Prediction for the Settlement of Concrete Face Rockfill Dams Using Optimized LSTM Model via Correlated Monitoring Data

Yating Hu [1,2,3], Chongshi Gu [1,2,3,*], Zhenzhu Meng [4], Chenfei Shao [1,2,3] and Zhongze Min [5]

1   State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China; huyating_hhu@163.com (Y.H.); shao_chen_fei@hhu.edu.cn (C.S.)
2   College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
3   National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Nanjing 210098, China
4   School of Civil Engineering, Sun Yat-Sen University, Zhuhai 519082, China; mengzhzh@mail.sysu.edu.cn
5   Guangdong Communication Planning & Design Institute Group Co., Ltd., Guangzhou 510507, China; 15950527738@163.com
*   Correspondence: csgu@hhu.edu.cn; Tel.: +86-13809043532

Abstract: Settlement prediction is of great importance for safety control of concrete-face rockfill dams (CFRDs) during the operation stage. However, the prediction accuracy achieved by the commonly used hydrostatic–seasonal–time (HST) methods, without the consideration of the previous conditions of influencing factors, is not competitive. Moreover, in most methods, settlement data at each monitoring point are modeled individually; the correlation relationships between settlements are neglected. In this paper, a method based on an optimized long short-term memory (LSTM) model is proposed to predict the settlement of CFRDs, modeling multiple monitoring data series with strong correlation relationships simultaneously. In the method, settlement data series are classified into several categories, firstly according to a global relevance measure. Then, the cuckoo search (CS) algorithm is applied to optimize the hyper-parameters in the neural network structure of LSTM. Ultimately, the LSTM model is utilized to predict the multiple settlement data series classified in the same category. Results indicate that the proposed method has a better prediction performance compared with the LSTM model, the back propagation neural network (BPNN) model, and the HST with single monitoring point.

Keywords: settlement prediction model; concrete-face rockfill dam; long short-term memory; cuckoo search algorithm; multiple monitoring point model

1. Introduction

The concrete-face rockfill dam (CFRD) has been extensively adopted in current dam construction because of its strong adaptability to geological and economical construction technology [1,2]. To ensure the long-term safe operation of the CFRD, multiple instruments are installed to monitor its structural response to environmental load changes [3,4]. Deformation and seepage are two critical monitoring quantities that can reflect the operation status of CFRD, in which a large or uneven deformation may cause cracks to the concrete panel and even imply a failure of the structure [5,6]. Therefore, the prediction of settlement is one of the most important issues in structural health monitoring of CFRD.

The statistical model [7,8] and hybrid model [9,10] are the most adopted methods to predict the settlement of CFRDs. Sun et al. [11] proposed a method combining an improved back propagation neural network (BPNN) and finite element method (FEM) to inverse the material parameters of the CFRD. The hybrid model can be used to achieve comprehensive analysis for the settlement of the CFRD. However, it is time consuming and requires sufficient topographic and geological data of the engineering project, which may not be satisfied in small-scale civil engineering.

The statistical model adopted mathematical methods to construct the relationship between the settlement of the CFRD and its explanatory variables. The HST model is one of the most acknowledged statistical models, which describes the settlement as a combination of hydro-statistic components, temperature components, and aging components. With the development of computer science, plenty of machine learning-based methods have been applied to predict the settlement of CFRDs [12,13]. Owing to the strong ability to map the nonlinear relationship between the settlement and its explanatory variables, the machine learning techniques improve the performance of the prediction model intensely. For instance, a novel prediction model based on the wavelet support vector machine was proposed to model the settlement of the dam and the prediction accuracy was improved significantly [14]. A method combining the artificial neural network and support vector regression was implemented to predict the settlement of the dam with the consideration of long-term thermal behavior [15]. However, in most models, previous values of influencing factors and settlement are not considered sufficiently. In fact, the influence from influencing factors and settlement conditions both have a lag effect on the current settlement. Therefore, it is essential to establish a previously memorable prediction model which can utilize the relationships between previous times of settlement to analyze the settlement of the CFRD.

With the increase of monitoring points installed in the dam, monitoring data become abundant. Based on that, researchers started to study the correlation relationship among monitoring data series [16]; the modeling method is transferring from a single monitoring point (SMP) model to a multiple monitoring points (MMP) model. Wang et al. [17] proposed a spatial association-coupled support vector machine model to predict the displacement of arch dams. Hu et al. [18] proposed a multiple monitoring point prediction method using a random coefficient model and GMM-ISODATA cluster algorithm. In many cases for the prediction of radial displacement for arch dams, the stability of the MMP model has been verified. In this work, we introduce the idea of integrating data of multiple monitoring points in the prediction model for the settlement of the CFRD.

We herein establish a previously memorable MMP prediction model for the settlement of CFRD. In the MMP model, the first step is to determine the correlation relationship between monitoring data series, and then classify the data series according to the correlation coefficient. Next, the MMP model calculates the data series of high correlation simultaneously; in this way, the correlation relationship among the data series can be considered. Here, the clustering index adopts the Pearson coefficient, which can reflect the linear correlation characteristics of the data series. The classification method applied in this research is an improved K-means algorithm [19,20].

To establish a previously memorable model, long short-term memory (LSTM), an improved recurrent neural network (RNN), is adopted in this work. LSTM was proposed by Hochreiter and Schmidhuber [21] and has been successfully applied in multiple fields [22,23]. Owing to the memorability of NN structures [24,25], LSTM is qualified to model the settlement of CFRD, which considers the historical process of the settlement condition and influencing factors. The cuckoo search (CS) algorithm [26,27] is an extensively utilized optimization algorithm because of its advantages of fewer parameters, a random search path, and strong optimization ability. To improve the performance of LSTM, we employ the CS algorithm to search the optimal parameters, i.e., hidden layers, hidden nodes, and learning rate in the neural network structures.

