

Received July 14, 2020, accepted July 30, 2020, date of publication August 7, 2020, date of current version August 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014919

# Improve the Model Stability of Dam's Displacement Prediction Using a Numerical-Statistical Combined Model

YATING HU<sup>1,2,3</sup>, CHONGSHI GU<sup>1,2</sup>, ZHENZHU MENG<sup>1,3</sup>, AND CHENFEI SHAO<sup>1,2</sup>

<sup>1</sup>College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China

<sup>2</sup>National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Nanjing 210098, China

<sup>3</sup>Laboratory of Environmental Hydraulics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Corresponding author: Chongshi Gu (csgu@hhu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0401601 and Grant 2017YFC0804607; in part by the Fundamental Research Funds for the Central Universities under Grant 2018B619X14; in part by the National Natural Science Foundation of China under Grant 51739003, Grant 51479054, Grant 51779086, Grant 51579086, Grant 51379068, Grant 51579083, Grant 51579085, and Grant 51609074; in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX18\_0588; in part by the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions under Grant YS11001; in part by the Jiangsu Natural Science Foundation under Grant BK20160872; in part by the Special Project Funded of National Key Laboratory under Grant 20145027612 and Grant 20165042112; in part by the Key Research and Development Program of Guangxi under Grant AB17195074; and in part by the Central University Basic Research Project under Grant 2017B11114.

**ABSTRACT** In most studies of dam's displacement prediction based on monitoring data, emphasis was given on improving the prediction accuracy, while the model stability was merely considered. This study proposed a numerical-statistical combined model which aims to improve the model stability. The displacement was modelled within three modules: recoverable displacement (i.e., displacement induced by the external load including the water pressure and temperature), non-recoverable displacement (i.e., displacement due to the inherent variations of the materials such as the creep and fatigue of the concrete), and measurement errors (i.e., instrument error and human error). To reduce the random errors and increase the model stability, we used the numerical simulation to constrain the coefficients of explanatory variables for the recoverable displacement. The non-recoverable displacement was estimated by empirical equations, and the measurement errors were given by Gaussian distributions. The randomness of coefficients in the model among all monitoring points are constrained further by random coefficient model. We adopted the root mean square error (RMSE) at varying time and the change ratio of the coefficients (CRC) to evaluate the model stability. Results indicated that the proposed model not only has better prediction accuracy but also has better model stability compared with the statistical model and coordinates-included statistical model proposed in previous studies.

**INDEX TERMS** Displacement prediction, dam, combined model, model stability, numerical simulation, statistical model.

## I. INTRODUCTION

In most early studies of dam's displacement prediction based on monitoring data, researchers used statistical models to estimate the future displacement from the past monitoring data, in which the displacement was quantified by three

influencing factors, i.e., hydrostatic pressure, temperature, and ageing [1]–[4]. With the development of computational technologies, machine learning methods were introduced to the field, and more and more complicate influencing factors were taken into consideration. Such methods include the artificial neural network method [5]–[8], support vector machine method [9], [10], extreme learning machine method [11], and etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Vlad Diaconita<sup>1</sup>.

Both statistical models and machine learning methods have shown very high prediction precisions. However, the model stability was rarely discussed, which is equally as important as the prediction precision, especially when a long-term prediction is involved [12]–[14]. Many unquantifiable factors such as the construction quality of pouring and material properties such as compression strength, elastic modulus, and Poisson ratio may induce uncertainties in predicting the displacement. These factors depend on the spatial positions of the monitoring points on the dam.

In order to enhance the model stability, recent studies have integrated the spatial correlations of the monitoring points into the statistical models, by classifying the monitoring data at different monitoring points into several groups [15], [16]. In the statistical models with monitoring data being classified, the spatial correlations were quantified by groups, whereas the correlations between monitoring points in one group were lacking. To reflect the overall spatial correlations, one method is to integrate the coordinates of monitoring points into the statistical models as explanatory variables [17]. The coordinates-included statistical model considered the spatial correlations between each monitoring point; however, the accuracy of the model was usually reduced due to the increasing number of explanatory variables and model complexity. Taking a simple power function as an example, the number of explanatory variables of the function that includes the coordinates  $(x, y, z)$  is increased by  $4^3$  times compared with that without coordinates. As the number of variables increases, the coordinates-included model exhibits an increasing possibility of autocorrelations between variables, which would weaken the prediction accuracy. The objective of the present study is to enhance the stability of the prediction model without increase the number of variables like coordinates-included statistical models.

In this study, we proposed a numerical-statistical combined model, which modelled the displacement via three modules: the recoverable displacement, non-recoverable displacement and measurement errors. The recoverable displacement is the most critical module, and it represents the displacement induced by the external loads including the water pressure and temperature. This component was quantified by numerical simulation [18]–[20] with reduced explanatory variables. The non-recoverable displacement includes the displacement resulting from the inherent variations of the materials such as the creep and fatigue of the concrete, and it was provided with an empirical formula. The measurement errors (i.e., instrument error and human error) were estimated with Gaussian distributions. To constrain the explanatory variables of the proposed model further, a statistical model called random coefficient model [21]–[23] were used to obtain the coefficients. To evaluate the proposed model, we first quantified the model's prediction accuracy for data of six different prediction periods varied from one month to six months. We then quantified the model stability with an indicator named change ratio of coefficient (CRC), which estimates the fluctuation of the coefficients of the explanatory variables.

This article was organized as follows. Section II presented how the proposed model was developed. We first introduced the statistical model and the coordinates-included statistical model, and then developed the statistical-numerical combined model. Section III exhibited the engineering project and the dataset. Section IV discussed the results including the prediction accuracy and the model stability. Concluding remarks completed the paper in Section V.

## II. MODEL DEVELOPMENT

### A. COORDINATES-INCLUDED STATISTICAL MODEL

Dam's displacement depends on many external and internal factors, such as the external load, the material properties of the dam body and the dam foundation, and the quality of the construction. Researchers developed statistical models to quantify the influence of these factors on dam's displacement. In most statistical models, the calculation of dam's displacement  $\delta$  were divided into three modules: hydraulic components  $\delta_H$ , temperature components  $\delta_T$ , and ageing components  $\delta_\theta$ .

$$\delta = \delta_H + \delta_T + \delta_\theta \quad (1)$$

The displacement  $\delta$  of a single monitoring point was written as:

$$\delta = a_0 + \sum_{i=1}^n a_i H^i + \sum_{i=1}^{m_3} (b_{1i} \sin \frac{2\pi it}{365} + b_{2i} \cos \frac{2\pi it}{365}) + c_1 \theta + c_2 \ln \theta \quad (2)$$

where  $a_0$  is a constant term which represents the initial conditions;  $H$  is the upstream water level;  $n$  is a coefficient with  $n=3$  for gravity dams and  $n=4$  for arch dams;  $t$  counts time from the starting date of the selected dataset;  $\theta = \frac{t}{100}$ ;  $a_i, b_i, c_i$  are unknown coefficients.

