict the uplift pressure value according to the influence factors, the average prediction accuracy is above 96%, compared with the support vector machine (SVM) model, random forest model has better robustness, better prediction precision and faster convergence speed, and the random forest model is more robust to missing data and unbalanced data. (2) The effect of water level on uplift pressure is the largest, and the influence of rainfall on the uplift pressure is the smallest compared with other factors.

1. Introduction
As the high dam reservoir and high dam constructed in the complex foundation is increasing, in order to ensure the safe operation of the dam and give full play to the role of water conservancy project, people pay more and more attention to the dam safety monitoring[1]. The monitoring data of deformation, seepage, stress, strain and temperature of the dam and dam foundation can be obtained through the dam safety monitoring. Analysis and comprehensive evaluation of these monitoring data can timely understand the working state of the dam, find hidden dangers and take appropriate engineering measures to avoid the occurrence or expansion of the accident. Analysis of monitoring data reveals the internal relationship between the actual operating state of the dam and the various factors. It provides scientific basis and practical information for the safety of flood control and maintenance of dam and is an important basis of evaluation the dam workability state [2].

Uplift pressure on the dam foundation refers to the uplift force produced by osmotic pressure and the tail water on the dam foundation which produced by the upstream and downstream head difference.
Uplift pressure on dam body and dam foundation can cause damage. It not only has the characteristics of it changes with time and space distribution, and it will also cause the changes of other forces, such as reduce the shear strength of the sliding surface, dissolution the cementing material of dam, erosion and dissolution the rock filling[3]. For concrete dam, the size of uplift pressure of dam foundation can judge the effect of grouting and drainage system. If the design value is exceeded, is a big danger for the dam safety. Dam foundation uplift pressure is an important load in the bottom of the dam. It has obvious influence on the stability, stress and deformation of the dam. It is very important to observe and analyse the uplift pressure of dam foundation due to the effect of upward uplift pressure can decrease the effective weight of the dam, decrease the sliding resistance and threaten the safe operation of the dam.

At present, the uplift pressure prediction models include statistical model, deterministic model, hybrid model, grey model, neural network model and so on. Wu Zhongru analyzed the effect of each component and the established a statistical model of concrete dam foundation uplift pressure[4]; On the basis of three-dimensional seepage finite element calculation and combined with safety monitoring data, Zhang Qianfei et al have put forward the dam seepage coefficient inversion method [5]; Based on the analysis of the influencing factors of uplift pressure, Jiang Yu and Wang Zuqiang established the monitoring and forecasting model of uplift pressure by stepwise regression[6]; Gu Chongshi et al. Studied the application of gray relational grade and fuzzy clustering analysis principle in the prediction of pressure in Xin'anjiang dam based on the pressure observation data[7]; WANG Wei et al. Improved particle swarm optimization for dam foundation pressure prediction by introducing incentive factors and penalty factors[8]; Zhou Jian et al. Using artificial neural network to quantitatively determine the influence ratio of the influence of the pressure on the dam [9]. Rankovic´ V et al. Developed the support vector regression identification model for prediction of dam structural behavior [10]; Mata J. used the artificial neural network and multiple linear regression models to interpret the dam behavior [11]; Demirkaya S. used ANFIS to analyze the deformation of arch dam [12].

In the previous model, grey prediction model in the system under the condition of stationary time series prediction accuracy is higher, under the system of the volatile situation prediction accuracy is low, and only for small samples to forecast, forecast time is short. The neural network model[13] has the advantages of nonlinear, self-adaptive and robust computing and information processing capabilities. Modelling and control of complex systems can be solved with uncertainty, serious nonlinear and time-varying lag, but this model has shortcomings of slow convergence speed and training time is too long. It is difficult to meet the system requirements in real-time, and in learning with limited samples to ensure accuracy. When learning a large number of samples, the model will fall into “dimension disaster” and the generalization is not high. Statistical forecasting model is a traditional forecasting method, which has a good adaptability, but the choice of the model factor has great randomness. The regression analysis [14] needs more original data, the amount of calculation is large, it is difficult to extrapolate and the prediction accuracy need to be improved. In view of the advantages and disadvantages of these models, this paper introduces the Random Forest (RF) algorithm to predict the dam uplift pressure. The basic idea of RF is multiple weak classifier together to form a strong classifier [15], which play a complementary role between weak classifier, this can reduce the impact of a single classifier error and improve classification accuracy and stability. As a natural nonlinear modelling tool, RF [16] has a very good effect on the prediction of multi variables, thus being applied to many fields. A large number of theoretical and practical examples show that the RF has strong data mining ability and very high prediction accuracy[17][18][19], and has very good tolerance to outliers and noise and not prone to fitting, even is known as one of the best algorithm.

