water

MDPI

Article

Concrete Dam Displacement Prediction Based on an ISODATA-GMM Clustering and Random Coefficient Model

Yating Hu [1,2,3], Chenfei Shao [1,2], Chongshi Gu [1,2,*] and Zhenzhu Meng [3]

1 College of Water Conservancy and Hydropower Engineering, Hohai University, 210098 Nanjing, China; huyating_hhu@163.com (Y.H.); chenfeishao.hhu@gmail.com (C.S.)
2 National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, 210098 Nanjing, China
3 Laboratory of Environmental Hydraulics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; zhenzhu.meng@epfl.ch
* Correspondence: csgu@hhu.edu.cn; Tel.: +86-13809043532

Received: 15 February 2019; Accepted: 2 April 2019; Published: 6 April 2019

Abstract: Displacement data modelling is of great importance for the safety control of concrete dams. The commonly used artificial intelligence method modelled the displacement data at each monitoring point individually, i.e., the data correlations between the monitoring points are overlooked, which leads to the over-fitting problem and the limitations in the generalization of model. A novel model combines Gaussian mixture model and Iterative self-organizing data analysing (ISODATA-GMM) clustering and the random coefficient method is proposed in this article, which takes the temporal-spatial correlation among the monitoring points into account. By taking the temporal-spatial correlation among the monitoring points into account and building models for all the points simultaneously, the random coefficient model improves the generalization ability of the model through reducing the number of free model variables. Since the random coefficient model supposed the data follows normal distributions, we use an ISODATA-GMM clustering algorithm to classify the measuring points into several groups according to its temporal and spatial characteristics, so that each group follows one distribution. Our model has the advantage of having a stronger generalization ability.

Keywords: dam safety; displacement; Gaussian mixture model; iterative self-organizing data analysing; random coefficient model

1. Introduction

Dam safety monitoring aims to understand the actual running status of the dam, so as to provide sufficient information to ensure the safety of the concrete dam [1]. Displacement is a dominant indicator of the safety of the dam. One of the most important topics in dam safety management is to forecast the dam's displacement from the displacement data obtained from the monitoring points laid inside the dam [2,3].

Researchers have established many displacement forecast models. At the very beginning, researchers developed statistical models, in which the displacement δ at each monitoring point can be approximated by: $\delta = \delta_H + \delta_\theta + \delta_t + k$, where δ_H, δ_θ and δ_t are displacements due to hydrostatic pressure, temperature and ageing, respectively; k is determined by regression analysis [4–6]. The δ_H is usually fitted by a polynomial equation of the upstream water level in reservoir H: $\delta_H = \alpha_1 H + \alpha_2 H^2 + \alpha_3 H^3 + \alpha_4 H^4$. However,

the prediction accuracy of the statistical models is limited due to its uncertainty and the multicollinearity caused by the high correlation between explanatory variables.

In recent years, many artificial intelligence models, such as artificial neural network [7], grey system model [8], support vector machines [9], and genetic algorithm [10] have been applied in the displacement data analysis and the prediction. Whether in statistical models or artificial intelligence models, the explanatory variables' coefficients at each monitoring points were analysed and predicted independently, herein the spatial correlation of each monitoring point is overlooked [11]. However, the displacement of adjacent monitoring points are correlative, as both the hydrostatic pressure and temperature acting on the dam are gradually varied. In addition, in order to obtain more accurate fitting results, redundant explanatory variables are adopted in the models which may weaken the generalization ability of the model.

Instead of modelling the data of each monitoring point individually, we introduced the random coefficient model of multi-dimensional data. The random coefficient model can model the data of several monitoring points synchronously, and make the explanatory variables' coefficients of each measuring point satisfy asymptotic normal distributions [12,13]. Of course, not all coefficients follow the same normal distribution. Therefore, we classified the measuring points into several clusters based on their structural and temporal characteristics. Then, we can assume that each cluster follows the same distribution.

The clustering methods mainly fall into two categories. One is based on similarity or dissimilarity distances such as hierarchical cluster analysis [14] and K-means algorithm [15]. Another is model-based method in which each cluster is represented by a parametric distribution such as Gaussian distribution, and the entire dataset is modelled by a mixture of these distributions [16–18]. Model based clustering provides a rigorous framework to assess the number of clusters and the role of each variable in the clustering process. In this study, we clustered the data using the Gaussian mixture model (GMM), which assumes a multivariate Gaussian distribution for each component. To avoid the divergence in the GMM model, an iterative self-organizing data analysis (ISODATA) was used to govern the number of individuals in each class.

This article is organised as follows. Section 2 introduces the classical statistical prediction model. The model developed in this study is presented in Section 3, where we first present the clustering method based on ISODATA-GMM in Section 3.1 and then a random coefficient model of multidimensional data in Section 3.2. Section 4 describes the data sets. The predicting results and comparison with statistical model are discussed in Section 5. Concluding remarks complete the paper in Section 6.

2. Statistical Prediction Model

The dam displacement δ includes displacement due to three components: temperature component δ, aging component δ_θ, and water pressure component δ_H—among which, the δ_H is mainly composed of deformation of three parts: the dam body itself δ_{1H}, the dam foundation δ_{2H} and the rotation of the dam bedrock δ_{3H} (see Figure 1):

$$\delta_H = \delta_{1H} + \delta_{2H} + \delta_{3H} \tag{1}$$

δ_H, δ_{2H} and δ_{3H} are mainly dependent with the upstream water level H, as exhibited in Equations (2)–(4):

$$\begin{aligned}\delta_{1H} =& \frac{\gamma_0}{E_c m^3}\left[(h-d)^2 + 6(h-H)\left(dln\frac{h}{d} + d - h\right) + 6(h-H)^2\left(\frac{d}{h} - 1 + dln\frac{h}{d}\right) - \frac{(h-H)^3}{h^2 d}(h-d)^2\right] \\ &+ \frac{\gamma_0}{G_c m}\left[\frac{h^2-d^2}{4} - (h-H)(h-d) + \frac{(h-H)^2}{2} ln\frac{h}{d}\right]\end{aligned} \tag{2}$$

$$\delta_{2H} = \left[\frac{3\left(1-\mu_r^2\right)\gamma_0}{\pi E_r m^2 h^2} H^3 + \frac{\left(1+\mu_r\right)\left(1-2\mu_r\right)\gamma_0}{2E_r m h} H^2 \right] (h-d) \tag{3}$$

$$\delta_{3H} = \alpha H \tag{4}$$

where H is the upstream water level; h is the height of the dam; m is the downstream slope; d is the distance between the observation point and the dam crest; E_c, G_c are elastic modulus and shear modulus of dam concrete, respectively; E_r, μ_r are elastic modulus and Poisson's ratio of foundation, respectively; γ_0 is the water density and α is the rotation angle of dam foundation surface at the dam heel.

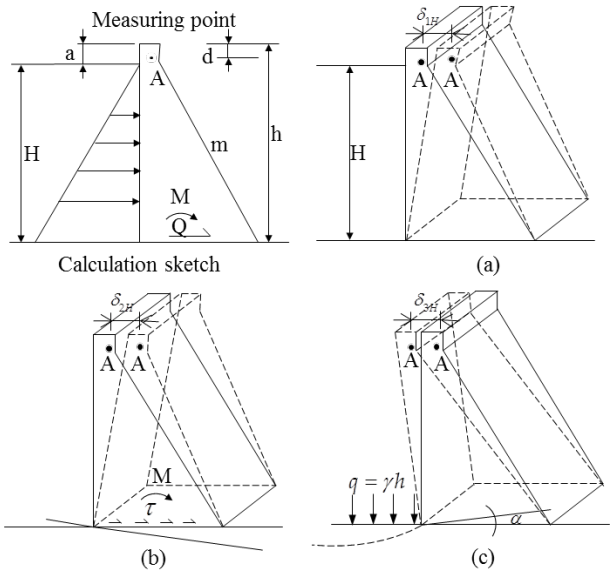

Figure 1. Three components of δ_H: (**a**) δ_{1H}; (**b**) δ_{2H}; (**c**) δ_{3H}.

Due to a lack of measured temperature data in most engineering projects, the temperature component δ_T can be expressed by a trigonometric function of different periods. The aging component δ_θ is commonly described by a fixed form of trend function. In addition to the above-mentioned deformation components, a random interference term k is often considered, which consists of the human errors, measurement errors, etc. According to the central limit theorem, k obeys a normal distribution with a mean of zero.