To consider the spatial correlations among all monitoring points, the coordinates of monitoring point were included as variables in the statistical model (see Figure 1). Then, the displacement  $\delta$  becomes:

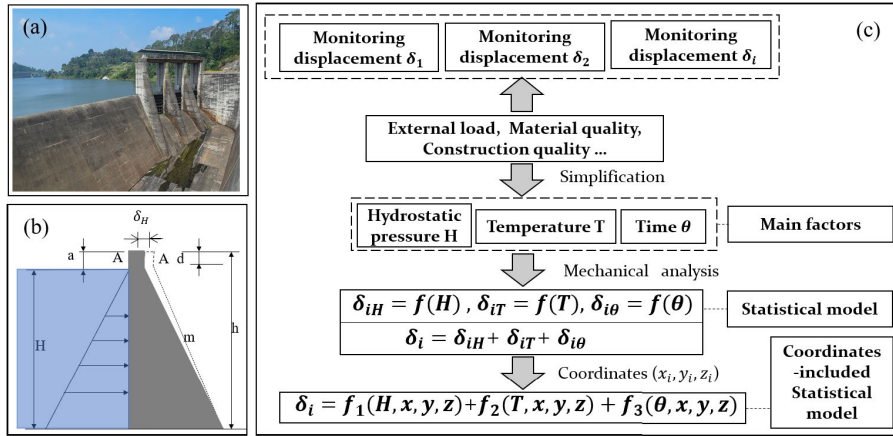
$$\delta = f(H, T, \theta, x, y, z) = f_1(H, x, y, z) + f_2(T, x, y, z) + f_3(\theta, x, y, z) \quad (3)$$

with  $H$  the upstream water level,  $T$  the temperature,  $\theta = \frac{t}{100}$ ,  $t$  the time, and  $x, y, z$  the coordinates of the monitoring points.  $f_1(H, x, y, z), f_2(T, x, y, z)$ , and  $f_3(\theta, x, y, z)$  are related to  $\delta_H, \delta_T$ , and  $\delta_\theta$ , respectively (see Equation(1)).

$$f_1(H, x, y, z) = f_1[f(H), f(x, y, z)] = f[\sum_{i=1}^{3(4)} a_i H^i, \sum_{l,m,n=0}^3 a_{lmn} x^l y^m z^n] \quad (4)$$

where  $f(H)$  is the hydrostatic pressure component at one monitoring point as expressed in Equation 2.  $f_1(x, y, z)$  represents the displacement field induced by the water load, and it can be expressed with a multivariate power function:

$$f_1(H, x, y, z) = \sum_{k=1}^{3(4)} \sum_{l,m,n=0}^3 A_{klmn} H^k x^l y^m z^n \quad (5)$$



**FIGURE 1.** (a) Photograph of a gravity dam, (b) sketch of the gravity dam, (c) flowchart of the statistical model and the coordinates-included statistical model.

The displacement field induced by the temperature component  $f_2(T, x, y, z)$  exhibits:

$$\begin{aligned}
 f_2(T, x, y, z) &= f_2[f(T), f(x, y, z)] \\
 &= \sum_{j,k=0}^1 \sum_{l,m,n=0}^3 B_{jklmn} \sin \frac{2\pi jt}{365} \cos \frac{2\pi kt}{365} x^l y^m z^n
 \end{aligned} \tag{6}$$

The displacement field induced by the ageing component  $f_3(x, y, z)$  is:

$$\begin{aligned}
 f_3(\theta, x, y, z) &= f_3[f(\theta), f(x, y, z)] \\
 &= \sum_{l,m,n=0}^3 C_{jklmn} \theta^j \ln \theta^k x^l y^m z^n \\
 &\quad (j = 0, k = 1 \text{ or } j = 1, k = 0)
 \end{aligned} \tag{7}$$

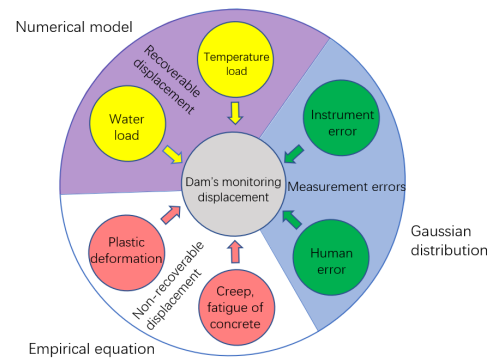
Then, the coordinates-included statistical model can be expressed as:

$$\begin{aligned}
 \delta &= \sum_{k=1}^{3(4)} \sum_{l,m,n=0}^3 A_{klmn} H^k x^l y^m z^n \\
 &+ \sum_{j,k=0}^1 \sum_{l,m,n=0}^3 B_{jklmn} \sin \frac{2\pi jt}{365} \cos \frac{2\pi kt}{365} x^l y^m z^n \\
 &+ \sum_{l,m,n=0}^3 C_{jklmn} \theta^j \ln \theta^k x^l y^m z^n
 \end{aligned} \tag{8}$$

Known the spatial coordinates  $(x, y, z)$  and displacement  $\delta(H, T, \theta, x, y, z)$  of each monitoring point, we can fit the  $A_i, B_i, C_i$  with the monitoring data using a stepwise regression method.

**B. NUMERICAL-STATISTICAL COMBINED MODEL**

With the coordinates  $(x, y, z)$  integrated into the statistical model, the numbers of independent variables are increased. The increasing variables may lead to an auto-correlation problem. The proposed numerical-statistical combined model



**FIGURE 2.** Framework of the proposed numerical-statistical combined model.

aims to increase the model stability without increase the number of independent variables. In contrast to previous studies which divided the displacement into water pressure component, temperature component, and ageing component, we model the displacement within three modules: recoverable displacement  $\delta_{i,r}$ , non-recoverable displacement  $\delta_{i,n-r}$ , and measurement errors  $\delta_{i,e}$  (see Figure 2). The recoverable displacement is estimated by integrating the numerical simulation into the statistical model, the non-recoverable displacement is evaluated with an empirical equation, and the measurement errors is assumed to follow a Gaussian distribution.

**1) RECOVERABLE DISPLACEMENT**

The module of the recoverable displacement  $\delta_{i,r}$  represents the dam's displacement induced by the external loads including the water pressure  $\delta_{iH}$  and temperature  $\delta_{iT}$ .  $\delta_{iH}$  and  $\delta_{iT}$  are regarded as linear elastic and satisfies small deformation assumption, they hence can be expressed by the equilibrium equation (9), geometric equation (10), and constitutive equation (11).

$$\sigma_{ij,j} + f_i = 0 \tag{9}$$

$$\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}) \quad (10)$$

$$\varepsilon_{ij} = \frac{1 + \mu}{E} \sigma_{ij} - \frac{\mu}{E} \delta_{ij} \sigma \quad (11)$$

where  $\sigma$ ,  $\varepsilon$  are the stress tensor and strain tensor, respectively;  $\mathbf{u}$  is the theoretical displacement field;  $\mathbf{f}$  is the volume force and  $\delta_{ij}$  is the Kronecker symbol,  $\delta_{ij} = 0$  when  $i \neq j$  and  $\delta_{ij} = 1$  when  $i = j$ .

The  $\delta_{iH}^0$  at different water levels and  $\delta_{iT}^0$  at different temperatures can be solved with Equation (9) to (11) using a finite element method. The relation between the designed and simulated displacement field  $\delta_{iH}$  and  $\delta_{iH}^0$  are:

$$\delta_{iH} = \frac{E_c^0}{E_{ic}} \cdot \delta_{iH}^0 = \zeta_i \cdot \delta_{iH}^0 \quad (12)$$

where  $i$  is the serial number of monitoring points,  $E_c^0$  is the designed elastic modulus,  $E_{ic}$  is the actual elastic modulus, and  $\zeta_i$  is the ratio of the designed elastic modulus  $E_c^0$  to the actual elastic modulus  $E_{ic}$ .

Similarly, the relation between the actual linear expansion  $\delta_{iT}$  and the designed linear expansion  $\delta_{iT}^0$  can be expressed as  $\delta_{iT} = \xi_i \cdot \delta_{iT}^0$ , in which  $\xi_i$  is the ratio of the designed linear expansion to the actual linear expansion. The displacement of the recoverable module exhibits:

$$\delta_{i,r} = \zeta_i \cdot \delta_{iH}^0 + \xi_i \cdot \delta_{iT}^0 \quad (13)$$

$\delta_{iH}^0$  and  $\delta_{iT}^0$  are provided by the numerical simulation,  $\zeta_i$  and  $\xi_i$  represent the fluctuations of the actual elastic modulus and the actual linear expansion coefficients at indicated position.