2. The principle of random forest algorithm
Random forest algorithm is a kind of integrated learning algorithm, which is put forward by master Leo Breiman of statistical learning[16]. It is a combination of a number of random trees, each of
which is independent of each other, and the introduction of randomness in the selection of training samples and the growth of trees in order to reduce the variance of tree structure classifier.

Random forest is a classifier composed of several classification decision trees, and the independent and identically distributed random vectors are used to determine the growth process in each classifier. The final result of the model is determined by the majority vote of all the trees[16].

As shown in Figure 1 and 2, the random forest is a multi-decision tree classifier, a subset of data subsets is randomly sampled from the original training set as the training subspace in each classifier’s construction. Then construction on this random subspace, in the process of construction, each of the feature selection is based on the characteristics of random subspace, and feature selection is based on the information gain metric, finally a decision tree model is formed on each bag. Then the final training model can be regarded as a collection of these trees, and the classification result is decided by the classification results of the majority tree model. The characteristics of random forests are mainly concentrated in three aspects: first, the training sample subset is randomly selected, that is the base classifier training set is formed by many times sampling with replacement. Second, the feature subset is selected randomly, that is the feature set is based on random sampling before considering the importance of feature. Third, let all the tree models grow freely, without pruning. Because of the randomness processing of these aspects, the random forest is improved greatly compared with the decision tree. If the sample set has a slightly larger amount of data in the original decision tree, it is easy to form a luxuriant foliage situation, the generated model will be very complex, that will inevitably lead to overfitting of the data. The idea of random forest is to train a decision tree has the decision ability in a certain aspect. The decision tree has no problem of over complexity and over fitting. Generally speaking, it is a weak decision tree, but many aspects of the weak classifier set can form a powerful classifier.

**Figure 1.** The generation steps of RF
3. Random forest model

Random forest is a classifier composed of a series of tree classifiers $h(x, \theta_k), k = 1, \ldots$, here $\theta_k$ is an independent and identically distributed random vector, and each tree votes an equal vote to get the most popular class for the input variable $x$, and $k$ is the number of trees, $x$ is the input sample vector, $\theta_k$ is the random vector actually the parameter vector of the $k$-tree, it was determined independently and identically through learning on the basis of independent and identically distributed bootstrap sets. When at the prediction stage the $k$-tree determined by $\theta_k$ is used to predict the input vector $x$. The prediction value $y$ of an input vector $x$ is determined by voting on the output of all the trees $h(x, \theta_k), k = 1, \ldots$.

Random forests can be directly used for regression analysis. The main research of statistical regression analysis is the dependency relationship of a variable to one or multiple independent variables. It is helpful to understand the typical values of dependent variables with independent variables when the other independent variables are fixed. In general, regression analysis is used to estimate the conditional expectation of the dependent variable in the case of a given independent variable. The goal of the estimator is a function called the regression function to describe the independent variables of the probability distribution. In other words, it is assumed that there exists a certain correlation between random variable $y$ and $x$ as a dependent variable. The $F(y|x)$ expresses the distribution function of $y$ when the $x$ get a certain value, if the regulation of $F(y|x)$ change with the value of $x$ is mastered, then we can fully grasp the relationship between the random variables $y$ and $x$. It is often not possible to obtain a complete distribution function, as a result we can investigate the mathematical expectation of $F(y|x)$ as an approximation. The mathematical expectation of random variable $y$ (that is, the mathematical expectation of $F(y|x)$) depends on the value of $x$, which is a function of $x$, the function is denoted as $\mu_y|x$. It is the regression function about $x$ and $y$. 