Equation (5) expresses the most commonly used statistical model of the dam deformation [19]:

$$\delta = \delta_H + \delta_\theta + \delta_T + k = \sum_{i=1}^{n} a_i H^i + (c_1 t + c_2 lnt) + \sum_{j=1}^{2} \left(b_{j1} \sin \frac{2\pi j t}{365} + b_{j2} \cos \frac{2\pi j t}{365} \right) + k \tag{5}$$

where n is a coefficient, $n = 3$ for gravity dam and $n = 4$ for arch dam; t is the time; a, b, c are coefficients; k is the random interference term.

In this paper, the coefficients in the statistical model were solved using an Ordinary Least Squares estimation method.

3. Model Development

We first clustered the measured displacement data obtained from each measuring point using Gaussian Mixture Model (GMM) and improved the model by an Iterative Self-Organizing Data Analysis (ISODATA). The displacement data of 24 measuring points we selected for the case were classified into five groups. We then used the random coefficient model to fit the data of each class.

3.1. Clustering of the Monitoring Data Based on ISODATA-GMM

As we have introduced in Section 2, the displacement is mainly induced by three components: water pressure component δ_H, temperature component δ_T and aging component δ_θ. To build a clustering criterion to represent the spatial and temporal characteristics of measuring points, we have to discuss these three factors, respectively. The water pressure component δ_H and the temperature component δ_T are mainly dependent on the location of the measuring point and the geometrical size of the dam. For concrete dams, the spatial relations between each measuring point can be represented by its distance to the dam foundation d. The temporal characteristic of measuring points mainly affects the aging component δ_θ. We first separated the aging component δ_θ from the time series measured data. The temporal characteristic can be described by two factors: one is the maximum absolute value of the aging sequence λ, and another is the the degree of convergence of the data series ξ, which is expressed by $\xi = \left|\frac{c_1}{c_2}\right|$, where c_1 and c_2 are the aging term coefficients. Therefore, we use d, λ and ξ as the clustering criteria, to represent the spatial and temporal characteristics of the measuring points.

Gaussian Mixture Model (GMM) based clustering assumes that data comes from several sub-datasets which are modelled separately, and the whole dataset is a mixture of these sub-datasets. The resulting model is a finite mixture model. When data are multivariate continuous observations, the parametrized component density is usually a multidimensional Gaussian density.

For a one-dimensional dataset, we assume that the probability distribution of a random variable x follows a mixture of two Gaussian distributions as described in Equation (6):

$$P\left(x|\mu_1,\mu_2,\sigma\right) = \sum_{k=1}^{2} p_k \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\left(x-\mu_k\right)^2}{2\sigma^2}\right) \tag{6}$$

where $k = 1$ and $k = 2$ represent the two Gaussian distributions; the kth prior probability is $\{p_1 = 1/2, p_2 = 1/2\}$; $\{\mu_k\}$ and σ are the mean and the standard deviation of the two Gaussian distributions, respectively. We use $\theta \equiv \{\{\mu_k\}, \sigma\}$ to simplify these parameters.

The dataset $\{x_n\}_{n=1}^N$ which contains N points is assumed as an independent sample from the distribution. k_n denotes the unknown class tag for the nth point.

In the case that $\{\mu_k\}$ and σ are known, the posterior probability of the class tag of the nth point k_n can be written as:

$$P\left(k_n|x_n,\theta\right) = \frac{1}{1+exp\left[-\left(\omega_1 x_n + \omega_0\right)\right]} \tag{7}$$

If the case of $\{\mu_k\}$ is unknown and σ is known, we may deduce the $\{\mu_k\}$ from the data series $\{x_n\}_{n=1}^N$. We hence derive an iterative algorithm of $\{\mu_k\}$ to maximize the likelihood estimation:

$$P\left(\{x_n\}_{n=1}^N \mid \{\mu_k\}, \sigma\right) = \prod_n P\left(x_n \mid \{\mu_k\}, \sigma\right) \tag{8}$$

The natural logarithm of the likelihood L derivation of $\{\mu_k\}$ is:

$$\frac{\partial}{\partial \mu_k} L = \sum_n p_{k|n} \frac{x_n - \mu_k}{\sigma^2} \tag{9}$$

where $p_{k|n} \equiv P(k_n = k|x_n, \theta)$ is the Gaussian density (see Equation (7)). Ignoring the items in $\frac{\partial}{\partial \mu_k} P(k_n = k|x_n, \theta)$, the second derivative versus $\{\mu_k\}$ can be approximated as:

$$\frac{\partial^2}{\partial \mu_k^2} L = -\sum_n p_{k|n} \frac{1}{\sigma^2} \tag{10}$$

Then, the initial μ_1, μ_2 are iterated to μ_1', μ_2' using the approximate Newton–Raphson steps:

$$\mu_k' = \frac{\sum_n p_{k|n} x_n}{\sum_n p_{k|n}} \tag{11}$$

We now expand to the multidimensional dataset (multiple Gaussian distribution). The Gaussian mixture density can be written as:

$$p_{k|n} = \frac{\pi_k \frac{1}{\prod_{i=1}^{I} \sqrt{2\pi}\sigma_i^{(k)}} exp\left(-\sum_i^I \left(\mu_i^{(k)} - x_i^{(n)}\right)^2 / 2\left(\sigma_i^{(k)}\right)^2\right)}{\sum_{k'} \pi_{k'} \frac{1}{\prod_{i=1}^{I} \sqrt{2\pi}\sigma_i^{(k')}} exp\left(-\sum_i^I \left(\mu_i^{(k')} - x_i^{(n)}\right)^2 / 2\left(\sigma_i^{(k')}\right)^2\right)} \tag{12}$$

where k is the serial number of the Gaussian distribution; i is the serial number of the data's dimension; n is the serial number of the data sequence; I is the total number of the data's dimension; π_k is the weight; $\mu_i^{(k)}$ is the mean of the Gaussian distribution; $\sigma_i^{(k)}$ is the variance of the Gaussian distribution; $x_i^{(n)}$ is the data point. The iterative formula of μ_i^k has been presented in Equation (11). The iterative formulas of the variance $\sigma_i^{(k)}$ and the weight π_k are as follows:

$$\sigma_i^{2(k)} = \frac{\sum_n p_{k|n} \left(x^{(n)} - \mu_i^{(k)}\right)^2}{\sum_n p_{k|n}} \tag{13}$$

$$\pi_k = \frac{\sum_n p_{k|n}}{\sum_k \sum_n p_{k|n}} \tag{14}$$

Once the iteration is in convergence, the GMM clustering classified the dataset into several classes.

However, there are still some defects in GMM clustering. The number of classes and the number of data points in one class are unknown before clustering; hence, the iteration may obtain a class with only one or two data points, which may result in the divergence of the final results.

To solve this problem, we introduce the Iterative Self-Organizing Data Analysis (ISODATA) to realize the following functions: (a) separate the class into two when the variance is too large, (b) delete the class when the number of samples below an indicated value, and (c) merge two classes when they are too close. Figure 2 shows the flow chart of ISODATA.

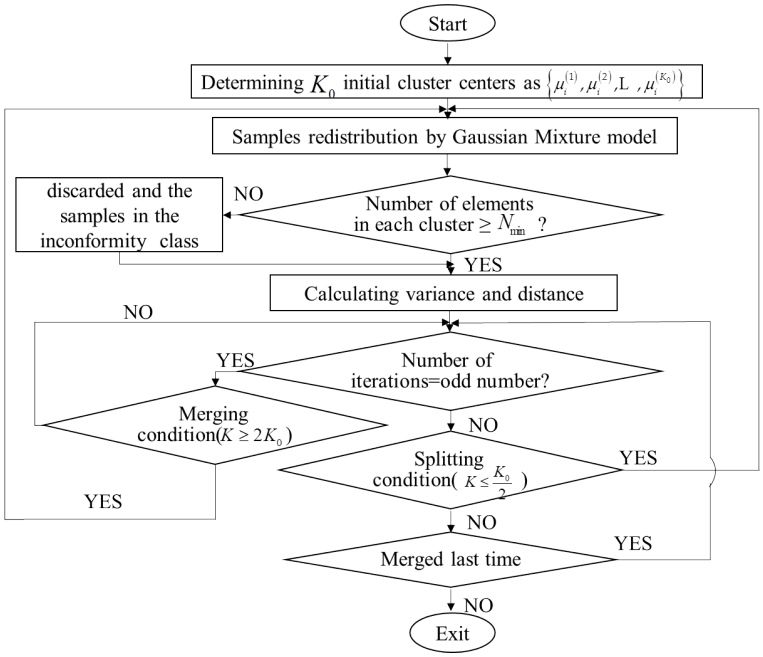

Figure 2. Flow chart of the ISODATA model.