### 2) NON-RECOVERABLE DISPLACEMENT

The non-recoverable displacement field  $\delta_{i,n-r}$  represents the displacement that results from the inherent variations of the materials such as plasticity, creep, fatigue of concrete, and etc. The physical mechanism of these influencing factors are unclear until now. Here, we use the linear function and the logarithmic function to characterize the divergence trend and the convergence trend of  $\delta_{i,n-r}$ , respectively. The expression of  $\delta_{i,n-r}$  is exhibited as:

$$\delta_{i,n-r} = \sum_{i=0}^L \sum_{m=0}^M (d_{1lm}\theta + d_{2lm}\ln\theta)x_i^l z_i^m \quad (14)$$

where  $x$  is the horizontal coordinate and  $z$  is the vertical coordinate,  $\theta$  is the time,  $d_{1lm}$  and  $d_{2lm}$  are pending coefficients.

### 3) MEASUREMENT ERRORS

The measurement errors  $\delta_{i,e}$  include the instrument errors and the human errors. The measurement errors of the displacement data at a monitoring point can be regarded as satisfying a Gaussian distribution  $N(0, \sigma^2)$ . The probability distribution function of  $\delta_{i,e}$  is:

$$f(\delta_{i,e}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\delta_{i,e}^2}{2\sigma^2}\right) \quad (15)$$

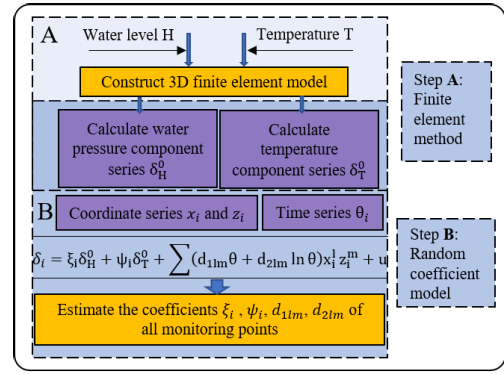


FIGURE 3. Flowchart of the proposed numerical-statistical combined model.

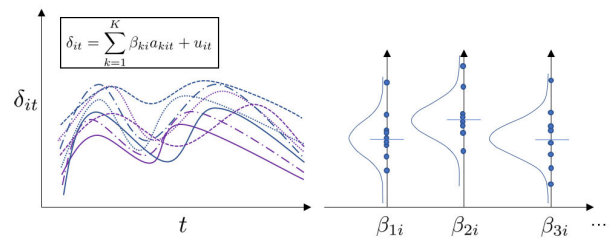


FIGURE 4. Sketch of the random coefficient model.

### 4) SOLVING THE MODEL

As presented in Section II-B1 to II-B3, the displacement field can be exhibited as:

$$\begin{aligned} \delta_i &= \delta_{i,r} + \delta_{i,n-r} + \delta_{i,e} \\ &= \zeta_i \cdot \delta_{iH}^0 + \xi_i \cdot \delta_{iT}^0 + \sum_{i=0}^L \sum_{m=0}^M (d_{1lm}\theta + d_{2lm}\ln\theta)x_i^l z_i^m + u_i \end{aligned} \quad (16)$$

Figure 3 illustrates the flowchart of the proposed numerical-statistical combined model.

With known spatial coordinates of the monitoring points, the geometric characteristics at each monitoring point can be determined from the numerical simulation. The advantage of the numerical method is that only two variables ( $\zeta_i$  and  $\xi_i$ ) are left in the module of recoverable displacement.

We use the random coefficient model to solve the numerical-statistical combined model, which assumes the regression coefficients are random variables and obey a Gaussian distribution (See Figure 4).

By introducing the random effects, the correlation between individual observations are taken into account, and the degree of freedom of the model are reduced. Then, Equation 16 can be exhibited as:

$$\delta_{it} = \sum_{k=1}^K \beta_{ki} a_{kit} + u_{it} = \sum_{k=1}^K (\beta_k + \gamma_{ki}) a_{kit} + u \quad (17)$$

where  $\delta_{it}$  is a two-dimensional data panel of the displacements containing temporal and spatial information;  $a_{kit}$  is



FIGURE 5. Photo of the selected dam.

a two-dimensional data panel of explanatory variables;  $t$  is time;  $i$  is the dam's cross-section index;  $k$  is the explanatory variables index and  $u$  is a random term. The pending coefficient  $\beta_{ki}$  includes  $\beta_k$  and  $\gamma_{ki}$ , with  $\beta_k = (\beta_1, \dots, \beta_K)'$  the common mean coefficient vector, and  $\gamma = (\gamma_{1i}, \dots, \gamma_{Ki})'$  the derivation from the individual data to the common mean value. According to the central limit theorem,  $\beta_{ki}$  approximately obeys an asymptotic Gaussian distribution.

III. DATASETS

In the present study, we used the monitoring displacement data of the concrete arch dam at the Jinping-I hydropower station, which is one of the highest concrete arch dam in the world (see Figure 5). The elevations of the crest and foundation of the dam are 1885m and 1580 m, respectively. The normal impounded water level and the level of dead water are 1880m and 1800m, respectively.

For the dataset, we selected the radial displacement monitoring data of 23 monitoring points distributed in six plumb lines (5#, 9#, 11#, 13#, 16# and 19#) from July 1, 2015 to December 31, 2018. Figure 6 shows the distribution of the monitoring points. Displacement to the downstream direction counts for positive, and the displacement to the upstream direction counts for negative. The displacement data were recorded once a day. After eliminated the miss-data, we obtained 914 validated data in total. The dataset was divided into two parts: data from July 1, 2015 to June 30, 2018 were selected as training dataset, whereas data from July 1, 2018 to December 31, 2018 were selected as testing dataset.

Figure 7 shows the time variation of the upstream water level and the monitoring data of the selected monitoring points.

IV. RESULTS AND DISCUSSIONS

A. RESULTS OF THE PROPOSED MODEL

We analysed the module of the recoverable displacement in ABAQUS. We firstly established a three-dimensional finite element model for the selected dam, which consists of the dam body, dam pedestal, and the surrounded mountain. The dam body contained 38537 elements and 31941 nodes.

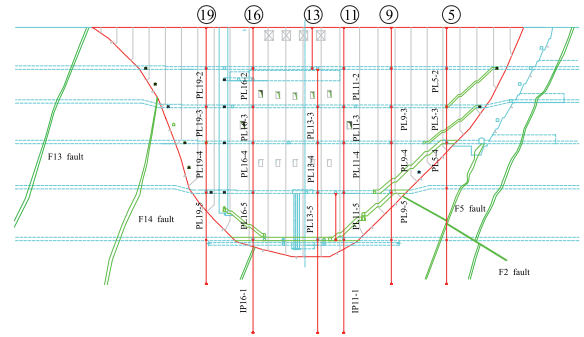


FIGURE 6. The distribution of monitoring points (view from the upstream side).

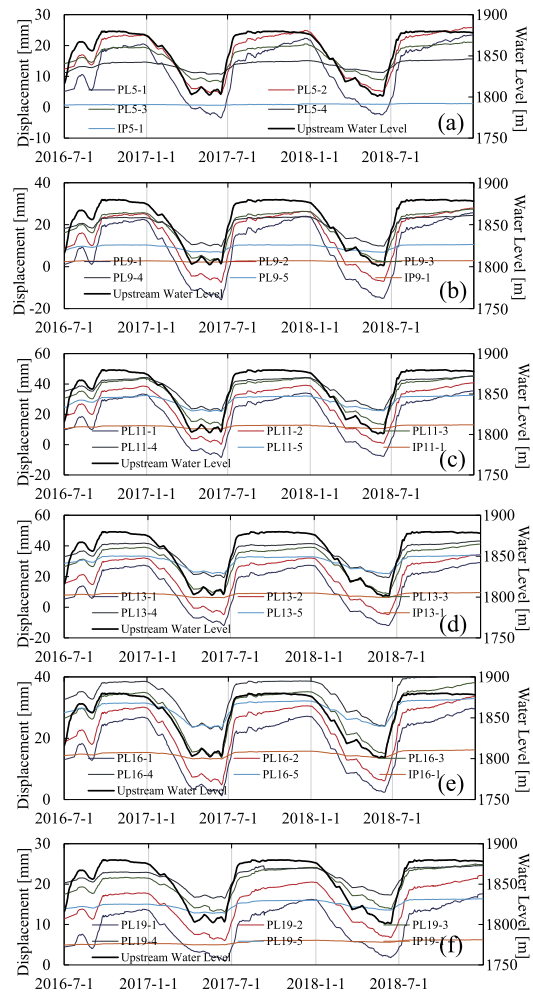
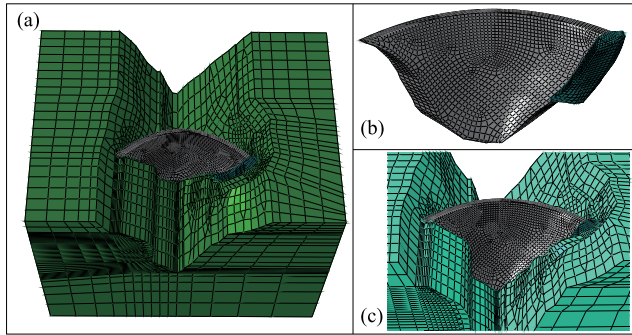