Figure 2. Single risk classification tree

Figure 3. Flow chart of uplift pressure prediction based on random forest
The random forest used for regression analysis is obtained by training the tree with random vector $\theta$ as parameters. The difference between the random forest for classification and the random forest for regression is that the target of the tree predictor $h(x \mid \theta)$ is a numerical value rather than a class label, the output of forest is numerical value.

For the random forest used for regression analysis, it is assumed that the training set is independent extracted from random variable $Y$ and $X$. Like all predictor its mean square random error can be expressed as $\text{E}_{X,Y}(Y - h(X)^2).

The random forest predictor is derived from the average value of these trees $h(x, \theta)$ about $k$. So when the number of trees in the forest increases indefinitely, almost everywhere we get $\text{E}_{X,Y}(Y - \text{av}_kh(X, \theta))^2 \rightarrow \text{E}_{X,Y}(Y - E_{\theta}h(X, \theta))^2$ (1)

Therefore, the random forest regression function is: $Y = E_{\theta}h(X, \theta)$, in practice, can be used in the case of sufficiently large $k$ approximation formula: $Y = \text{av}_kh(X, \theta)$ instead.

At this point, the generalization error of random forest is represented by $\text{PE}^*$. The average generalization error of a decision tree is:

$$\text{PE(tree)} = E_{\theta}\text{E}_{X,Y}(Y - h(X, \theta))^2$$ (2)

Assuming that for all the $\theta$ we can get $EY = E_{\theta}h(X, \theta)$, there is

$$\text{PE(forest)} \leq \rho \text{PE(tree)}$$ (3)

Here $\rho$ is the weighted correlation coefficient between the residue of $Y - h(x, \theta)$ and $Y - h(x, \theta)$, $\theta$ and $\theta$ are independent of each other. Prove:

$$\text{PE(forest)} = E_{X,Y}\left[E(Y - h(X, \theta))\right]^2 = E_{\theta}E_{\theta}E_{X,Y}(Y - h(X, \theta))(Y - h(X, \theta))$$ (4)

On the right side of the above formula is a convergent series, it can be written in the following form:

$$E_{\theta}E_{\theta}\left(\rho(\theta, \theta)sd(\theta)sd(\theta)\right)$$

Here $sd(\theta)\left(E_{X,Y}(Y - h(X, \theta))^2\right)^{1/2}$, the weighted correlation coefficient can be defined as:

$$\rho = E_{\theta}E_{\theta}\left(\rho(\theta, \theta)sd(\theta)sd(\theta)\right)(E_{\theta}sd(\theta))^2$$ (5)

Then we can get: $\text{PE(forest)} = \rho(E_{\theta}sd(\theta))^2 \leq \rho \text{PE(tree)}$

It shows that the generalization error of random forest is less than $\rho$ times than that of the membership tree, by introducing random variables x and y through bagging integration can reduce $\rho$ to obtain a better generalization error. The modeling process of the uplift pressure prediction model based on random forest algorithm is shown in Figure 3.

4. Case study
A dam is a concrete gravity arch dam, the maximum dam height is 76.3m. The dam is divided into 28 dam block (from left to right numbered #3~#30). In order to monitor the dam foundation uplift pressure distribution and water blocking effect of dam foundation curtain, 55 uplift pressure holes are fixed up in the dam foundation. In order to verify the feasibility of the proposed algorithm in this paper, the uplift pressure of the dam in the #8 dam block is taken as an example, the proposed method and support vector machine model are respectively used to analyze the uplift pressure. A total of 230
sets of measured data of the effect and the environment were collected from January 2003 to May 2005. The first 200 sets of data are used as the modeling data of the prediction model, and the latter 30 sets of data are used as the prediction interval. Only part of the data was listed in the table 1 due to the limited space of this paper.