3.2. *Random Coefficient Model*

As shown in Figure 3, the monitoring data is two-dimensional, which contains time series data and cross-sectional data. The data on one panel represents the cross section displacement data at a certain time, and each grid on the panel stands for a monitoring point.The monitoring data of the dam's cross section at an indicated time can be considered as a two-dimensional panel.

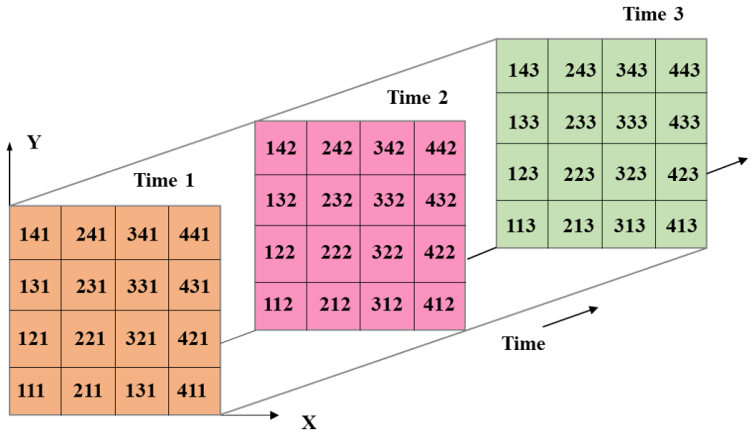

Figure 3. Schematic diagram of the temporal and spatial representation of the monitoring data.

Here, Equation (15) expresses the regression coefficients of a panel without time variation:

$$y_{it} = \sum_{i=1}^{K} \beta_{ki} x_{kit} + u_{it} = \sum_{k=1}^{K} (\beta_k + \gamma_{ki}) x_{kit} + u \tag{15}$$

where y_{it} is the two-dimensional dam displacement data; x_{kit} is the two-dimensional data of explanatory variables; t is the time index; i is the cross section index; k is the explanatory variables index; β_{ki} is independent with time and can be divided into β_k and γ_{ki}; $\beta = (\beta_1, \cdots, \beta_K)^{'}$ is the common mean coefficient vector, $\gamma = (\gamma_{1i}, \cdots, \gamma_{Ki})^{'}$ is the derivation from individual data to the common mean value; u is a random interference term. Ref. [20] assumed that $\beta_i = \beta + \gamma_i$ is a random variable and deduced the following assumptions in Equation (16):

$$\begin{aligned} &E(\gamma_i) = 0 \\ &E(\gamma_i \gamma_j^{'}) = \begin{cases} \Delta, & i = j \\ 0, & i \neq j \end{cases} \\ &E(x_{it} \gamma_j^{'}) = 0 \\ &E(u_i u_j^{'}) = \begin{cases} \sigma_i^2 I_T, & i = j \\ 0, & i \neq j \end{cases} \end{aligned} \tag{16}$$

By integrating the NT observation data, we can obtain the equation in a matrix format (Equation (17)):

$$y = X\beta + \widetilde{X}\gamma + u \tag{17}$$

where

$$y_{NT\times 1} = (y_1^{'}, \cdots, y_N^{'})^{'}, X_{NT\times K} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \widetilde{X}_{NT\times NK} = \begin{pmatrix} X_1 & & & 0 \\ & X_2 & & \\ & & \ddots & \\ 0 & & & X_N \end{pmatrix}$$

$u = (u_1', \cdots, u_N')', \gamma = (\gamma_1', \cdots, \gamma_N')'$, N is the number of panel, T is the number of data in each panel, the compound error term $\widetilde{X}\gamma + u$ is a diagonal matrix, and the i-th diagonal block is $\psi_i = X_i \Delta X_i' + \sigma_i^2 I_T$. According to [20], the estimation of β from OLS is biased. Once $\frac{1}{NT} X'X$ converges to a non-zero constant matrix, we can hence obtain a consistent non-effective estimation. The optimal linear unbiased estimator of β is the generalized least squares estimation:

$$\begin{aligned} &\hat{\beta}_{GLS} = \left(\sum_{i=1}^{N} X_i^{'} \psi_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i^{'} \psi_i^{-1} y_i \right) = \sum_{i=1}^{N} W_i \hat{\beta}_i \\ &W_i = \left\{ \sum_{i=1}^{N} \left[\Delta + \sigma_i^2 (X_i^{'} X_i)^{-1} \right]^{-1} \right\}^{-1} \left[\Delta + \sigma_i^2 (X_i^{'} X_i)^{-1} \right]^{-1} \\ &\hat{\beta}_i = \left(X_i^{'} X_i \right)^{-1} X_i^{'} y_i \end{aligned} \tag{18}$$

The variance of the estimator is:

$$Var\left(\hat{\beta}_{GLS}\right) = \left(\sum_{i=1}^{N} X_i^{'} \psi_i^{-1} X_i \right)^{-1} = \left\{ \sum_{i=1}^{N} \left[\Delta + \sigma_i^2 (X_i^{'} X_i)^{-1} \right]^{-1} \right\}^{-1} \tag{19}$$

$\hat{\beta}_{GLS}$ follows an asymptotic normal distribution and it is the effective estimation of β. The random coefficient model can dominate the explanatory variables coefficients, which makes the coefficients following asymptotic normal distributions instead of being free variables, and hence represents the correlation between adjacent monitoring points.

The distribution density of one monitoring point is strongly dependent on its features such as the location of the monitoring point. Hence, we cluster the measuring points based on its spatial and temporal characteristics. Using the ISODATA-GMM method introduced in Section 3.1, the measuring points with similar spatial and temporal characteristics are classified into the same group. Then, the coefficients in the same cluster can be considered as following the same normal distribution.

4. Data Sets

We selected the concrete dam in the Jinping-I Hydropower Station as an example to validate the model. The station is located at the Yalong River in China (Figure 4). The main feature of the station is generating electricity, with a maximum capacity of 3600 MW. Another feature is flood-control, the gross capacity of reservoir and flood regulation storage capacity are 77.6×10^8 and 49.1×10^8 m^3, respectively. The dam is the world's tallest arch dam at present, it is a double-curved arch dam with the height of 305 m, the crest width of 16 m, the bottom thickness of 63 m and the volume of dam 474×10^6 m^3 [21,22]. The storage of water started from 30 November 2012, while the construction of the dam body were accomplished in June 2013. During this period, the water level was fairly low and hence the associated dam deformation was ignorable. The water level reached a normal value on 23 August 2014.

In this study, we selected radical displacement data (to the downstream is positive, to the upstream is negative) from 16 June 2013 to 25 August 2015 for analysis. The data were measured by plumb lines (PL) and inverted lines (IP) installed at dam crest and dam body. Note that some measuring points had not yet been installed during this period, we selected 24 measuring points. Figure 5 exhibits the distribution map of the measuring points selected in this study. These 24 points are distributed in six perpendiculars on the same cross section.

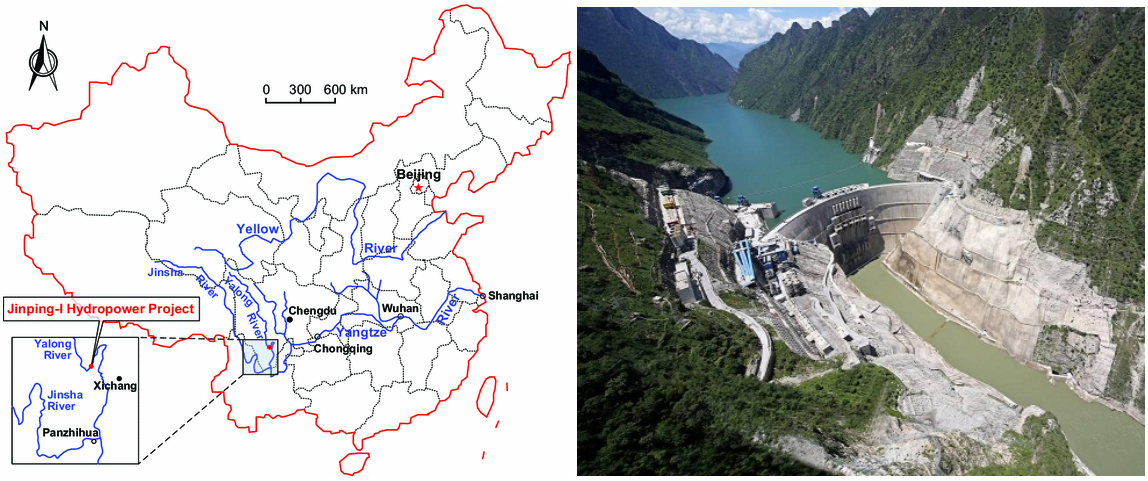

Figure 4. A sketch map for the geographical location of Jinping-I Hydropower Station and its design drawing.