FIGURE 7. Time variation of the upstream water level and monitoring data of the monitoring points at different plumb lines: (a) 5#, (b) 9#, (c) 11#, (d) 13#, (e) 16#, (f) 19#.

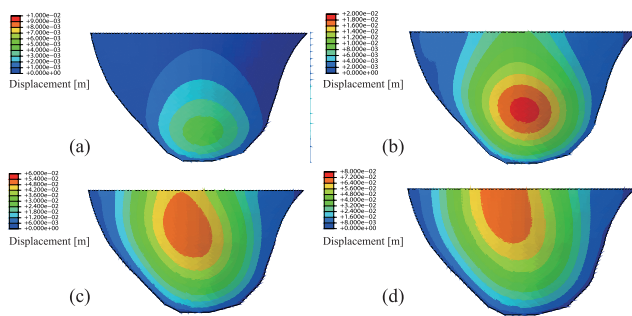
The model was constrained in the normal direction for all lateral boundaries, and was fixed in all directions for the bottom boundary (see Figure 8). The parameters of the material properties used the designed value (see Table 1). The density, Young's modulus and Poisson ratio were used to calculate the hydrostatic pressure-induced displacement,



**FIGURE 8.** The finite element model of the selected dam: (a) the overall structure, (b) the dam body and dam pedestal, (c) the dam body and the surrounded mountain.

**TABLE 1.** The settings of material properties in the finite element model.

	Density [kg/m <sup>3</sup> ]	Young's modulus[GPa]	Possion ratio [I]	Expansion coefficient [1/°C]
Concrete	2400	30	0.2	10 <sup>-5</sup>
Rock	2400	20	0.2	0

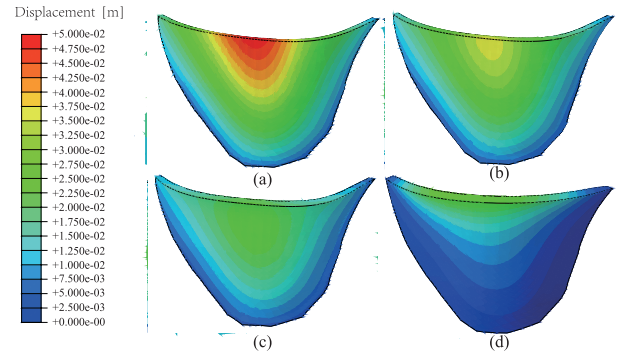


**FIGURE 9.** The simulated radial displacement at different upstream water levels: (a) 1700 m, (b) 1780 m, (c) 1860 m, (d) 1880 m.

and the expansion coefficient was used to calculate the temperature-induced displacement. We simplified the dam body as concrete, and the dam pedestal and foundation as rock.

As the upstream water level varied between 1700 m to 1880 m in the real world, we simulated the dam's displacement at six different upstream water levels (see Figure 9). Results showed that the displacement field was approximately horizontal symmetry, with the displacement at the midline larger than at the border areas.

Similarly, as the temperature in the real world varied from 4 °C to 24 °C, we selected 4°C, 8°C, 12°C, 16°C, 20°C and 24°C as the boundary conditions of the downstream dam surface and the dam crest. The boundary conditions of the upstream dam surface were set to 3°C, 4°C, 5°C, 8°C, 11°C and 14°C, which are the average temperatures of the water body in the above six configurations. Figure 10 exhibits the displacement field with the temperature of the upstream dam surface varied systematically from 4°C to 24°C. The dam's crest deforms toward downstream at low temperature and toward upstream at high temperature.



**FIGURE 10.** The simulated radial displacement at different upstream dam surface temperatures: (a) 4°C, (b) 12°C, (c) 20°C, (d) 24°C.

**TABLE 2.** The coordinates of the selected monitoring points.

Monitoring point	X [m]	Z [m]
IP11-1	996.99	1601
PL11-2	997.37	1830
PL11-3	996.94	1778
PL11-4	996.02	1730
PL11-5	996.46	1664.25
PL13-3	1029.84	1778
PL13-4	1029.58	1730
PL13-5	1029.42	1664.25
IP16-1	1112.39	1601
PL16-2	1111.76	1830
PL16-3	1112.34	1778
PL16-4	1113.13	1730
PL16-5	1112.7	1664.25
PL19-2	1170.83	1830
PL19-3	1169.89	1778
PL19-4	1169.34	1730
PL19-5	1168.19	1664.25
PL5-2	874.64	1830
PL5-3	880.35	1778
PL5-4	882.56	173
PL9-3	936.98	1778
PL9-4	943.41	1730
PL9-5	942.3	1664.25

With the numerical simulations of the configurations at different water level and temperature, we obtained the regression relationship between  $H$  and  $\delta_H^0$  (i.e.,  $\delta_H^0 = \sum_{i=0}^4 a_i H^i$ ), and the relation between  $T$  and  $\delta_T^0$  (i.e.,  $\delta_T^0 = \sum_{i=0}^2 b_i T^i$ ). Table 2 exhibits the coordinates of all the selected monitoring points.

With the random coefficient model, the coefficients of the explanatory variables in Equation 16 can be obtained, which are exhibited in Table 3.

### B. PREDICTION ACCURACY

In order to evaluate the proposed model, we calculated the coefficients of determination  $R^2$  of the training data and the root of mean square error RMSE of the testing dataset:

$$R^2 = \frac{\sum_{i=1}^n (\hat{\delta}_i - \bar{\delta}_i)^2}{\sum_{i=1}^n (\delta_i - \bar{\delta}_i)^2} \quad (18)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{\delta}_i - \bar{\delta}_i)^2}{n}} \quad (19)$$

**TABLE 3.** The coefficients of independent variables of the proposed model at each monitoring point.

Monitoring point	$\xi_i$	$\psi_i$	$d_1$	$d_2$	cons
IP11-1	0.725	0.07	0.519	0.545	6.341
PL11-2	17.177	2.264	4.966	-6.59	-50.628
PL11-3	5.329	1.025	5.46	-10.325	-8.998
PL11-4	1.276	0.211	4.077	-7.827	5.821
PL11-5	0.509	0.074	0.705	0.481	10.835
PL13-3	2.346	0.431	5.509	-11.418	-15.66
PL13-4	11.305	1.072	3.978	-8.64	-1.222
PL13-5	2.415	0.417	2.192	-5.099	21.365
IP16-1	0.715	0.07	1.345	-0.958	10.262
PL16-2	1.085	0.384	7.427	-10.773	-2.392
PL16-3	0.58	0.142	7.164	-10.844	7.853
PL16-4	0.421	0.078	6.255	-9.138	18.43
PL16-5	0.346	0.051	2.804	-3.764	19.894
PL19-2	1.644	0.362	5.168	-2.856	-2.595
PL19-3	0.622	0.187	4.755	-3.286	12.438
PL19-4	0.188	0.037	2.895	-2.975	15.93
PL19-5	0.054	0.01	2.008	-0.937	11.984
PL5-2	0.399	0.109	4.131	-4.092	-2.787
PL5-3	0.352	0.09	3.476	-4.636	5.904
PL5-4	0.314	0.097	1.166	0.29	8.8
PL9-3	0.903	0.245	5.183	-12.304	11.002
PL9-4	0.236	0.052	3.027	-7.092	11.862
PL9-5	0.095	0.018	0.23	-0.488	6.586

**TABLE 4.** The coefficient of determination  $R^2$  of the coordinates-included statistical model (C-S model) and the statistical model (S model).