The objective of this work is to present a prediction method for the settlement of the CFRD, which is based on the LSTM model, on groups of datasets with similar regular variation. We primarily interpret the settlement of CFRD during the operation stage, determining the explanatory variables used in the following LSTM training framework. The settlement data series are clustered into several categories by an improved K-means clustering approach. The CS algorithm is utilized to optimize the learning rates and neural network structures, so as to enhance the performance of the LSTM model. Ultimately, LSTM is applied to model the multiple monitoring settlement data series in the same category simultaneously.

This article is organized as follows. Section 2 presents the interpretation model for the settlement of CFRD. Section 3 illustrates the framework of the proposed optimized M-LSTM method; we first introduce the clustering process and then the optimization for the LSTM model with CS algorithm. Section 4 exhibits the engineering project and the settlement dataset. Training and predicting results are exhibited and discussed in Section 5. Concluding remarks complete the paper in Section 6.

2. The Interpretation Model for the Settlement of CFRD

Settlement of CFRD refers to the vertical-direction displacement under load, which can be mainly divided into initial settlement, consolidation settlement, and secondary consolidation settlement. In this work, we put emphasis on the settlement of the CFRD during the operation period. At this stage, the initial settlement and consolidation settlement of the soil are basically completed; the secondary consolidation process during the operation stage is therefore the core component to be analyzed in the following. Figure 1 exhibits the sketch of the CFRD and the main components of the settlement.
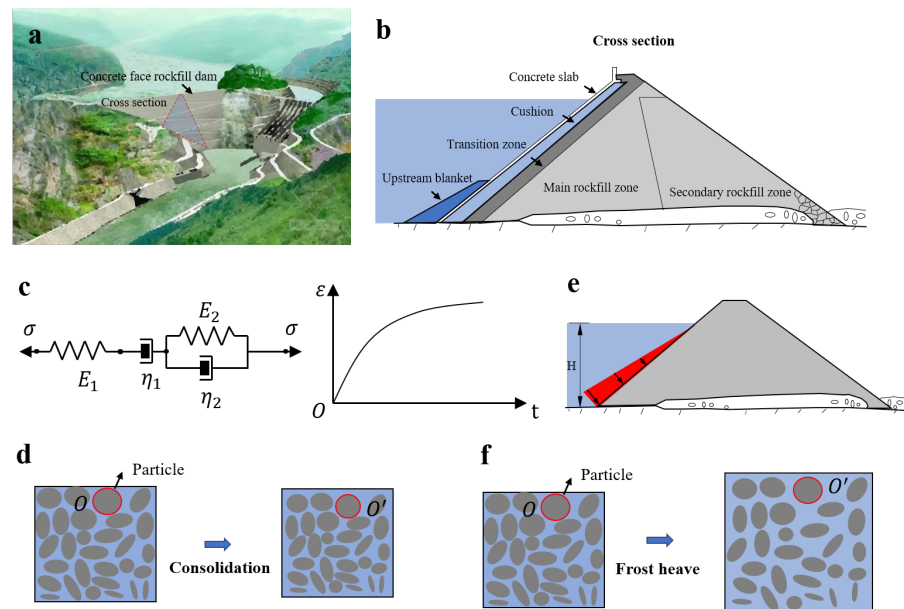

Figure 1. Schematic diagram of the CFRD and the main components of the settlements: (**a**) sketch of the CFRD, (**b**) structure of the main body of the CFRD, (**c**) creep deformation component, (**d**) consolidation of soil particles, (**e**) hydrostatic load component, (**f**) frost heave of soil particles.

In the process of the secondary consolidation of the soil, effective stress on the soil particles remains almost unchanged, but the volume of the soil still grows with time. The deformation produced in the above-mentioned second consolidation can be called creep deformation. A number of indoor tests and on-site measurements have been applied to the soil; results imply that the creep deformation of the secondary consolidation is related to the time. In previous work, multiple empirical calculation methods for the creep settlement δ_θ of different kinds of soil have been proposed, and the expression widely used in the engineering is presented as follows [28–31]:

$$\delta_\theta = \begin{cases} C_1\theta + C_2 ln\theta \\ \dfrac{\theta}{C_1\theta + C_2} \\ C_1 e^{\frac{C_2}{\theta}} \end{cases} \tag{1}$$

where t is the time, $\theta = \frac{t}{100}$, and C_1 and C_2 are pending coefficients.

We will adopt the suitable forms from above optional equations as explanatory variables when analyzing the creep deformation of the CFRD during the operation period.

Except for the settlement induced by the secondary consolidation, soil settlement occurs when the soil bears external loads. During the operation stage of the CFRD, the main external load comes from the upstream water head. The hydrostatic-related deformation has a relationship with the stress–strain level of the material, which is related to the historical external load put on the CFRD. Therefore, the settlement of the CFRD during the operation period is related to the current and the historical stress state, simultaneously. Based on the analysis above, the settlement induced by the hydrostatic load is concluded as:

$$\delta_H = f(H, \overline{H}) \tag{2}$$

Specifically, it can be expanded as:

$$\delta_H = \sum_{i=0}^{3} a_{1i} H^i + \sum_{i=0}^{m_1} a_{2i} \overline{H}^i \tag{3}$$

where H is the upstream water level, $\overline{H}$ is the average value of upstream water levels in the early stage, and m_1 represents the dates considered as the previous stage.