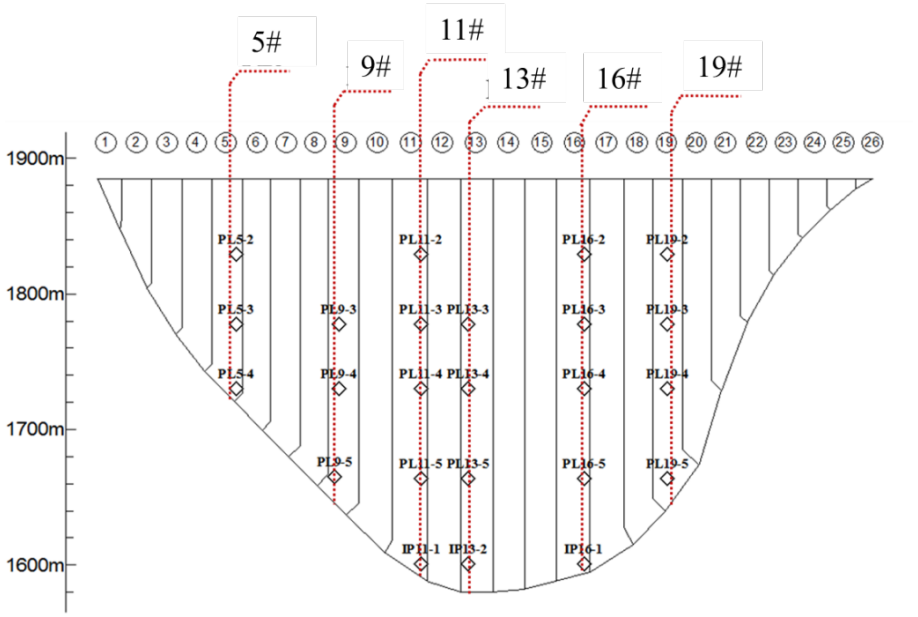

Figure 5. Distribution map of the measuring points.

The data collecting during the flood season is more frequent than in other period. During the storage period and the flood season, the radical displacement data were collected three times each day. For other periods, the data were collected once a week. For the case that there are three pieces of data in one day, we calculated the mean of the three values. For the case that there is only one piece of data in a week, we limited the number of missing data in the panel using the generalized least squares method, and finally obtained 274 validated time frames. The time variation of the water level and the displacement data for all the measuring points are shown in Figure 6. It is obvious that the displacement data of all measuring point is strongly dependent with the water level, which also indicated the importance of taking the temporal characteristics into consideration in data clustering. In addition, many noisy data exist in the time variation of the radical displacement, especially during the period from January 2014 to May 2014 and the period from April 2015 to August 2015. The noisy data commonly came from the measurement errors or human errors, it may reduce the accuracy of the prediction; in another aspect, the prediction results at the noisy point can serve as an indicator to evaluate whether the model is over fitting.

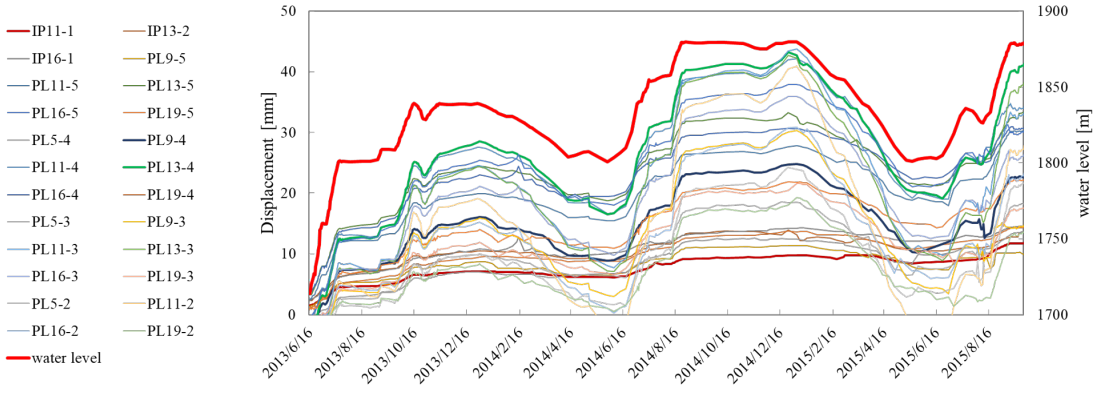

Figure 6. Time evolution of water level and radical displacement (to the downstream is positive) of each measuring point (data provided by Jinping I hydropower station with permission).

As shown in Figure 7, the annual variation of the radical displacement at each measurement point in 2014 (from 1 January 2014 to 31 December 2014) has a strong relevance to its spatial location. More specifically, the variation at one point on the dam is dependent with its distance to the dam foundation. The displacement at the marginal position of the cross section is significantly smaller than the displacement at the central position. Therefore, the distance from the measuring point to the dam's foundation d was selected as the spatial indicator in the clustering criteria.

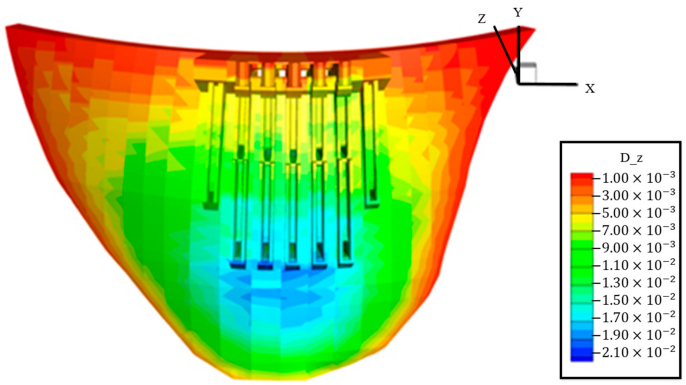

Figure 7. Annual variation of cross section's radical displacement in 2014.

The datasets were divided into two groups: 16 June 2013 to 15 June 2015 as fitting datasets, and 16 June 2015 to 28 September 2015 as testing datasets. We first used the fitting datasets to develop the predicting model, and then used the testing datasets to check the prediction capacity of the developed model.

5. Results and Discussion

5.1. Clustering Results

As introduced in Section 3.1, we clustered the monitoring data obtained from 24 measuring points based on their spatial and temporal characteristics. We used the distance from the measuring point to the dam foundation d as an indicator of the spatial characteristics. The temporal characteristics are represented by two indicators: one is the maximum absolute value of the aging sequence λ; another is the degree of convergence of the data series ξ. As the first step of the clustering, we calculated the three indicators d, λ and ξ for each measuring point as a criteria. The values of d, λ and ξ are shown in Table 1.

Table 1. Indicators d, ξ and λ calculated based on the ISODATA-GMM clustering method.

Measuring Point	d (m)	λ (mm)	ξ (/)	Measuring Point	d (m)	λ (mm)	ξ (/)
IP11-1	8	3.26	2.31	PL11-5	58	6.16	3.49
IP13-2	21	7.18	0	PL13-3	163	3.8	0.43
IP16-1	7	10.08	0.3	PL13-4	130	9.39	0.32
PL5-2	69	5.03	0.7	PL13-5	80	11.46	0.26
PL5-3	40	6.67	0.63	PL16-2	131	4.71	0.68
PL5-4	8	10.85	0.53	PL16-3	113	9.88	0.22
PL9-3	96	5.44	3.49	PL16-4	98	19.93	0.13
PL9-4	62	2.65	3.49	PL16-5	60	13.96	0.32
PL9-5	14	4.22	0.2	PL19-2	76	6.51	0
PL11-2	175	18.49	0	PL19-3	55	4.36	0
PL11-3	139	0	0	PL19-4	38	10.47	1.2
PL11-4	105	10.42	0	PL19-5	13	9.79	0.57

As we can see from Table 1, the dataset of indicators are represented at different scales. The spatial indicator d ranges from 7 to 175, the temporal indicator λ ranges from 0 to 19.93, and the values of ξ are located in the range of 0 to 3.49. In order to eliminate the dimensional influence between different indicators, we normalized the values of indicators d, λ and ξ to $0-1$. Then, we set the initial parameters of the clustering. The initial class number was set to 4; the initial weight parameter π_k was 0.25; the initial variance $\sigma_i^{(k)}$ was 1; the minimum element number N_{min} was 2; the maximum allowable variance σ_{max} was 3; the minimum allowable distance d_{min} was 0.1. Using the ISODATA-GMM method, the clustering results of the measuring points based on the values of d, λ and ξ are shown in Figure 8.