Monitoring point	C-S model	S model	the proposed model
IP11-1		0.9951	0.9939
PL11-2		0.9834	0.9851
PL11-3		0.9956	0.9911
PL11-4		0.9942	0.9939
PL11-5		0.9939	0.9943
PL13-3		0.9868	0.9854
PL13-4		0.9954	0.9943
PL13-5		0.9854	0.9845
IP16-1		0.9947	0.9935
PL16-2		0.9949	0.9942
PL16-3		0.9848	0.9849
PL16-4	0.9718	0.9947	0.9944
PL16-5		0.994	0.9954
PL19-2		0.9909	0.9872
PL19-3		0.9881	0.9885
PL19-4		0.9908	0.9905
PL19-5		0.9871	0.9851
PL5-2		0.9966	0.9943
PL5-3		0.9853	0.9864
PL5-4		0.9954	0.9964
PL9-3		0.9948	0.9933
PL9-4		0.9928	0.9936
PL9-5		0.9978	0.9929

where  $\bar{\delta}_i$  is the average of the monitoring data and  $n$  is the number of displacement data.

As shown in Table 4, we compared the coefficient of determination  $R^2$  of the proposed model with the statistical model (S model) and coordinates-included statistical model(C-S model). The coefficients of the explanatory variables in the S model and C-S model are shown in Appendix (Table 6 and 7). The number of variables of a single monitoring points of the S model, C-S model, and the proposed model were 9, 160, and 5, respectively. For the displacement prediction of single monitoring point, the number of variables of the C-S model was more than those of the S model and the proposed model. Here, to establish model for all monitoring points, the number of variables were 207, 160 and 115 for the S model, C-S model, and the proposed model, respectively. For the whole dataset, S model had the most numerous variables (207) and thus had the best fitting results but also a larger possibility of having the over-fitting problem. The  $R^2$  of the C-S model was slightly smaller than those of all monitoring points in S model and the proposed model. Overall, the coefficient of determination  $R^2$  of all these three models exceeded 0.95 for the prediction of all the monitoring points, which means all these models performed well in fitting the training data.

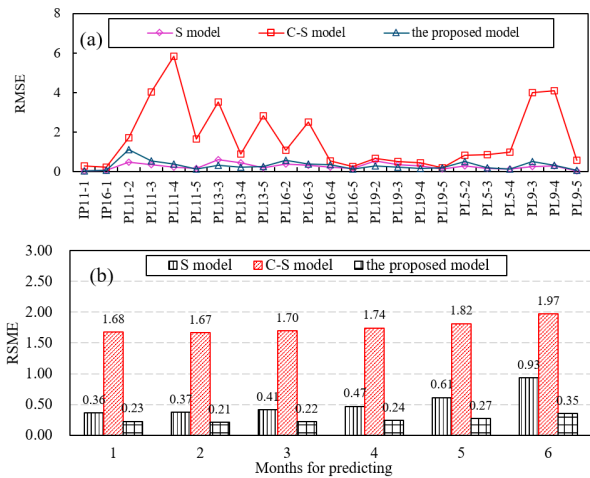
Figure 11 (a) presents the RMSE of three models for each monitoring point, the average RMSE of all monitoring points were 0.315, 1.679 and 0.270 for the S model, the C-S model, and the proposed model, respectively. For the whole 23 monitoring points, S model had the lowest RMSE in 8 monitoring points, the proposed model had the lowest RMSE in 15 monitoring points. Since more than half of the variables in the C-S model were fairly small (almost close to zero), the number of effective variables is the smallest.

Therefore, C-S model performed the worst of three models in fitting and predicting displacement. The prediction accuracy of the proposed model was as good as S model, because it considered the deterministic relation between the variables and the dam's displacement using the numerical simulation.

Further, to evaluate the prediction accuracy at different time period, we calculated the RMSE from one to six months for the testing dataset. Figure 11 (b) exhibits the average RMSE of 23 monitoring points, which varied from 1.67 to 1.97 for the C-S model, from 0.36 to 0.93 for the S model, and from 0.21 to 0.35 for the proposed numerical-statistical combined model. For all these three models, the average RMSE of all monitoring points kept increasing with the prediction period last longer. When the prediction time rose from five months to six months, the RMSE increased intensely. Compared with the other two models, the S model had a more obvious increase of RMSE, especially at the sixth month, the increase rate was 52.4 % (increased from 0.61 for five months' prediction time to 0.93 for six months). The average RMSE of the C-S model for all predicting period was the highest but its increment was smaller than the S model. To conclude, the prediction accuracy of all these three models decline when the predicting time lasts longer. By comparing the RMSE and its increments for the prediction time varied from one month to six months, we noticed that the proposed model had the most steady prediction accuracy. This is because the proposed model limits the variables with random coefficient model.

### C. MODEL STABILITY

Here, we select change ratio of coefficient (CRC) as the indicator to evaluate the model stability, which represents the



**FIGURE 11.** The RMSE of the S model, the C-S model and the proposed model (a) at different monitoring point and (b) with different prediction time.

sensitivity of the coefficients of the model to the varied inputs. The expression of CRC is exhibited as:

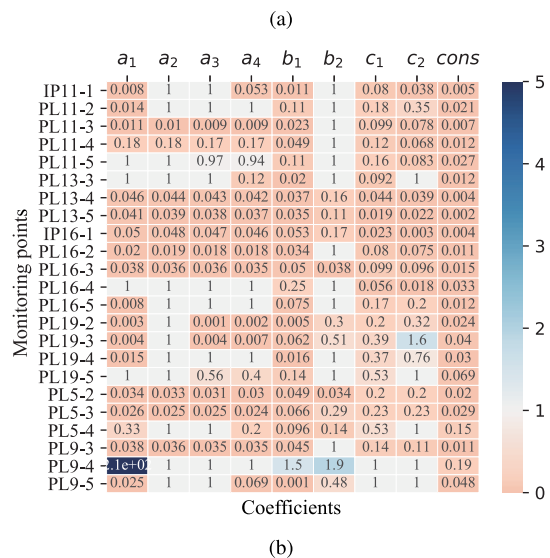
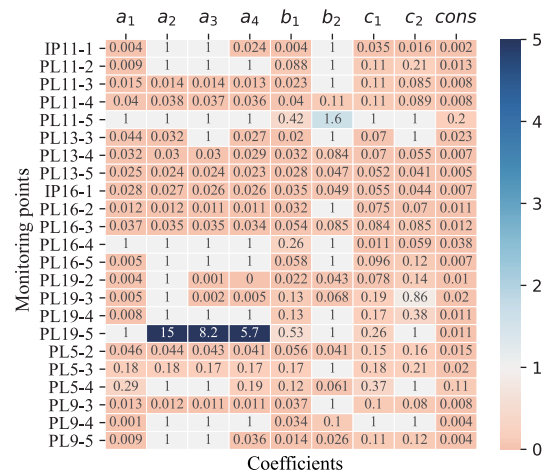
$$CRC = \frac{|Coef_{ini} - Coef'|}{|Coef_{ini}|} \quad (20)$$

where CRC denotes the change ratio of coefficient,  $Coef_{ini}$  and  $Coef'$  are the coefficients calculated with the initial inputs and varied inputs, respectively. A larger CRC signifies that the coefficients calculated with varied inputs have greater fluctuation, thus a more unstable model.