In addition to hydrostatic load, temperature loads may induce settlement in certain conditions. Theoretically, the linear expansion of the soil caused by temperature changes has little effect on the settlement of the CFRD. However, in some extremely cold zones, where the temperature usually falls below zero, the soil will have a relatively significant frost heave settlement. Since the emergence and duration of the freezing period almost has an annual change cycle, we can apply a periodic function to express the relationship between the temperature and the temperature-related settlement. In summary, for the interpretation of the settlement for CFRD, we can ignore the influence from the temperature load generally, but for the CFRD in extremely cold zones, the temperature-related settlement can be expressed by the following periodic function:

$$\delta_T = \sum_{i=1}^{m_2} (b_{1i} \sin \frac{2\pi it}{365} + b_{2i} \cos \frac{2\pi it}{365}) \tag{4}$$

where m_2 is selected as 1 for CFRD, b_{1i}, b_{2i} are pending coefficients, and t denotes the time.

Based on the above analysis, the interpretation model of the settlement for CFRD has three main components, including the creep deformation component δ_θ, the hydrostatic load component δ_H, and the temperature load component δ_T (for CFRDs in extremely cold zones), and the expression of the settlement for CFRD δ can be expressed as:

$$\delta = \delta_\theta + \delta_H + \delta_T + cons \tag{5}$$

where *cons* is the constant term.

3. LSTM-CS MMP Prediction Model for the Settlement of CFRD

In this section, we will construct the MMP prediction model based on the LSTM neural network of the settlement of the CFRD.

3.1. Data Clustering Based on K-Means++ Algorithm

To recognize the monitoring data with similar variation, we select the Pearson correlation coefficient to measure the correlation relationship between the monitoring points. The expression of the Pearson correlation coefficient is:

$$\rho_{ij} = \frac{\sum_{t=1}^{T} (\delta_{it} - \bar{\delta}_i)(\delta_{jt} - \bar{\delta}_j)}{\sqrt{\sum_{t=1}^{T} (\delta_{it} - \bar{\delta}_i)^2 \sum_t (\delta_{jt} - \bar{\delta}_j)^2}} \tag{6}$$

where δ_{it} and δ_{jt} denote the data series of monitoring point i and j, respectively, and $\bar{\delta}_i$ and $\bar{\delta}_j$ denote the mean value of the two monitoring data series.

Then, ω_i is calculated to represent a global relevance measure of a certain monitoring point, which includes the attributes of the correlation relationship between the monitoring data series δ_i and all other monitoring data series. The expression of ω_i is

$$\omega_i = [\rho_{i1}, \rho_{i2}, \cdots, \rho_{ij}, \cdots, \rho_{iN}]^T \tag{7}$$

where N is the number of monitoring points.

Based on the global relevance measure ω_i, we adopt the *K*-means++ algorithm to classify monitoring data. Suppose there are N monitoring data samples $\Omega = \{\omega_1, \omega_2, \cdots, \omega_N\}$ in total; our goal is to classify N samples into K clusters $C = \{C_1, C_2, \cdots, C_K\}$. The clustering process implemented by the *K*-means++ algorithm is presented as follows.

Step 1: Select K initial cluster centers $\mu_{pj}^{(1)}$ ($1 \leq p \leq K, 1 \leq j \leq N$). The improved algorithm selects the initial cluster centers by two calculation procedures: the first procedure is to select one sample from the dataset indiscriminately as one initial cluster center; the second procedure is to determine the shortest distance between each sample and the existing cluster centers. Then, the roulette method is applied to determine the next cluster center, and the probability that each sample could be selected as the next cluster center P_i is expressed as:

$$P_i = \frac{minD(\omega_i, \mu_{pj}^{(1)})^2}{\sum_{i=1}^{N} minD(\omega_i, \mu_{pj}^{(1)})} \tag{8}$$

where $D(\omega_i, \mu_{pj}^{(1)})$ denotes the Euclidean distance between ω_i and $\mu_{pj}^{(1)}$. The expression of $D(\omega_i, \mu_{pj}^{(1)})$ is exhibited below.

$$D(\omega_i, \mu_{pj}^{(1)}) = \sqrt{\sum_{j=1}^{N} (\omega_{ij} - \mu_{pj}^{(1)})^2} \tag{9}$$

Calculate Equations (8) and (9) cyclically until all initial cluster centers are decided.

Step 2: Calculate the Euclidean distance between ω_i ($1 \leq i \leq N$, N denotes the number of samples) and $\mu_{pj}^{(n)}$ in the *n*-th iteration.

Step 3: Classify ω_i into the cluster according to the smallest distance $D(\omega_i, \mu_{pj}^{(n)})$.

Step 4: Update the cluster centers. The updated cluster center is the mean vector of all samples in a certain cluster, and the expression is:

$$\mu_{pj}^{(n+1)} = \frac{1}{N_p} \sum_{\omega_i \in C_p^n} \omega_i \tag{10}$$

where $\mu_{pj}^{(n+1)}$ represents the *p*th cluster center in the (*n*+1)-th iteration, N_p represents the number of samples in the *p*th cluster, and $\omega_i \in C_p^n$ represents all ω_i that belong to the *p*th cluster in the *n*-th iteration.

Step 5: Calculate the square error E_n of all samples.

$$E_n = \sum_{p=1}^{K} \sum_{\omega_i \in C_p^n} ||\omega_i - \mu_p^n||^2 \tag{11}$$

where μ_p^n represents the *p*th cluster center in the *n*-th iteration.

If $|E_n < E_{n-1}| < \alpha$, the program will terminate; otherwise, repeat Step 2 to Step 5.

3.2. CS Algorithm Optimized LSTM

LSTM, an improved type of recurrent neural network (RNN), is capable of recognizing long-term dependence experience. LSTM is utilized to model the monitoring settlement data and predict the unknown settlement of the CFRD in this work.