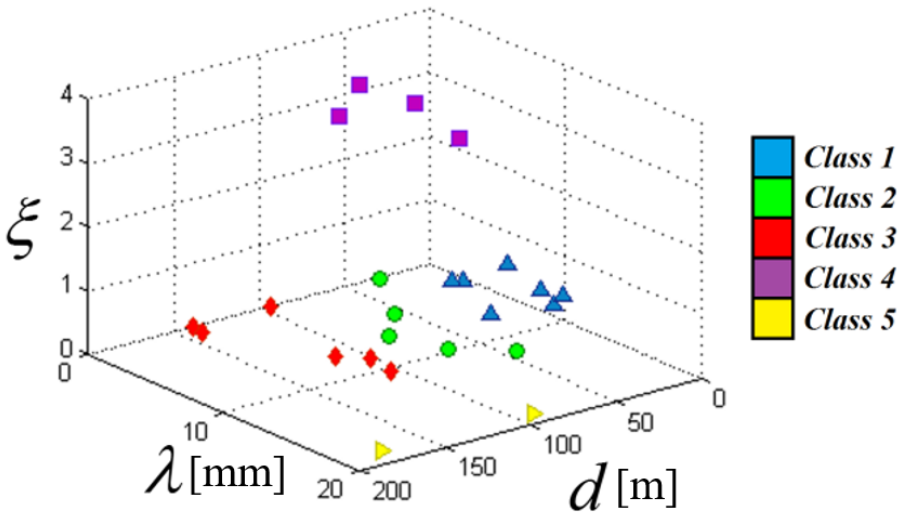

Figure 8. Classification of the measuring points after clustering.

The 24 measuring points were classified into five groups. Figure 9 shows the clustering map of the measuring points. The measuring points IP 13-2, IP 16-1, IP 19-5, PL 19-4, PL 9-5, PL5-4 and PL 5-3 were categorized to Class 1. PL 19-2, PL 19-3, PL 16-5, PL 5-2, PL13-5 were categorized to Class 2. PL 16-3, PL 16-2, PL 11-4, PL 11-3, PL 13-3 and PL 13-4 were belongs to Class 3. PL 11-5, PL 9-4, PL 9-3 and IP 11-1 were classified to Class 4. PL 11-2 and PL 16-4 were Class 5.

It is obvious that the classification roughly corresponds to its spatial location, e.g., all the measuring points in Class 1 were located on the edge area of the dam, and all the measuring points in Class 3 were located in the center part (see Figure 9). Of course, the results are not strictly dependent with their spatial location due to the influence of the temporal indicators—for instance, one of the measuring points in Class 4 was located on the edge while the other three were in the center area. However, these four points in Class 4 were relatively located in adjoining areas. Therefore, it can be thought that the clustering results represent the spatial characteristics of the measuring points. In addition, according to the indicator d exhibited in Table 1, the intervals of indicator d in five classes are [7,40], [55,80], [105,163], [8,96], [98,175], respectively. The extreme difference of the values of d from Class 1 to Class 5 are 33, 25, 58, 88 and 77, respectively. Note that the spatial similarities of measuring points in Classes 1, 2 and 3 are significantly better than those in Classes 4 and 5.

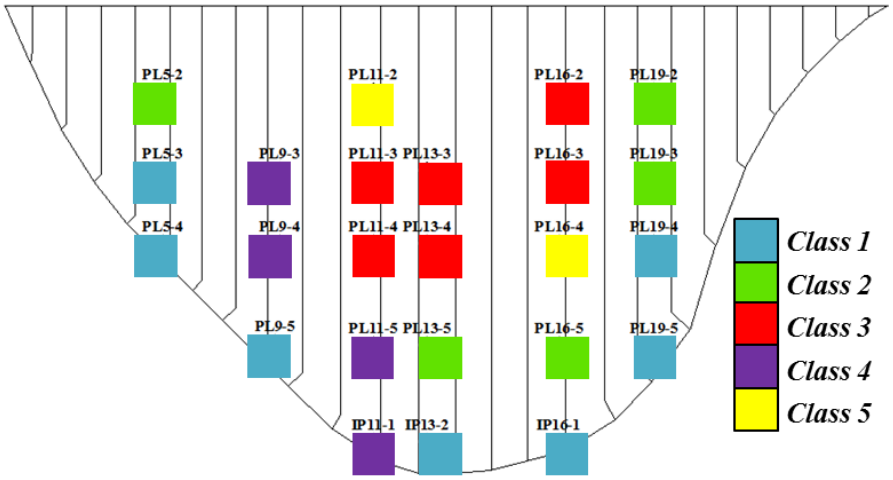

Figure 9. Map of measuring points indicated clustering results.

In Figure 10, the clustering results were exhibited relating to temporal indicators (λ and ξ). It is interesting to note that measuring points in Class 5 (PL11-2 and PL16-4) and Class 4 (IP11-1, PL9-3, PL9-4 and PL11-5) gathered in centers far away other points, respectively. It means that the temporal similarities of measuring points in Classes 4 and 5 are more significant than those in Classes 1, 2, 3, which is opposite to the spatial similarities where measuring points in Classes 1, 2, 3 were significantly better than those in Classes 4 and 5. In general, we can see that the clustering model took both the effects of temporal and spatial factors into consideration.

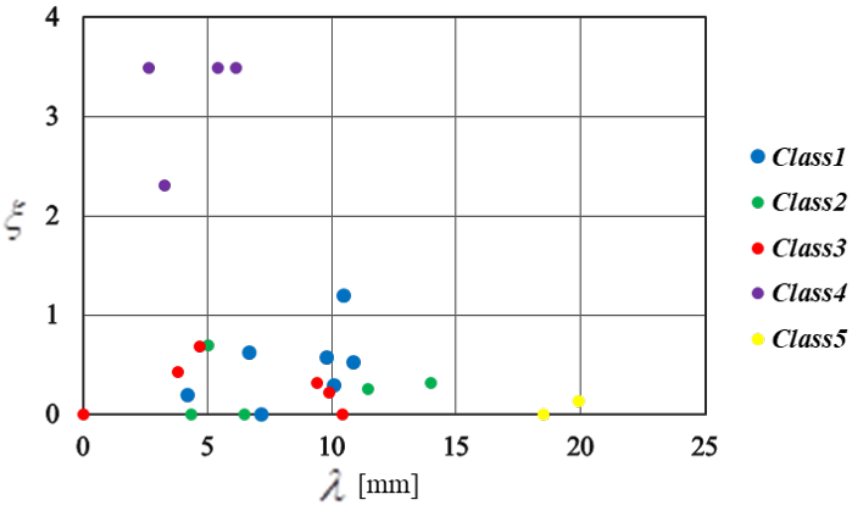

Figure 10. Clustering results of measuring points indicated by temporal characteristics.

5.2. Predicting Results

After we classified the measuring points into five classes using clustering analysis based on the ISODATA-GMM method, we developed a random coefficient model for each class. Here, in order to establish models of monitoring data, we selected the explanatory variables relating to upstream level, temperature and age which include H, H^2, H^3, H^4, $\sin\frac{2\pi jt}{365}$, $\cos\frac{2\pi jt}{365}$, t and lnt, where H is the upstream

water level and t is the time. The water level is expressed by a several exponential function of H, which inspired from the statistical model. The temperature is represented by trigonometric functions of time, by assuming that the temperature follows the same tendency each year. The aging component is described by time t and its natural logarithm directly.

Using the ISODATA-GMM method and random coefficient model, we fitted the displacement data from 16 June 2013 to 15 June 2015 to develop the prediction model. Then, we validated the model with the dataset from 16 June 2015 to 25 August 2015. Figure 11 shows the fitting and forecast results of the seven measuring points in Class 1. The modelling datasets are located in the white area and the testing dataset are located in the blue area. The red dots represent the measured data and the black lines are the fitted and predicted results. Even though there are always some noisy points exist in the measured data, the predicted data fit well with the measured data in the whole. In addition, the fitting data for the other 17 measuring points in Classes 2, 3, 4, 5 are illustrated in Appendix A.

After the prediction model had been developed, we then evaluated the performance of the model and compared it with the statistical model. We used correlation coefficient R and residual standard deviation s as criteria of model performance. Their expressions are as follows:

$$R = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}} \tag{20}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-1}} \tag{21}$$

where $\hat{y}_i$ is the series of fitting data; y_i is the series of measuring data; $\bar{y}_i$ is the mean of measuring data series; and n is the number of measuring data.