We adopted the varied training datasets and varied measurement errors ( $\delta_{i,e}$ ) to obtain the CRC in this section. To calculate CRC of varied training datasets, we defined a comparison scheme and used the coefficients calculated with the whole training dataset as a control group. We then reduced 5% and 10% of the training dataset, respectively, and obtained the differences of the coefficients calculated with these three datasets. Similarly, CRC of varied  $\delta_{i,e}$  was obtained by the coefficients calculated with the initial  $\delta_{i,e}$  and varied  $\delta_{i,e}$  according to Equation 20. Here, we set  $\delta_{i,e}$  fluctuated in the ranges of [-10%, 10%], [-20%, 20%], [-30%, 30%], [-40%, 40%] and [-50%, 50%], respectively as varied inputs.

### 1) VARIED TRAINING DATASETS

Figure 12, Figure 13 and Figure 14 exhibit the  $CRC_{5\%}$  and  $CRC_{10\%}$  of the S model, the C-S model and the proposed model ( $CRC_{5\%}$  and  $CRC_{10\%}$  denotes the CRC with 5% and 10% reduced training datasets, respectively). In Figure 12, 20 monitoring points had a higher average CRC of all coefficients ( $a_i, i=1,2,3,4; b_i, i=1,2; c_i, i=1,2$  and constant term) with the 10% reduced training dataset than with the 5% reduced training dataset. The average  $CRC_{5\%}$  of PL19-5 (3.677) was maximum, and the fluctuations of the coefficients of PL9-4 (0.571), PL11-5 (0.917), PL16-4 (0.596) and PL19-4 (0.521) were relatively more obvious. When the training dataset was reduced by 10%, the coefficients of PL9-4,



**FIGURE 12.** The change ratio of coefficient CRC of the S model with (a) 5% and (b) 10% reduced training dataset.

PL9-5, PL11-5, PL16-4, PL19-4 and PL19-5 had the most obvious fluctuations whose  $CRC_{10\%}$  were above 0.5.

In Figure 13, the  $CRC_{5\%}$  of  $A_{1ln}, A_{2ln}$  and  $B_{11ln}$  ( $l=0,1,2,3; n=0,1,2,3$ ) and  $CRC_{10\%}$  of  $A_{1ln}, A_{2ln}, B_{01ln}, B_{10ln}$  and  $B_{11ln}$  ( $l=0,1,2,3; n=0,1,2,3$ ) were relatively large. As more than half coefficients were not used in the C-S model, the CRC of these coefficients were considered as 0. However, in the C-S model, the maximum  $CRC_{5\%}$  and  $CRC_{10\%}$  were 67 and 55, which were 18.26 and 59.98 times as much as the maximum CRC in the S model. In this case, the coefficients in the C-S model were quite sensitive when the training dataset varies, and thus the model was unstable.

In the proposed model, the maximum average  $CRC_{5\%}$  was 0.146 at the monitoring point PL19-5, and the average  $CRC_{5\%}$  of the monitoring points PL9-5, PL13-4, PL19-3 and PL19-4 exceeded 0.05. The CRC of over 80% monitoring points (19 of 23) had a trend of increase with the reduced training datasets, in which  $CRC_{10\%}$  of PL11-5 and PL19-5 exceeded 0.2. Compared with the S model and the C-S model,



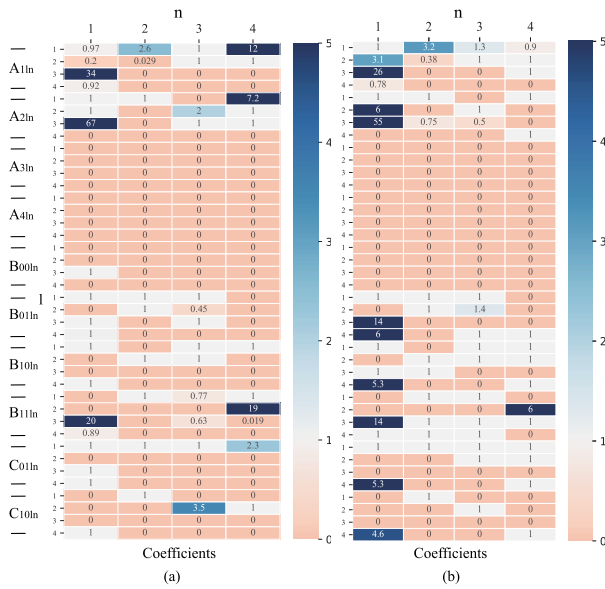


FIGURE 13. The CRC of the C-S model with (a) 5% and (b) 10% reduced training dataset.

TABLE 5. Maximum CRC of different measurement errors inputs of S model, C-S model and the proposed model.

	S model	C-S model	the proposed model
$[-10\%, 10\%]$	15.3235	15.8318	0.0436
$[-20\%, 20\%]$	103.5521	119.3915	0.1019
$[-30\%, 30\%]$	415.7012	117.9608	0.1267
$[-40\%, 40\%]$	446.9219	329.9919	0.1483
$[-50\%, 50\%]$	448.5635	285.4802	0.2361

the proposed model was more stable when the training dataset changes. From Figure 12 and 14, it is noted that the monitoring points around the dam foundation (i.e., PL19-5, PL19-4 and PL9-5) had a more significant influence by varying the training dataset.

2) VARIED MEASUREMENT ERRORS

Table 5 represents the maximum CRC of the three models, in which, the maximum CRC of S model, C-S model and proposed model increased from 15.3235 to 448.5635, from 15.8318 to 285.4802 and from 0.0436 to 0.2361, respectively. The coefficients in the proposed model fluctuated much less in response to the varied  $\delta_{i,e}$  compared with the other two models.

In Figure 15, the average CRC of S model, C-S model and the proposed model varied from 0.1898 to 3.8137, from 0.8939 to 4.9890 and from 0.0085 to 0.0209, respectively. In general, the coefficients of all three models fluctuated more obviously with the measurement errors inputs varied in a wider range. The coefficients in S model and C-S model were sensitive to the variation of measurement errors inputs, on average, a change that is nearly three times as much as the

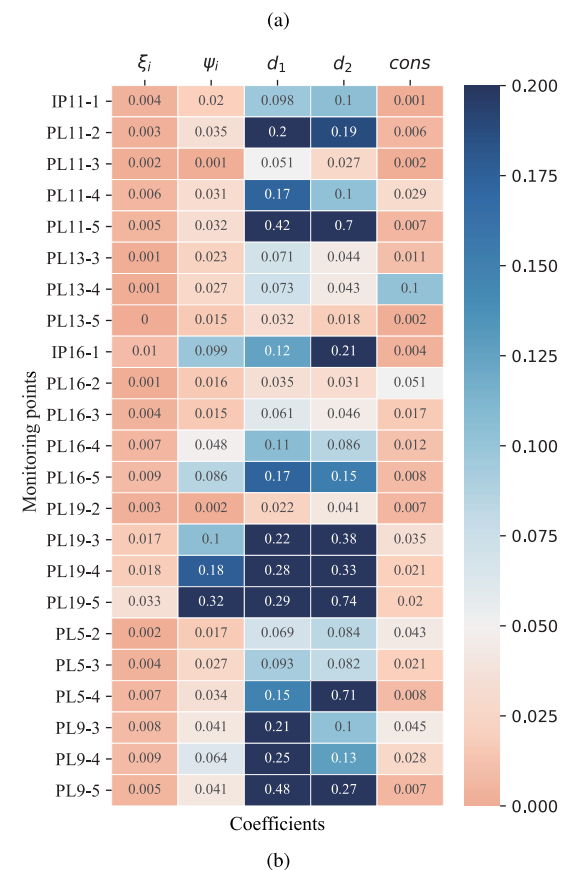
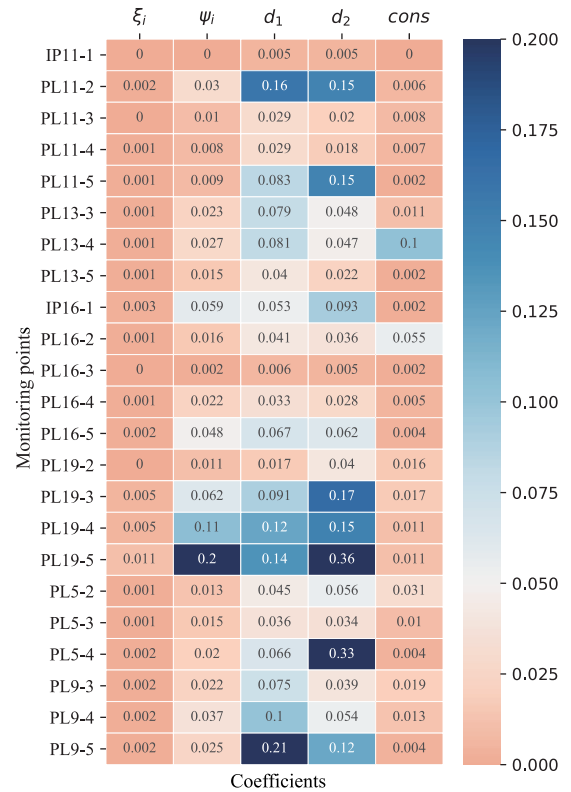