| Date     | Upstream water level(m) | Downstream water level(m) | Temperature(°C) | Rainfall(mm) | Uplift pressure(m) |
|----------|-------------------------|---------------------------|-----------------|--------------|-------------------|
| 2003/1/6 | 111.23                  | 58.35                     | 0.6             | 0            | 112               |
| 2003/1/13| 111.01                  | 58.42                     | 8.0             | 0            | 112               |
| 2003/1/20| 110.84                  | 58.49                     | 5.7             | 0            | 112               |
| 2003/1/27| 110.79                  | 58.49                     | 2.1             | 0            | 112               |
| 2003/2/3 | 110.78                  | 58.49                     | 4.7             | 0            | 112               |
| 2003/2/10| 110.95                  | 58.49                     | 7.4             | 10           | 112               |
| 2003/2/17| 110.89                  | 60.42                     | 7.0             | 0            | 112               |
| 2003/2/24| 110.03                  | 58.47                     | 8.5             | 0            | 110.98            |

The following 18 parameters [20][21] are selected as the input layer of the RF regression prediction model in this paper: (1) water pressure component $H_1$: Upstream water level of monitoring day; (2) water pressure component $H_2$: the average upstream water level of 1~4 days before the monitoring day; (3) water pressure component $H_3$: the average upstream water level of 5~10 days before the monitoring day; (4) water pressure component $H_4$: the average upstream water level of 11~20 days before the monitoring day; (5) water pressure component $H_5$: the average upstream water level of 21~30 days before the monitoring day; (6) water pressure component $H_6$: Downstream water level of monitoring day; (7) temperature component $T_1$: $\sin\left(\frac{2\pi t}{365}\right) - \sin\left(\frac{2\pi t_0}{365}\right)$; (8) temperature component $T_2$: $\cos\left(\frac{2\pi t}{365}\right) - \cos\left(\frac{2\pi t_0}{365}\right)$; (9) temperature component $T_3$: $\sin\left(\frac{4\pi t}{365}\right) - \sin\left(\frac{4\pi t_0}{365}\right)$; (10) temperature component $T_4$: $\cos\left(\frac{4\pi t}{365}\right) - \cos\left(\frac{4\pi t_0}{365}\right)$; (11) rainfall component $P_1$: rainfall of monitoring day; (12) rainfall component $P_2$: rainfall 1 days before the monitoring day; (13) rainfall component $P_3$: rainfall 2 days before the monitoring day; (14) rainfall component $P_4$: the average rainfall of 3~4 days before the monitoring day; (15) rainfall component $P_5$: the average rainfall of 5~15 days before the monitoring day; (16) rainfall component $P_6$: the average rainfall of 16~30 days before the monitoring day; (17) time effect $S_1$: $\theta - \theta_0$; (18) time effect $S_2$: $\ln \theta - \ln \theta_0$.

Model parameter selection: in this paper, the model of dam uplift pressure based on random forest is constructed by calling the random forest package in R. The model mainly needs to set two parameters: (1) $n_{tree}$ is the number of the tree, which represents the number of single tree in the model. The greater value of the $n_{tree}$, the smaller the fitting effect of the prediction. In general, the $n_{tree}$ is not less than 100. (2) $m_{try}$ is the split nodes of the decision tree of the variable which indicates the number of the candidate variables are selected randomly in each split. The greater the value of $m_{try}$, the smaller the difference between the sub prediction models. This parameter is important. In general, the default value of the parameter is the variable number for secondary root in the classification model.
of the number of variables in the regression model. In order to achieve better prediction results, in this paper, we select the optimal values of \( m_{try} \) and \( n_{tree} \) through successive calculation in the process of model building. The obtained results through calculation shown in Fig.4 and Fig.5.

![Figure 4.](image1.png)  
**Figure 4.** Relationship between \( m_{try} \) value and mean square deviation

![Figure 5.](image2.png)  
**Figure 5.** Relationship between \( n_{tree} \) value and OOB error

From the figure, it is known the error is small when \( m_{try} = 12, n_{tree} = 400 \), so the parameters of prediction model take the above value to calculate.