The correlation coefficient R and residual standard deviation s represent the strength of the relationship between the measured dataset and predicted dataset. The R and s of each measuring point's testing dataset (from 16 June 2015 to 28 September 2015) are indicated in Table 2. Generally, the model can be validated once the R above 0.9. Here, the R for all the monitoring points are located in the range 0.958–0.999, which represents a fairly well fitting between the predicting data and the measuring data. The maximum R is 0.999 for the measuring points PL9-4 and PL16-4, which means that these two datasets best fit with the model. The s ranges from 0.121 to 1.344.

Table 2. Correlation coefficient R and residual standard deviation s of each measuring point.

Measuring Point	R (-)	s (mm)	Measuring Point	R (-)	s (mm)
IP13-2	0.992	0.313	PL16-3	0.978	1.344
IP16-1	0.991	0.496	PL16-2	0.988	0.302
PL19-5	0.989	0.123	PL13-4	0.991	0.221
PL19-4	0.984	0.432	PL13-3	0.99	0.321
PL9-5	0.989	0.121	PL11-4	0.982	1.27
PL5-4	0.995	0.283	PL11-3	0.992	0.211
PL5-3	0.991	0.568	PL11-5	0.995	0.142
PL19-3	0.998	0.268	PL9-4	0.999	0.542
PL19-2	0.987	0.423	PL9-3	0.958	0.363
PL16-5	0.983	0.752	IP11-1	0.992	1.116
PL5-2	0.998	0.433	PL16-4	0.999	0.215
PL13-5	0.992	0.193	PL11-2	0.985	1.809

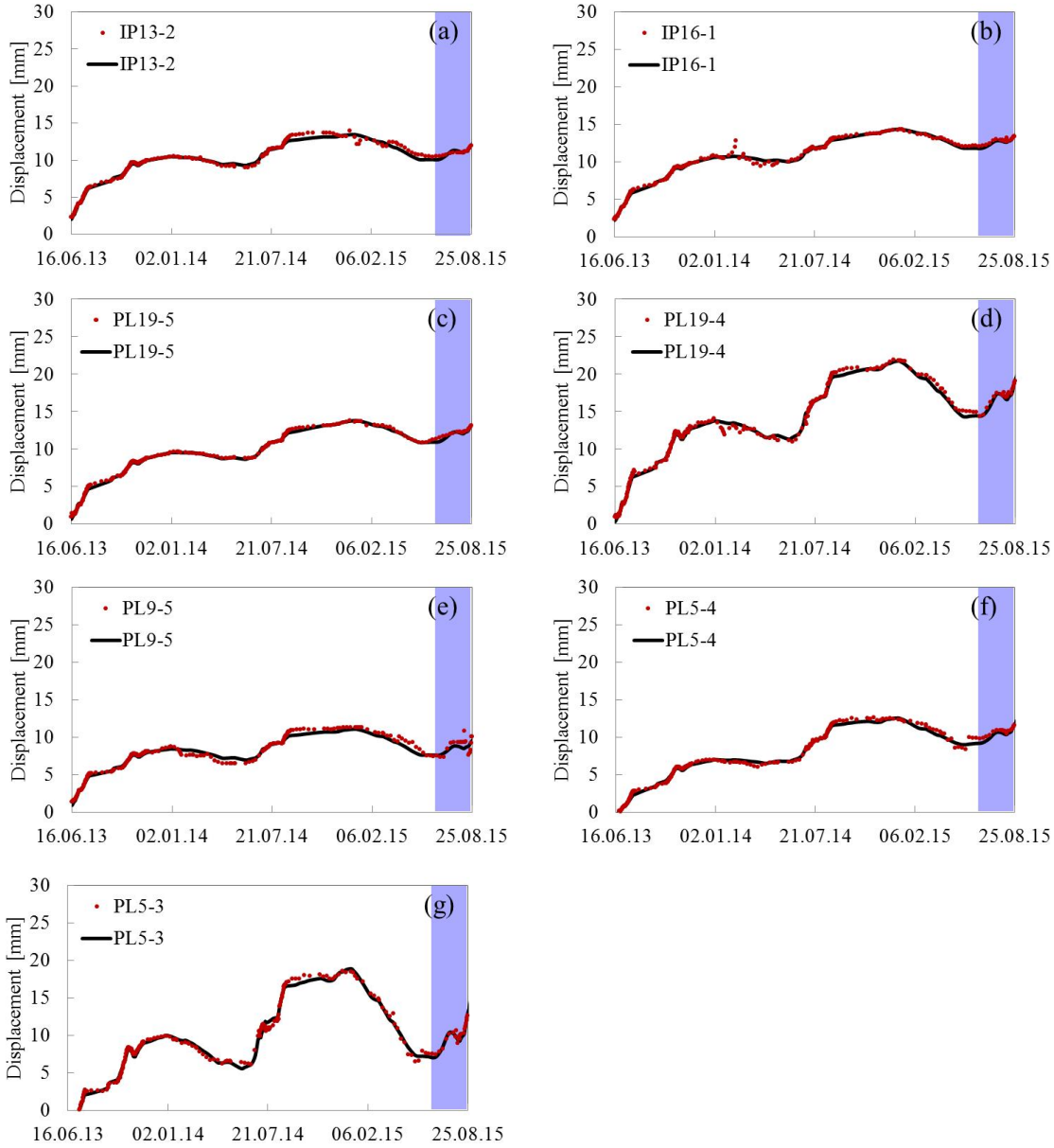

Figure 11. The fitting and predicting results of Class 1: (**a**) IP 13-2; (**b**) IP 16-1; (**c**) PL 19-5; (**d**) PL 19-4; (**e**) PL 9-5; (**f**) PL 5-4; (**g**) PL 5-3. The red dots represent the measured data, and the black lines are fitted and predicted data. The model was developed with the data in the white area and validated by the data in the blue area. The positive direction is downstream.

5.3. Comparison with the Statistical Model

We then compared our model with the statistical model which has been introduced in Section 2. To evaluate the prediction performance of two models, the correlation coefficients R and the residual standard deviation s calculated from the testing dataset are represented in Figure 12. It is obvious that the random coefficient model with ISODATA-GMM clustering has a better performance than the statistical model. A total of 24 measuring points were modelled, and the R of the random coefficient model is larger than that of the statistical model for 22 of them. For the measuring point PL9-3 and PL16-3, the statistical

model performs better than our model. If we take a look on the data series of the fitting results PL9-3 on Figure A3c, the deviation from the predicting data to the measuring data is much more significant than that of other data series. It is interesting to note that, for the statistical model, the values of R for PL16-3 and PL9-3 are lower than other measuring points.

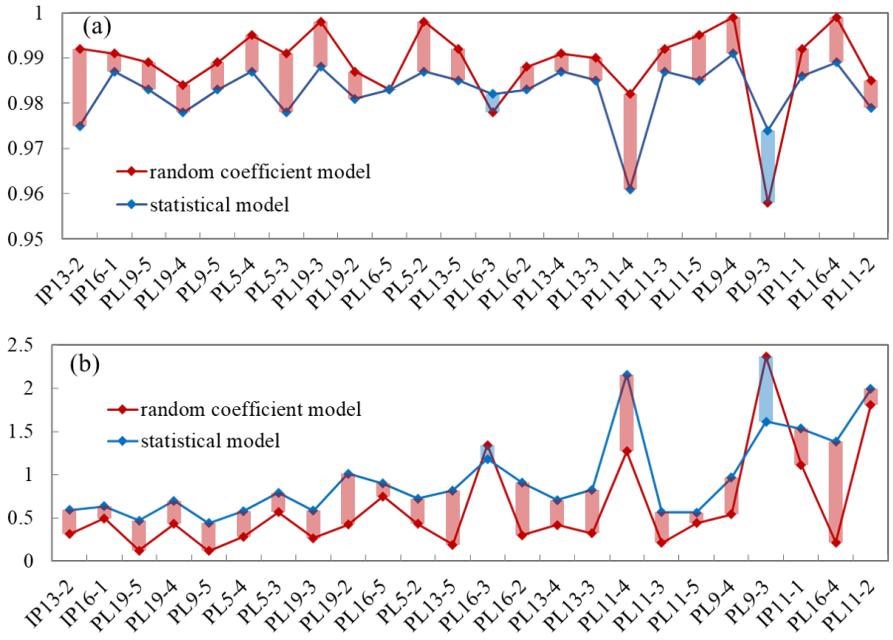

Figure 12. Comparison of (**a**) R and (**b**) s for a random coefficient model and statistical model.

In addition, regarding the statistical model, correlation coefficients R are always above 0.95 for all the measuring points. Hence, even though less accurate than the machine learning model, the accuracy of the statistical model is still validated for predicting the dam displacement in common occasions.