3.2.1. Long Short-Term Memory Structure

LSTM is conducted by a combination of neural network (NN) modules, as presented in Figure 2. Each NN module consists of a memory cell and three gates (input gate, forget gate, and output gate).
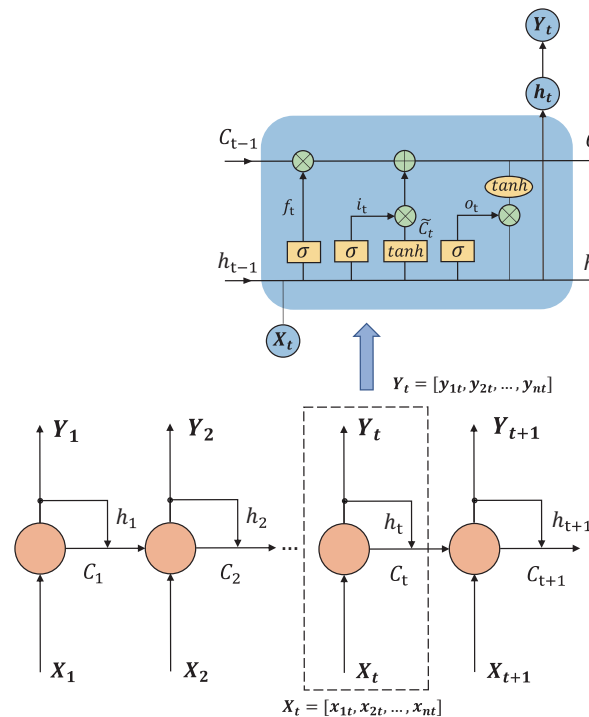

Figure 2. Structure of long short-term memory model.

The forget gate controls the ratio of the information flow from the output of the previous time step to be retained and to be discarded. The input gate decides the information flow from the previous step to be transferred into the memory cell. The output gate controls the information flow to be output, which is relative to vectors of cell memory output, previous output, and current input. The corresponding functions of each gate structure are as follows.

$$f_t = \sigma(W_{f1}x_t + W_{f2}y_{t-1} + b_f) \tag{12}$$

$$I_t = \sigma(W_{i1}x_t + W_{i2}y_{t-1} + b_i) \tag{13}$$

$$O_t = \sigma(W_{o1}x_t + W_{o2}y_{t-1} + b_o) \tag{14}$$

where f_t, I_t, and O_t represent the vector of the forget gate, input gate, and output gate of a cell of LSTM neutral network at t-th time, separately; σ represents the sigmoid activation function, mapping the real number to [0,1]; b_f, b_i, and b_o represent the bias weights for the forget gate, input gate, and output gate, respectively; x_t represents the input (explanatory variable of settlement) at t-th step time; y_{t-1} represents the output (settlement) at (t–1)-th step time; W_{f1}, W_{f2}, W_{i1}, W_{i2}, W_{o1}, and W_{o2} represent the connection weight between the input nodes and the hidden nodes, and between the hidden nodes and the output nodes, of the forget gate, input gate, and output gate, respectively.

h_t represents the output of the hidden layer at *t*-th step time, and the expression is:

$$h_t = O_t \tan h(C_t) \tag{15}$$

where C_t represents the vector of memory cell at *t*-th step time, which is obtained according to Equations (16) and (17), and $\tan h$ represents a hyperbolic tangent function, mapping the real number to $[-1, 1]$.

$$\widetilde{C}_t = \tan h(W_{c1}x_t + W_{c2}y_{t-1} + b_c) \tag{16}$$

$$C_t = f_t C_{t-1} + I_t \widetilde{C}_t \tag{17}$$

where W_{c1} and W_{c2} represent the connection weights between the input nodes and the hidden nodes and between the hidden nodes and the output nodes of the memory cell, and b_c represents the bias weights for the memory cell. y_{t-1} represents the output at the (*t*–1)-th time step. f_t and I_t are exhibited in Equations (12) and (13).

Ultimately, output (settlement) at *t*-th time step y_t can be obtained by the expression:

$$y_t = W_{hy}h_t + b_y \tag{18}$$

where W_{hy} represents the connection weight between the hidden nodes and the output vectors, b_y represents the bias weights of W_{hy}, and h_t is represented as Equation (15).

3.2.2. Cuckoo Search Algorithm

The CS algorithm is an optimization algorithm that models brood parasitism of cuckoo birds. The algorithm imitates cuckoos' behavior by Lévy flights, which has a good global search. The CS algorithm is applied to optimize the parameters of the LSTM neural network. Compared with isotropic random movement, Lévy flight has the possibility of long-distance movement. Figure 3 exhibits the searching path of Lévy flight and isotropic random walk in 2000 iterations; it can be noted that the step length of Lévy flight varies alternately, and the frequency of large strides occurring in Lévy flight is higher than isotropic random walk. The efficiency for global optimum searching of Lévy flight benefits from this combination of long and short steps.
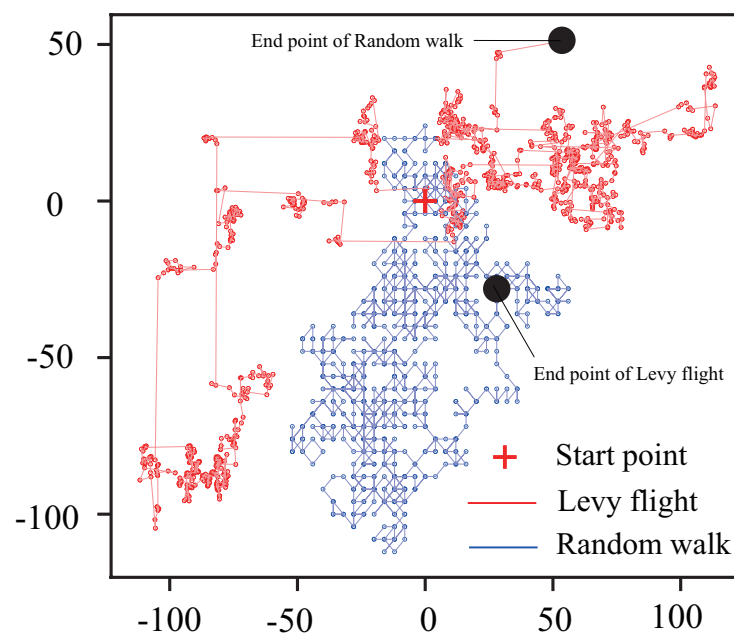

Figure 3. Schematics for typical searching process of Lévy flights and random walk in 2000 iterations.