In order to evaluate the accuracy of RF in dam safety monitoring of uplift pressure prediction, the model of support vector machine (SVM) is selected as a comparison. SVM has a very strong nonlinear processing ability[2][10][22], is also a kind of efficient and reliable artificial intelligence algorithm, the comparison of the two models are more meaningful. Through the calculation of R, the process graph of learning and prediction effect of random forest model is shown in figure 6, and the residuals graph of the RF model is shown in figure 7; the process graph of learning and prediction effect of SVM model is shown in figure 8, and the residuals graph of the SVM model is shown in figure 9. The degree of importance of the independent variables of the RF prediction model calculated by the built-in function is shown in figure 10.

![Figure 6.](image3.png)  
**Figure 6.** Fitting prediction process line of RF model
Figure 7. Fitting prediction residuals of RF model

Figure 8. Fitting prediction process line of SVM model

Figure 9. Fitting prediction residuals of SVM model

Figure 10. Independent variable importance degree of RF model
In this paper, we compare the fitting and prediction accuracy of the two models by using the mean square error $\sigma$ of the residuals $s$. The expression of $\sigma$ [23] is shown in equation (6), the results of comparison is shown in table 2.

$$\sigma = \left( \frac{1}{n} \sum_{i=1}^{n} (s_i - \bar{y}_i)^2 \right)^{1/2}$$

(6)

$n$ is the number of samples, $y_i$ is the monitoring data, $\bar{y}_i$ is the calculation value of the model.

| Time interval                  | Mean square error of model |         |
|-------------------------------|---------------------------|---------|
|                               | RF model                  | SVM model |
| Mean square error of fitting  | 0.175                     | 0.292   |
| time                          |                           |         |
| Mean square error of          | 0.817                     | 1.109   |
| prediction time               |                           |         |

From the figures and table above:

(1) It can be drawn from the importance degree figure of variables: The upstream water level has the greatest influence on uplift pressure, the upstream water level rises, the uplift pressure increases, the upstream water level drops, the uplift pressure decreases. The effect of rainfall on uplift pressure is the minimum, this is also consistent with the actual situation. However, due to the lack of the measured data of the non-load factors, in this case only considered the influence of load factors. If the non-load factors are taken into account, the prediction results will be closer to the actual situation.

(2) The residual of the SVM model is larger than that of RF model from the fitting and prediction effect of the two models. As can be seen from table 2, the mean square deviation of RF model is less than the corresponding value of SVM model, which indicates that the RF model has a better fitting accuracy. Moreover, the prediction mean square error of the RF model is smaller than that of the SVM model, which shows the RF model has a better prediction effect.

Footnotes should be avoided whenever possible. If required they should be used only for brief notes that do not fit conveniently into the text.

5. Conclusion

In this paper, the random forest algorithm is applied to the uplift pressure prediction model of dam safety monitoring, construction of uplift pressure prediction model by RF algorithm, to explore the new way to predict the uplift pressure, in order to provide reference for dam safety management, normal operation of the dam and disaster early warning. It can be drawn from the practical examples and analysis:

(1) The uplift pressure prediction model based on random forest, the training error is taken as the constraint condition of the optimization problem, and the minimum confidence range is used as the optimization objective, the theoretical basis is clear, and the generalization ability is obviously superior to other machine learning methods such as support vector machine. It can capture the global nonlinear dynamic characteristics of the uplift pressure in the process of dam work, and obtain excellent results in fitting and forecasting applications.

(2) Random forests method also can be used to pre-treatment of dam safety monitoring data, it can extract the implicit rules of between independent variables and dependent variable depends on the model of the random forest, and extract the independent variables that have great effect on the dependent variables in the case of large data sets, remove the independent variables that have little effect to simplify the input variables and improve the computational efficiency.

(3) Due to the lack of measured data of non-load factors, only the influence of load factors is considered in the process of building an uplift pressure prediction model. If the measured data of non-
load factors can be obtained in the future, in the process of modelling, the non-load factors can be taken into account, the model can get a higher prediction precision, and the warning effect of the uplift pressure prediction model will be greatly improved.