5.4. Limitations

One limitation of the current model is that the fitting and prediction results were vulnerable to noisy data. However, with an effective pre-processing technique to reduce the noisy, this defect could be overcome. Another issue should be noted in the current model is that correctness of monitoring data used for modelling should be guaranteed. Different from establishing models for each measuring point separately, the random coefficient model analyses all the data series in one class simultaneously and considers the relations between data series at different measuring points. Therefore, any noisy or error data series at one measuring point may affect the predicting results of other measuring points.

6. Conclusions

Dam displacement monitoring is one of the most efficient methods to manage and forecast the safety of the dam. As the monitoring points are limited in most dams, researchers and engineers commonly modelled displacement data at different monitoring points individually, and ignored the correlations among each points. However, more and more ultra high dams have been constructed in recent years, in which the uncertainty and multicollinearity increase significantly with the increasing of the number of monitoring points laid in dams.

With the objective of solving the multicollinearity problem in commonly used models, we built a random coefficient model of multi-dimensional data in this paper, which models multi-points simultaneously by making one explanatory variable coefficient at different points following the same asymptotic normal distribution. Measuring points following the same normal distribution are supposed to have similar spatial and temporal characteristics.

The second work is taking the correlations among data at different measuring points into account, by classifying the measuring points with a Gaussian mixture model according to structural attributes and the temporal characteristics. We selected the distance from the measuring point to dam foundation (d) as a spatial indicator and selected the maximum absolute value of the aging sequence (λ), degree of convergence of data series (ξ) as temporal indicators. The Gaussian mixture model has high flexibility (i.e., the shape of the multidimensional Gaussian distribution can fits well the sample points), which may induce high risk of over-fitting and fall into a local optimal solution. To find the optimal solution in a wider space, we introduced the Iterative self-organizing data analysis method to improve the Gaussian mixture model's annealing ability.

In this study, we validated the model using radical displacement data of the concrete arch dam in Jinping-I Hydropower Station as an example. We calculated a dataset of 24 measuring points, and evaluated the model using correlation coefficient (R) and residual standard deviation (s). It turned out that the predicted model fits well with the monitoring data, where the correlation coefficients for all the measuring points are above 0.9. We then compared our model with the statistical model and found that our model has better performance than the statistical model.

Using the clustering algorithm, the correlation between the measuring points can be considered when evaluating displacement of the dam, which significantly improves the accuracy of the prediction model. For the perspectives of the research, besides the dam displacement data, many other kinds of monitoring data exist in hydraulic engineering such as crack monitoring, slope deformation data, etc. In these monitoring projects, correlation exists between measuring points data. Therefore, we expect to apply the model in further structure monitoring projects. In addition, the current model takes the spatial characteristics of the measuring points into account; however, it can not yet predict the data at one point from measured data of adjacent points. Hence, another future orientation can predict the displacement at an arbitrary point on the dam from the measured data at limited monitoring points by combining the finite element method and the prediction model based on the machine learning method.

Author Contributions: Conceptualization, C.S.; methodology, C.S. and Y.H.; software, C.S.; validation, Y.H., C.S. and Z.M.; formal analysis, C.G.; data curation, Y.H. and Z.M.; writing—original draft preparation, C.S.; writing—review and editing, Y.H. and Z.M.; supervision, C.G.

Funding: The research was funded by the National Key R&D Program of China (2016YFC0401601, 2017YFC0804607), the National Natural Science Foundation of China (Grant Nos. 51739003, 51479054, 51779086, 51579086, 51379068, 51579083, 51579085, 51609074), the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (YS11001), the Jiangsu Natural Science Foundation (Grant Nos. BK20160872), the Special Project Funded by the National Key Laboratory (20145027612, 20165042112), the Key R&D Program of Guangxi (AB17195074), and the Central University Basic Research Project (2017B11114).

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

Abbreviations

The following abbreviations are used in this manuscript:

GMM	Gaussian Mixture Model
ISODATA	Iterative Self-Organizing Data Analysis

Appendix A. Fitting and Predicting Results

Figures A1–A4 show the fitting and predicting results of the measuring points in Classes 2, 3, 4 and 5, respectively. It is obvious that the fitting data of the measuring point PL 9-3 have a large derivation from the measuring data (see Figure A3), which results in the *R* of PL 9-3 being the smallest figure.

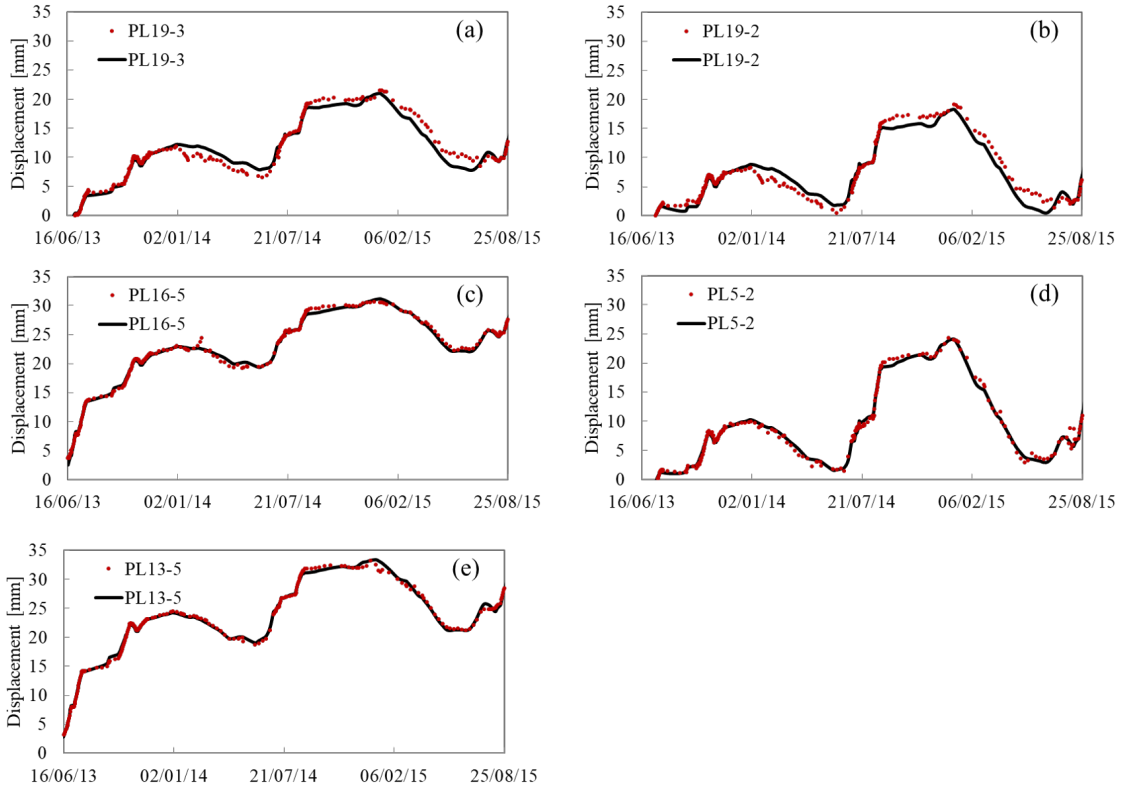

Figure A1. The fitting and forecast results of Class 2: (**a**) IP 19-3; (**b**) IP 19-2; (**c**) PL 16-5; (**d**) PL 5-2; and (**e**) PL 13-5.