The CS algorithm is applied in this work to search the best hyper−parameters in the LSTM network. The search procedure can be summarized as the following:

Step 1: Establish the objective function f.

$$f = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\delta}_i - \delta_i)^2}{N}} \tag{19}$$

where $\hat{\delta}_i$ denotes the modeled settlement data, δ_i denotes the monitoring settlement data, and N denotes the number of the data in the datasets.

Step 2: Generate the initial generation of cuckoos $X_1, X_2, ..., X_N$ randomly. Each cuckoo $X_i = (N_{li}, N_{ni}, LR_i)$ represents a set of solutions of the hyper−parameters. The number of hidden layers is N_{li}, the number of hidden nodes is N_{ni}, and the learning rate is LR_i, which is expressed as follows:

$$X = \begin{bmatrix} X_1 & F(N_{l1}, N_{n1}, LR_1) \\ X_2 & F(N_{l2}, N_{n2}, LR_2) \\ \cdots & \cdots \\ X_N & F(N_{lN}, N_{nN}, LR_N) \end{bmatrix} \tag{20}$$

where N is the number of cuckoos. According to Equation (19), the values of the objective function for each cuckoo are calculated, and the best cuckoo x_b^t can be determined.

Step 3: Update solutions through a Lévy flight. The expression is:

$$X_i^{t+1} = X_i^t + \alpha \oplus Levy(s, \lambda)(i = 1, 2, \cdots, N) \tag{21}$$

where α is the scaling factor of step size related to the scale of the problem, $\oplus$ denotes the entry−wise multiplications, and X_i^t and X_i^{t+1} are the nest locations of t-th and $(t+1)$-th generation of cuckoos, respectively. $Levy(s, \lambda)$ is a random searching vector obeying a Lévy distribution, which is expressed as

$$Levy(s, \lambda) = \frac{\lambda\Gamma(\lambda)sin(\pi\lambda/2)}{\pi}\frac{1}{s^{1+\lambda}}(s >> s_0 > 0) \tag{22}$$

where Γ is the Gamma function; s is the step size of Lévy flight which could be generated by Mantegna algorithm; s_0 is the minimum step size.

$$s = \frac{u}{|v|^{1/\beta}} \tag{23}$$

where u and v are random numbers following a Gaussian distribution, β is the parameter of Lévy flight, $u \sim N(0, \sigma_u^2)$, and $v \sim N(0, \sigma_v^2)$, in which, $\sigma_u = \left\{ \frac{\Gamma(1+\beta)sin\left(\frac{\beta}{2}\right)}{\Gamma\left[\frac{1+\beta}{2}\beta \times 2^{(\beta+1)/2}\right]} \right\}^{1/\beta}$; $\sigma_v = 1$.

Step 4: Throw out the alien eggs with a probability $P_a \in [0, 1]$. This process is simulated by the following equation:

$$X_i^{t+1} = \begin{cases} X_i^t + r \cdot \left(X_{r_1}^t - X_{r_2}^t\right), & if\ r < P_a \\ X_i^t, & otherwise \end{cases} \tag{24}$$

where $X_{r_1}^t$ and $X_{r_2}^t$ are two random nest locations in the t-th generation; r is a random number between 0 and 1.

Step 5: Calculate the value of the objective function of renewal nest locations. The optimal cuckoo of $(t+1)$-th generation X_b^{t+1} is obtained. After comparing the objective function value of X_b^{t+1} and X_b^t, the smaller function value can be kept as the $(t+1)$-th optimal cuckoo.

Step 6: Proceed with step 3 to step 4 recursively. The process ends when the termination criteria is satisfied.

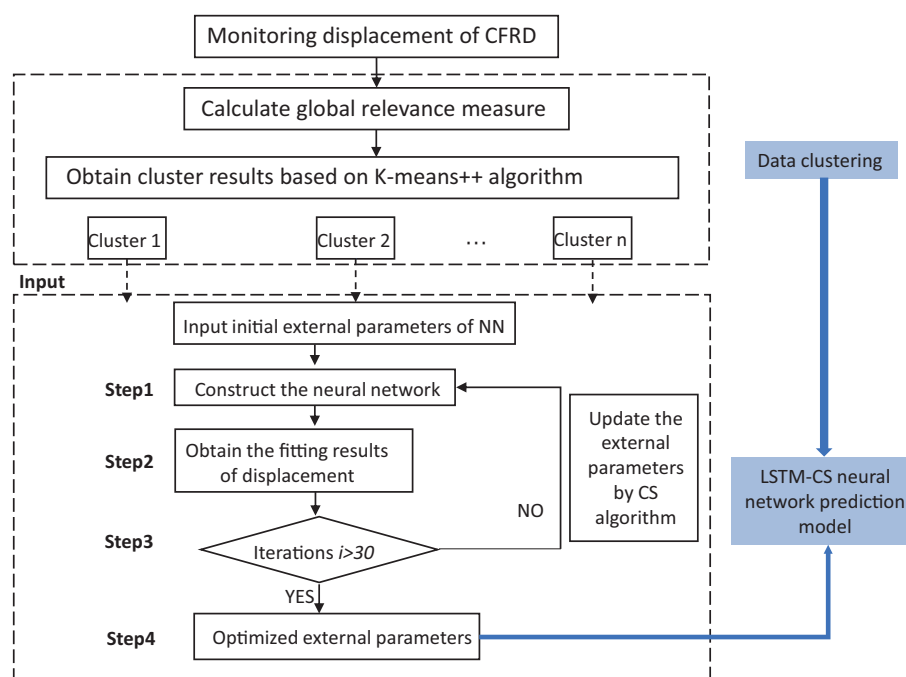Figure 4 exhibits the framework of the proposed method for predicting monitoring settlement of CFRD.