Acknowledgements
This research has been partially supported by National Natural Science Foundation of China (SN: 51579083, 51409167, 41323001).

References
[1] Wu Zhong-ru, W.A.H.S., Dam health diagnosis and evaluation. Smart Materials and Structures, 2005. 14(3): p. S130.
[2] Su, H., Z. Chen and Z. Wen, Performance improvement method of support vector machine-based model monitoring dam safety. Structural Control and Health Monitoring, 2016. 23(2): p. 252-266.
[3] Zhu, K., et al., The Analysis of the Concrete Gravity Dam’ s Foundation Uplift Pressure under the Function of Typhoon. Mathematical Problems in Engineering, 2016. 16: p. 1-9.
[4] Wu Zhong-ru. Safety Monitoring Theory & Its Application of Hydraulic Structures. Higher Education Press, 2003.
[5] Zhang Qian-fei, Wang Jian, Wu Zhong-ru. Divisional back analysis of dam percolation parameter based on artificial neural network model. Water Resources and Power, 2001. 19(4): 4-7.
[6] Jiang Yu, Wang Zu-qiang, Li Yu-qi. Data analysis method of uplift pressure monitoring for concrete gravity dam. Pearl River, 2011, 32(2): 47-50.
[7] Jia Cai-hong,Gu Chong-shi. Cluster analysis-based new method for prediction of uplift pressure of dams. Journal of Hohai University (Natural Sciences), 2002, 30(3): 110-112.
[8] Wang Wei, Shen Zhen-zhong. Prediction method of dam foundation pressure based on improved particle swarm optimization algorithm. Yellow River, 2009 (7): 121-123.
[9] Zhou Jian, Song Han-zhou. Application of improved neural network model into analysis of uplift pressure under dam-foundation. Hydrogeology and Engineering Geology, 2005, 32(3): 38-41.
[10] Ranković, V., et al., Development of support vector regression identification model for prediction of dam structural behaviour. Structural Safety, 2014. 48: p. 33-39.
[11] Mata J. Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. Eng Struct 2011;3 (3):903–10.
[12] Demirkaya S. Deformation analysis of an arch dam using ANFIS. In: Proceedings of the second international workshop on application of artificial intelligence and innovations in engineering geodesy, Braunschweig, Germany; 2010. p. 21–31.
[13] Kao, C. and C. Loh, Monitoring of long-term static deformation data of Fei-Tsui arch dam using artificial neural network-based approaches. Structural Control and Health Monitoring, 2013. 20(3): p. 282-303.
[14] De Sortis, A. and P. Paoliiani, Statistical analysis and structural identification in concrete dam monitoring. Engineering Structures, 2007. 29(1): p. 110-120.
[15] Breiman, L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA, 2002. 1.
[16] Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5-32.
[17] DONG, L., X. LI and K. PENG, Prediction of rockburst classification using Random Forest. Transactions of Nonferrous Metals Society of China, 2013. 23(2): p. 472-477.
[18] Chen, X. and H. Ishwaran, Random forests for genomic data analysis. Genomics, 2012. 99(6): p. 323-329.
[19] Mihai Mihailescu, D., Computer aided diagnosis method for steatosis rating in ultrasound images using random forests. Medical Ultrasonography, 2013. 15(3): p. 184-190.
[20] Su, H.Z., J. Hu and Z.R. Wu, A study of safety evaluation and early-warning method for dam global behavior. Structural Health Monitoring, 2012. 11(3): p. 269-279.
[21] Su, H., et al., Time-varying identification model for dam behavior considering structural reinforcement. Structural Safety, 2015. 57: p. 1-7.
[22] Kuang, F., W. Xu and S. Zhang, A novel hybrid KPCA and SVM with GA model for intrusion detection. Applied Soft Computing, 2014. 18: p. 178-184.
[23] Su, H., Z. Wen and Z. Wu, Study on an Intelligent Inference Engine in Early-Warning System of Dam Health. Water Resources Management, 2011. 25(6): p. 1545-1563.