FIGURE 14. The CRC of the proposed model with (a) 5% and (b) 10% reduced training dataset.

TABLE 6. The coefficients of explanatory variables in the statistical model.

Monitoring point	$a_1$	$a_2$	$a_3$	$a_4$	$b_1$	$b_2$	$c_1$	$c_2$	cons ( $a_0$ )
IP11-1	0.026	0	0	0	0.174	0	0.028	0.226	-0.416
IP16-1	0.03	0	0	0	0.157	0	0.163	-0.306	-0.893
PL11-2	-51.644	0.31	-0.001	0	4.513	0	0.953	-4.839	-11.696
PL11-3	-57.881	0.346	-0.001	0	3.377	-0.386	0.961	-4.875	-10.683
PL11-4	-39.268	0.235	-0.001	0	2.134	-0.43	0.654	-3.185	-7.809
PL11-5	-0.368	0.001	0	0	0.787	0	0.078	0	-1.85
PL13-3	-74.279	0.444	-0.001	0	3.665	-0.6	1.174	-6.24	-9.665
PL13-4	-58.525	0.35	-0.001	0	2.496	-0.706	0.875	-4.698	-7.208
PL13-5	-36.054	0.215	-0.001	0	1.304	-0.494	0.503	-2.657	-3.347
PL16-2	-61.298	0.365	-0.001	0	3.112	0	1.317	-5.585	-8.053
PL16-3	-62.583	0.373	-0.001	0	2.478	-0.521	1.224	-4.99	-7.07
PL16-4	-48.815	0.291	-0.001	0	1.565	-0.56	1.023	-3.884	-5.405
PL16-5	0.109	0	0	0	0.461	0	0.355	-1.092	-3.258
PL19-2	1.553	0	0	0	1.328	0.388	0.585	-1.258	-4.286
PL19-3	1.496	0	0	0	0.637	0.208	0.411	-0.357	-3.556
PL19-4	0.089	0	0	0	0.293	0	0.283	-0.473	-2.813
PL19-5	0	0.002	0	0	0.093	0	0.174	0	-1.055
PL5-2	-33.587	0.201	-0.001	0	1.96	0.573	0.639	-2.449	-5.389
PL5-3	-29.775	0.178	0	0	1.066	0.137	0.515	-1.953	-3.653
PL5-4	0.018	0	0	0	0.398	0.15	0.127	0	-0.483
PL9-3	-49.935	0.299	-0.001	0	2.056	0	0.752	-3.799	-8.17
PL9-4	0.174	0	0	0	0.473	0.285	0	0	-6.069
PL9-5	0.037	0	0	0	0.214	-0.031	0.024	-0.087	-0.587

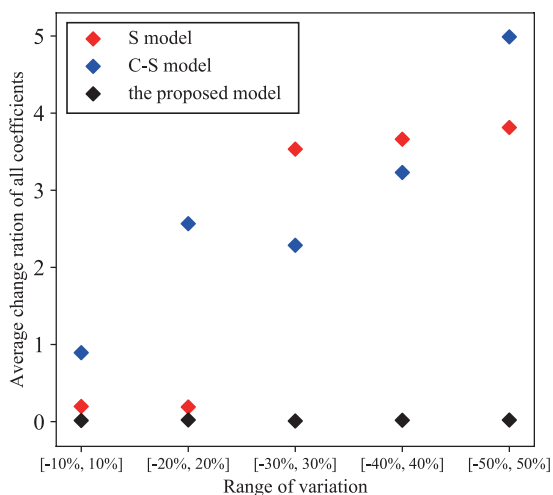


FIGURE 15. The average CRC of S model, C-S model and the proposed model with different varied measurement errors series.

initial inputs. Whereas the coefficients of the proposed model had little variation toward varied measurement errors inputs.

V. CONCLUSION

In the domain of dam's displacement prediction based on monitoring data, most previous studies gave emphasis on improving the prediction accuracy and paid less attention on the model stability. In this study, we proposed a numerical-statistical combined model to enhance the model stability. The proposed model considered the spatial correlations of different monitoring points and the randomness of the coefficients of explanatory variables. We quantified the

dam's displacement via three modules: the recoverable displacement, the non-recoverable displacement, and the measurement error. Numerical simulation was used to construct the coefficients of the explanatory variables in the module of recoverable displacement. The randomness of the coefficients of the explanatory variables was constrained with a random coefficient model.

We used the monitoring displacement data of a concrete arch dam at the Jinping-I hydropower station to validate the proposed model. The coefficients of determination  $R^2$  of the proposed model were above 0.9 for the training dataset of all monitoring points. Compared with the statistical model (S model) and the coordinates-introduced statistical model (C-S model), the proposed model had a better prediction ability: the smallest increase of RMSE with the increase of the prediction time. In addition, the proposed model had the best stability: the average change ratio of coefficient when reduce 5% and 10% of the training dataset ( $CRC_{5\%}$  and  $CRC_{10\%}$ ) were the lowest compared with the S model and the C-S model, and the lowest CRC in response to five different varied  $\delta_{i,e}$  series. With the better model stability, the proposed model is more suitable for the long-term displacement prediction of large dams.

ABBREVIATIONS

The following abbreviations are used in this manuscript:

- S model      Statistical model
- C-S model    Coordinates-introduced statistical model
- RMSE        Root Mean Squared Error
- CRC         change ratio of the coefficient
- $R^2$          Coefficient of determination

TABLE 7. The coefficients of the explanatory variables in the coordinates-included statistical model.

$A_{kln}$	$A_{1ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$				$A_{2ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$			
	24.448	1.096	-4.587	-4.067	-1.432	2.926	-6.713	-
	-1.432	2.926	-6.713	-	-0.001	-	-0.003	-
	-3.472	-	-	-	-0.001	0.004	-0.006	-
	43.576	-	-	-	-	-	-	-
$A_{3ln}$	$A_{3ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$				$A_{4ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$			
	-	-	-	-	0	-	0	0
	-	-	-	-	-	0	-	-
	-	-	0	-	-	0	-	0
	0	0	0	-	0	0	0	0
$B_{jkl n}$	$B_{00ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$				$B_{01ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$			
	-	-	-	-	-187.453	490.826	-1226.969	-
	-	-	-	-	-	-	-220.975	-
	-	-	-	-	-69.102	-	-	-
	-	-	-	-	-1720.687	-	-	-
$B_{10ln}$	$B_{10ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$				$B_{11ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$			
	2590.05	-	13439.857	2276.397	-	1165.04	-3916.481	-
	-	-	2167.973	-	-	-	-	-1557.888
	-	-	-	-	-2932.152	-	-1169.806	1373.378
	31275.161	-	-	-	-9.824	-	-	-
$C_{jkl n}$	$C_{01ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$				$C_{10ln}(l = 0, 1, 2, 3; n = 0, 1, 2, 3)$			
	-0.313	-2.265	-1.316	-0.593	-	0.055	-	-
	-	-	-	-	-	-	-0.008	-
	-	-	-	-	-	-	-	-
	-9.824	-	-	-	0.154	-	-	-

APPENDIX  
COEFFICIENTS RESULTS OF THE S MODEL AND THE C-S MODEL

Tables 6–7 show the coefficients of explanatory variables in the statistical model and the coordinates-introduced statistical model, respectively.