Figure 4. Flowchart of the proposed method for predicting the settlement of CFRD.

4. Project Overview

Langyashan pumped−storage power station (PSPS) is located in Chuzhou, Anhui Province. The power station is composed of main structures such as the CFRD of the upper reservoir, concrete dam of the lower reservoir, channel systems, etc. Figure 5a exhibits the location of the CFRD to be analyzed in this study. As shown in Figure 5a, the pumped−storage power station (PSPS) contains an upper reservoir and a lower reservoir; it stores energy in the form of gravitational potential energy of water. Excess power from the electrical grid is utilized to pump water from the lower reservoir, and the water stored in the upper reservoir can be utilized to generate electricity back to the grid as necessary. The upstream reservoir of Langyashan PSPS has a total storage capacity of 18.04 million m^3, a regulation storage of 12.38 million m^3, a normal storage level of 171.8 m, and a design flood level of 172.40 m.
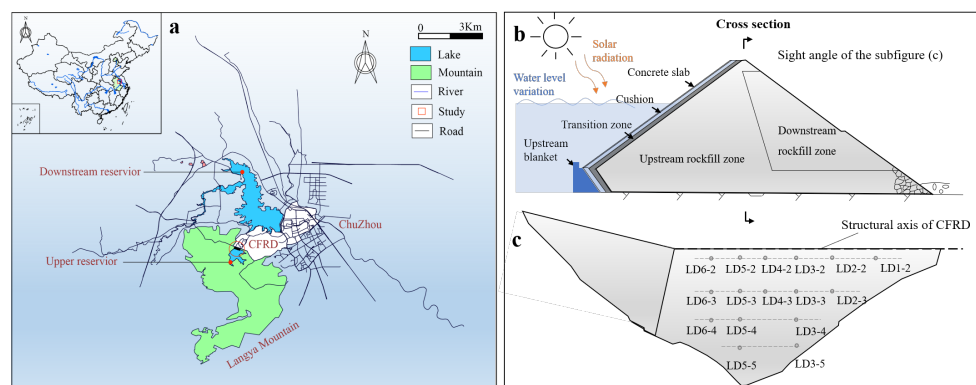

Figure 5. Schematic of: (**a**) location, (**b**) structure, (**c**) distribution of monitoring points of settlement, of the CFRD for Langyashan PSPS.

Langyashan CFRD consists of a concrete slab, cushion area, transition area, upstream blanket, upstream rockfill area, and downstream rockfill area, as exhibited in Figure 5b. The construction of Langyashan CFRD started in July 2002, and the main body was completed in March 2005. In order to measure the settlement of the CFRD, a total of 37 monitoring instruments were installed on the surface of a downstream dam slope, parallel to

the structural axis of the CFRD with different elevations (as shown in Figure 5c). Considering the failure of some instruments, we ultimately selected a monitoring data series of 16 settlement gauges for analysis.

In this study, we selected the settlement data (the vertical down is positive, the vertical up is negative) of all 16 monitoring points from 17 April 2005 to 28 October 2018 for analysis. After the pre−process of settlement data, i.e., removing the outliers, interpolating, compensating, etc., we obtained one settlement data point per week, a total of 700 data points in every monitoring dataset, as shown in Figure 6. Figure 7 presents the upstream water level and the environmental temperature in Langyashan CFRD. It is noted that the vertical settlement of all monitoring points have similar variation rules with time, increase slowly, and tend to converge, without obvious periodic variations.
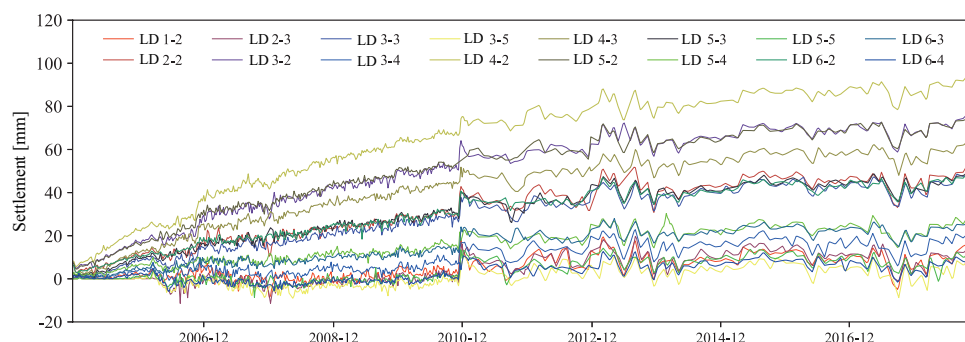