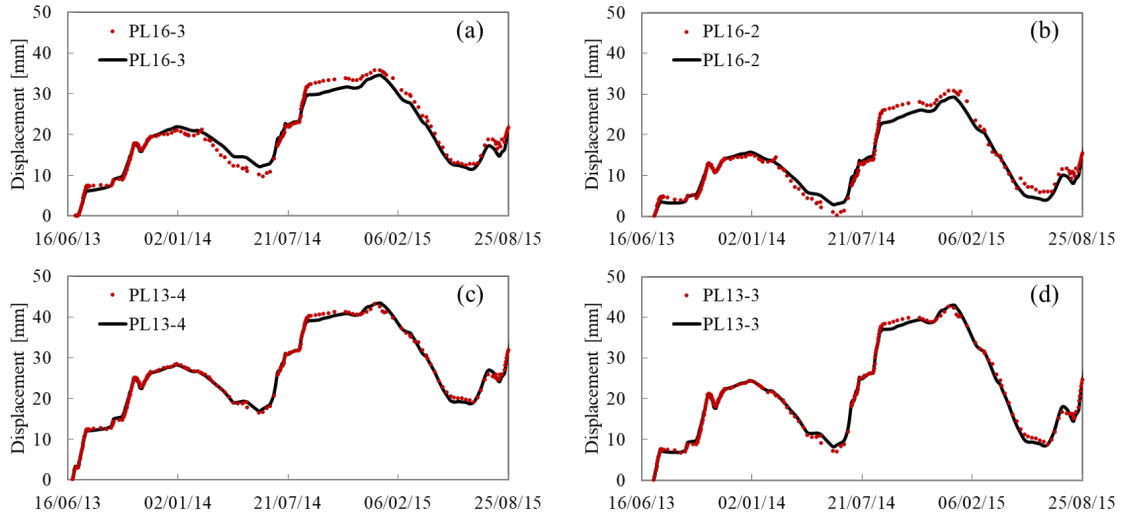

Figure A2. *Cont.*

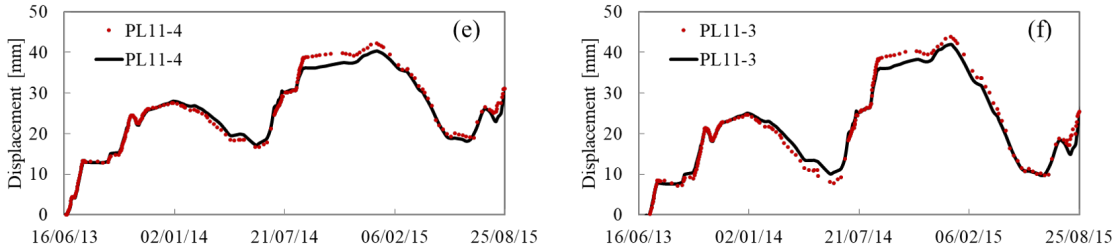

Figure A2. The fitting and forecast results of Class 3: (**a**) IP 16-3; (**b**) IP 16-2; (**c**) PL 13-4; (**d**) PL 13-3; (**e**) PL 11-4; (**f**) PL 11-3.

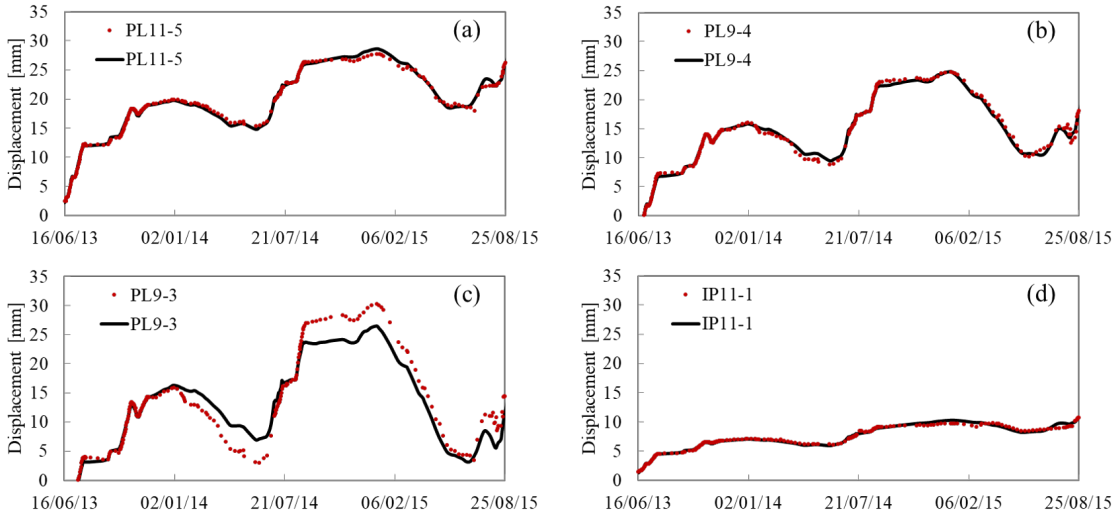

Figure A3. The fitting and forecast results of Class 4: (**a**) IP 11-5; (**b**) IP 9-4; (**c**) PL 9-3; (**d**) PL 11-1.

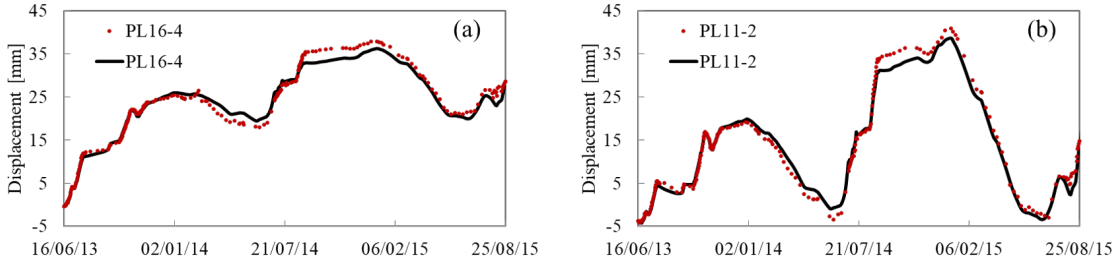

Figure A4. The fitting and forecast results of Class 5: (**a**) IP 16-4; (**b**) IP 11-2.

References

1. Johansson, S. Seepage Monitoring in Embankment Dams. Ph.D. Thesis, Institutionen för anläggning och miljö, Uppsala, Sweden, 1997.
2. Bonaldi, P.; Fanelli, M.; Giuseppetti, G. Displacement forecasting for concrete dams. *Int. Water Power Dam Constr.* **1977**, *29*, 42–45.
3. Brownjohn, J.M. Structural health monitoring of civil infrastructure. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2006**, *365*, 589–622. [CrossRef] [PubMed]
4. Léger, P.; Leclerc, M. Hydrostatic, temperature, time-displacement model for concrete dams. *J. Eng. Mech.* **2007**, *133*, 267–277. [CrossRef]
5. De Sortis, A.; Paoliani, P. Statistical analysis and structural identification in concrete dam monitoring. *Eng. Struct.* **2007**, *29*, 110–120. [CrossRef]

6. Mata, J.; Tavares de Castro, A.; Sá da Costa, J. Constructing statistical models for arch dam deformation. *Struct. Control Health Monit.* **2014**, *21*, 423–437. [CrossRef]
7. 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. [CrossRef]
8. Wang, L.; Zhang, S.C.; Li, Y.H. Application of dynamic gray forecast model in dam deformation monitoring and forecast. *J. Xi'an Univ. Sci. Technol.* **2005**, *3*, 014.
9. Xu, F.; Xu, W. Prediction of displacement time series based on support vector machines-Markov chain. *Rock Soil Mech.* **2010**, *31*, 944–948. [CrossRef]
10. Zhang, H.; Xu, S. Multi-scale dam deformation prediction based on empirical mode decomposition and genetic algorithm for support vector machines (GA-SVM). *Chin. J. Rock Mech. Eng.* **2011**, *30*, 3681–3688.
11. Chen, B.; Hu, T.; Huang, Z.; Fang, C. A spatio-temporal clustering and diagnosis method for concrete arch dams using deformation monitoring data. *Struct. Health Monit.* **2018**. [CrossRef]
12. Wooldridge, J.M. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Rev. Econ. Stat.* **2005**, *87*, 385–390. [CrossRef]
13. Shao, C.; Gu, C.; Yang, M.; Xu, Y.; Su, H. A novel model of dam displacement based on panel data. *Struct. Control Health Monit.* **2018**, *25*, e2037. [CrossRef]
14. Milligan, G.W.; Cooper, M.C. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar. Behav. Res.* **1986**, *21*, 441–458. [CrossRef] [PubMed]
15. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
16. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, *49*, 803–821. [CrossRef]
17. Celeux, G.; Govaert, G. Gaussian parsimonious clustering models. *Pattern Recognit.* **1995**, *28*, 781–793. [CrossRef]
18. Maugis, C.; Celeux, G.; Martin-Magniette, M.L. Variable selection for clustering with Gaussian mixture models. *Biometrics* **2009**, *65*, 701–709. [CrossRef] [PubMed]
19. Wu, Z. *Safety Monitoring Theory and Its Application of Hydraulic Structures*; Higher Education: Beijing, China, 2003.
20. Swamy, P.A. Efficient inference in a random coefficient regression model. *Econom. J. Econom. Soc.* **1970**, *38*, 311–323. [CrossRef]
21. Wu, S.; Shen, M.; Wang, J. Jinping hydropower project: Main technical issues on engineering geology and rock mechanics. *Bull. Eng. Geol. Environ.* **2010**, *69*, 325–332.
22. Xu, N.; Tang, C.; Li, L.; Zhou, Z.; Sha, C.; Liang, Z.; Yang, J. Microseismic monitoring and stability analysis of the left bank slope in Jinping first stage hydropower station in southwestern China. *Int. J. Rock Mech. Min. Sci.* **2011**, *48*, 950–963. [CrossRef]