REFERENCES

[1] P. Bukenya, P. Moyo, H. Beushausen, and C. Oosthuizen, "Health monitoring of concrete dams: A literature review," *J. Civil Struct. Health Monitor.*, vol. 4, no. 4, pp. 235–244, Nov. 2014.

[2] A. De Sortis and P. Paoliani, "Statistical analysis and structural identification in concrete dam monitoring," *Eng. Struct.*, vol. 29, no. 1, pp. 110–120, Jan. 2007.

[3] C. Shao, C. Gu, Z. Meng, and Y. Hu, "Integrating the finite element method with a data-driven approach for dam displacement prediction," *Adv. Civil Eng.*, vol. 2020, pp. 1–16, Apr. 2020.

[4] P. Léger and M. Leclerc, "Hydrostatic, temperature, time-displacement model for concrete dams," *J. Eng. Mech.*, vol. 133, no. 3, pp. 267–277, Mar. 2007.

[5] J. Mata, "Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models," *Eng. Struct.*, vol. 33, no. 3, pp. 903–910, Mar. 2011.

[6] Y.-S. Kim and B.-T. Kim, "Prediction of relative crest settlement of concrete-faced rockfill dams analyzed using an artificial neural network model," *Comput. Geotechnics*, vol. 35, no. 3, pp. 313–322, May 2008.

[7] C.-Y. Kao and C.-H. Loh, "Monitoring of long-term static deformation data of Fei-Tsui arch dam using artificial neural network-based approaches," *Struct. Control Health Monitor.*, vol. 20, no. 3, pp. 282–303, Mar. 2013.

[8] Z. Meng, Y. Hu, and C. Ancey, "Using a data driven approach to predict waves generated by gravity driven mass flows," *Water*, vol. 12, no. 2, p. 600, Feb. 2020.

[9] A. Hipni, A. El-shafie, A. Najah, O. A. Karim, A. Hussain, and M. Mukhlisin, "Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS)," *Water Resour. Manage.*, vol. 27, no. 10, pp. 3803–3823, Aug. 2013.

[10] M. A. Hariri-Ardebili and F. Pourkamali-Anaraki, "Support vector machine based reliability analysis of concrete dams," *Soil Dyn. Earthq. Eng.*, vol. 104, pp. 276–295, Jan. 2018.

[11] F. Kang, J. Liu, J. Li, and S. Li, "Concrete dam deformation prediction model for health monitoring based on extreme learning machine," *Struct. Control Health Monitor.*, vol. 24, no. 10, p. e1997, Oct. 2017.

[12] F. Salazar, R. Morán, M. Á. Toledo, and E. Oñate, "Data-based models for the prediction of dam behaviour: A review and some methodological considerations," *Arch. Comput. Methods Eng.*, vol. 24, no. 1, pp. 1–21, Jan. 2017.

[13] D. Santillán, E. Saleté, D. J. Vicente, and M. Á. Toledo, "Treatment of solar radiation by spatial and temporal discretization for modeling the thermal response of arch dams," *J. Eng. Mech.*, vol. 140, no. 11, Nov. 2014, Art. no. 05014001.

[14] R. Tomás, M. Cano, J. García-Barba, F. Vicente, G. Herrera, J. M. Lopez-Sanchez, and J. J. Mallorquí, "Monitoring an earthfill dam using differential SAR interferometry: La Pedrera dam, Alicante, Spain," *Eng. Geol.*, vol. 157, pp. 21–32, May 2013.

[15] C. Shao, C. Gu, M. Yang, Y. Xu, and H. Su, "A novel model of dam displacement based on panel data," *Struct. Control Health Monitor.*, vol. 25, no. 1, p. e2037, Jan. 2018.

[16] Y. Hu, C. Shao, C. Gu, and Z. Meng, "Concrete dam displacement prediction based on an ISODATA-GMM clustering and random coefficient model," *Water*, vol. 11, no. 4, p. 714, Apr. 2019.

[17] C. Shao, C. Gu, Z. Meng, and Y. Hu, "A data-driven approach based on multivariate copulas for quantitative risk assessment of concrete dam," *J. Mar. Sci. Eng.*, vol. 7, no. 10, p. 353, Oct. 2019.

[18] S. Khalid, B. Singh, G. C. Nayak, and O. P. Jain, "Nonlinear analysis of concrete face rockfill dam," *J. Geotech. Eng.*, vol. 116, no. 5, pp. 822–837, May 1990.

[19] I. Karimi, N. Khaji, M. T. Ahmadi, and M. Mirzayee, "System identification of concrete gravity dams using artificial neural networks based on a hybrid finite element-boundary element approach," *Eng. Struct.*, vol. 32, no. 11, pp. 3583–3591, Nov. 2010.

[20] B. Wu, J. Niu, H. Su, M. Yang, Z. Wu, and X. Cui, "An approach for deformation modulus mechanism of super-high arch dams," *Struct. Eng. Mech.*, vol. 69, no. 5, pp. 557–566, 2019.

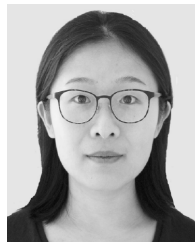
[21] P. A. Swamy, "Efficient inference in a random coefficient regression model," *Econometrica, J. Econ. Soc.*, vol. 38, pp. 311–323, Mar. 1970.

[22] P. D. Bliese and R. E. Ployhart, "Growth modeling using random coefficient models: Model building, testing, and illustrations," *Org. Res. Methods*, vol. 5, no. 4, pp. 362–387, Oct. 2002.

[23] L. D. Avendaño-Valencia and S. D. Fassois, "Gaussian mixture random coefficient model based framework for SHM in structures with time-dependent dynamics under uncertainty," *Mech. Syst. Signal Process.*, vol. 97, pp. 59–83, Dec. 2017.



**YATING HU** was born in Nanchang, Jiangxi, China, in 1994. She received the B.Eng. degree in water conservancy and hydropower engineering from Hohai University, Nanjing, China, in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include early warning and safety evaluation for hydraulic structures, monitoring data analysis, and artificial intelligence algorithms.



**ZHENZHU MENG** was born in Dezhou, Shandong, China, in 1991. She received the B.Eng. and M.Sc. degrees in water conservancy and hydropower engineering from Hohai University, in 2012 and 2015, respectively. She is currently pursuing the Ph.D. degree with the Environmental Hydraulics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL). Her research interests include landslide generated waves and structural monitoring data analysis.



**CHONGSHI GU** was born in Qidong, Jiangsu, China, in 1962. He received the B.Eng. and Ph.D. degrees in water conservancy and hydropower engineering from Hohai University, Nanjing, China, in 1985 and 1997, respectively.

Since 2004, he has been a Professor with the Water Conservancy and Hydropower Engineering College, Hohai University. From 2010 to 2020, he was the Dean of the College. He is the author of six books and more than 150 articles. His research

interests include early diagnosis for hydraulic structures, dam safety evaluation, and structural mechanics. He was a recipient of the Award of Scientific and Technological Progress of the State, in 2005, 2007, and 2015. He is the Chief Editor of *Advances in Science and Technology of Water Resources*.



**CHENFEI SHAO** was born in Nanjing, Jiangsu, China, in 1989. He received the B.Eng. degree in water conservancy and hydropower engineering from Hohai University, Nanjing, in 2011, where he is currently pursuing the Ph.D. degree. His research interests include early warning models for hydraulic structures, monitoring data analysis, computational mechanics, and engineering software.

...