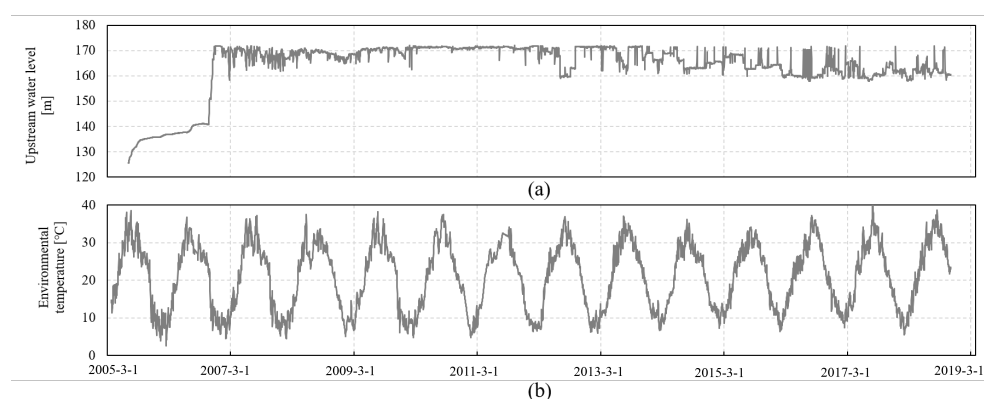**Figure 6.** Settlement data series of selected monitoring points.

Figure 7. Ambient data series of Langyashan dam: (**a**) upstream water level, (**b**) environmental temperature.

5. Results and Discussion

5.1. Clustering Results of Monitoring Data Series

According to the equations proposed in Section 3.1, we calculate the correlation coefficients ρ_{ij} between every two training datasets, δ_i and δ_j, of monitoring points; the results are exhibited in Figure 8. We set the variation range of the axis to 0.6–1.0, and the corresponding color varies from dark red to white to dark blue. The lower the correlation coefficient, the closer the color will be to dark red, and conversely, the higher the correlation coefficient, the closer the color will be to dark blue. Results illustrate that the correlation coefficients are all greater than 0.6, which indicates that the correlation relationships between the settlement datasets of monitoring points are relatively high. This result is consistent with the characteristics in the settlement process lines shown in Figure 6. Among all sixteen monitoring points, LD3-5 has the smallest average value of the correlation coefficient (0.728) compared with other monitoring points, and LD3-3 has the largest average value of the correlation coefficient (0.890).
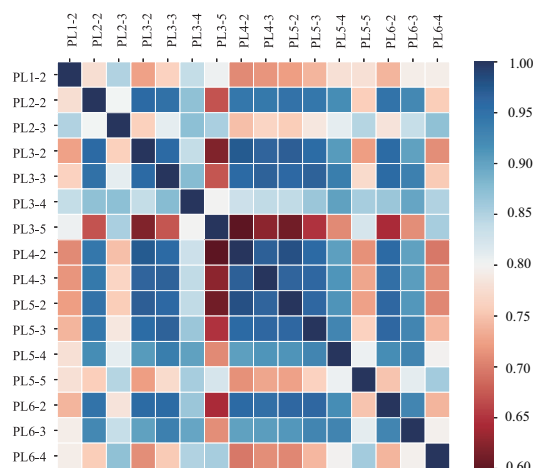
Figure 8. Results of correlation coefficients of the monitoring settlement data.

Then, we construct the global correlation measure ω_i and classify the monitoring data series by the *K*-means++ algorithm. The number of clusters is set as 4. The results of clustering are presented in Figure 9. The monitoring points LD1-2 and LD3-5 are categorized in Cluster 1. LD2-2, LD3-3, LD5-4 and LD5-4 are categorized in Cluster 2. LD2-3, LD3-4, LD5-5, and LD6-4 are classified to Cluster 3. LD3-2, LD4-2, LD4-3, LD5-2, LD5-3, and LD6-2 are classified to Cluster 4. It is noted that the classification is roughly related to the spatial locations; monitoring points located at the boundary (i.e., LD1-2, LD3-5, LD2-3, LD5-5, and LD6-4) are categorized in Cluster 1 and Cluster 3 and are separated apart with other inner monitoring points.
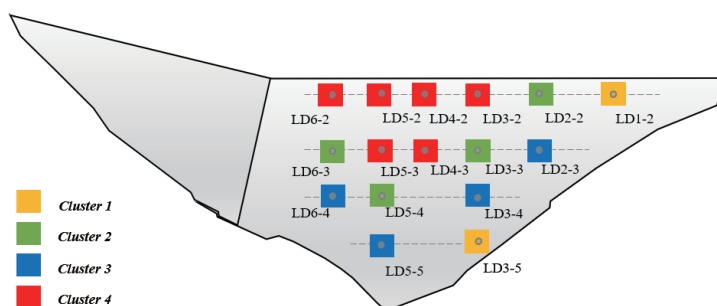

Figure 9. Monitoring points distribution indicated clustering results.

5.2. Optimization of Parameters in LSTM

In this work, we adopt the CS algorithm to optimize the parameters (including the number of hidden layers and hidden nodes and the learning rate) of the neural network in the LSTM model.

In the CS algorithm, we set population size as 20, β as 1.5, and number of iterations as 30. Figure 10 exhibits the objective function value of the best nest at each iteration step, and the algorithm finds the best objective function value as 4.333 at the 13th iteration step. The set of upper bound and lower bound of the search area for the three parameters are presented in Table 1. After calculation, we obtain the results of the optimized parameters as presented in Table 1.
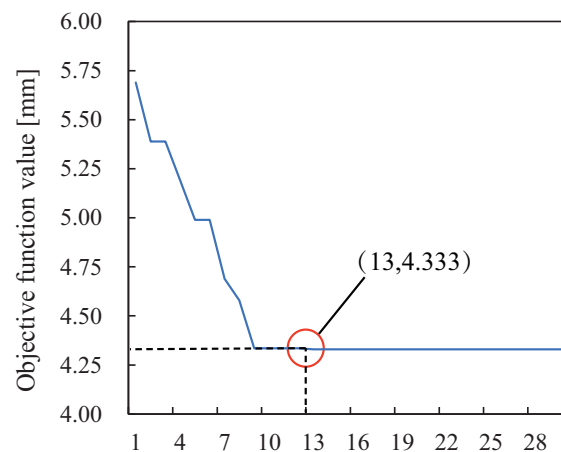
Figure 10. Plot of the objective function values versus the number of iterations.

Table 1. Upper bound and lower bound of the parameters set in the CS algorithm, and the results of optimized parameters of the NN structure in LSTM model.

	Hidden Layers	**Hidden Nodes**	**Learning Rate**
Upper bound	10	20	1×10^{-1}
Lower bound	1	1	1×10^{-5}
Optimized parameters	2	7	3×10^{-2}

5.3. Fitting and Prediction Results

5.3.1. Selection of the Input Variables in the Model

In the interpretation model for the settlement of CFRDs, environmental temperature commonly has a weak impact on the deformation of the CFRD, and a significant settlement variation occurred only when frost heaving happened. Hereby, the expressions of settlement model of the PSPS considering different temperature conditions are exhibited as follows.

(1) When the lowest temperature is below zero, then

$$\delta = \sum_{i=0}^{3} a_{1i}H^i + \sum_{i=0}^{m_1} a_{2i}\overline{H}^i + \sum_{i=1}^{m_2} b_1 T_i + \sum_{i=1}^{m_3} (b_{1i}\sin\frac{2\pi it}{365} + b_{2i}\cos\frac{2\pi it}{365}) + C_1\theta + C_2\ln\theta$$
$$+ \frac{\theta}{C_3\theta + C_4} + C_5 e^{\frac{C_6}{\theta}} + cons \tag{25}$$

(2) When the lowest temperature is above zero, then

$$\delta = \sum_{i=0}^{3} a_{1i}H^i + \sum_{i=0}^{m_1} a_{2i}\overline{H}^i + C_1\theta + C_2\ln\theta + \frac{\theta}{C_3\theta + C_4} + C_5 e^{\frac{C_6}{\theta}} + cons \tag{26}$$

The variables in Equations (25) and (26) have the same meaning as the variables in Equations (1)–(5). Owing to the lowest temperature being above zero at the region of Langyashan PSPS according to the ambient data, we adopt variables in the second expression above as the input variables in the training models.

In Equation (26), the average values of the previous upstream water level is adopted to describe the previous conditions, which will act as input variables, and the expression is:

$$\overline{H}_{pre(\Delta d)} = \frac{\sum_{i=0}^{d-1} H(t-i)}{\Delta d} \tag{27}$$

where $\overline{H}_{pre(\Delta d)}$ denotes the average value of the upstream water level of d days before the monitoring date and t denotes the current date. In this case, we apply $\overline{H}_{pre(\Delta 7)}$, $\overline{H}_{pre(\Delta 28)}$, $\overline{H}_{pre(\Delta 84)}$, and $\overline{H}_{pre(\Delta 168)}$ as the input variables in the models.

5.3.2. Fitting and Prediction Performance of M–LSTM Model

We apply the CS–LSTM model for establishing the settlement prediction model of multiple monitoring points in this section. Based on the clustering results, we select monitoring data series belonging to the same cluster as output variables to the proposed model for calculation. Monitoring data series are divided into a training dataset (80% of the data) and testing dataset (20% of the data). We fit the settlement data from 17 April 2005 to 10 January 2016 to develop the prediction model. Then, we test the proposed model using data from 17 January 2016 to 16 September 2018. Figure 11 exhibits the fitting and prediction results of four monitoring points: PL1-2 in Cluster 1 (lines in purple), PL2-2 in Cluster 2 (lines in green), PL3-4 in Cluster 3 (lines in blue), and PL3-2 in Cluster 4 (lines in red). The training datasets are exhibited in the white area, and the testing datasets are located in the gray area. Even though the monitoring settlement data fluctuates intensely, the results of the proposed model are close to the monitoring data, and the fitting and prediction results of all monitoring points are fairly good. In addition, the fitting and prediction results of the proposed model of the rest of the 12 monitoring points are illustrated in Appendix A (Figures A1–A4).
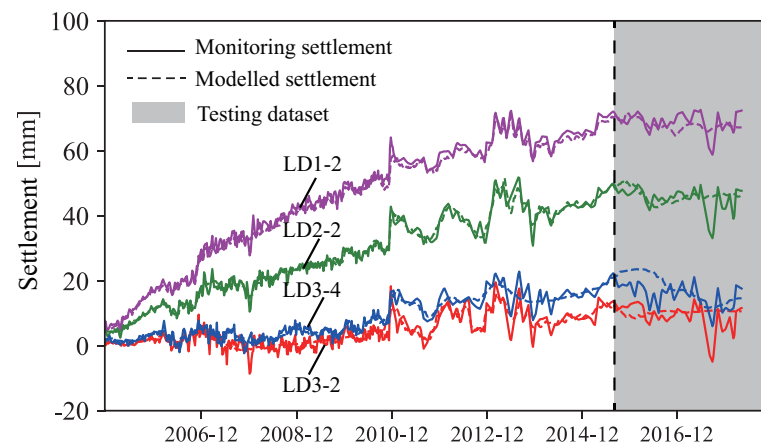

Figure 11. Fitting and prediction results of PL1-2, PL2-2, PL3-2, and PL3-4.

Furthermore, to quantify the fitting and prediction performance of the proposed model, we adopt the coefficient of determination (R^2) of the training dataset and the root mean square error ($RMSE$) of the testing dataset for evaluation. The expressions of the two indicators are as follows:

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{\delta}_i - \overline{\delta})}{\sum_{i=1}^{N}(\delta_i - \overline{\delta})} \tag{28}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(\delta_j - \hat{\delta}_j)^2} \tag{29}$$

where δ_i denotes the monitoring settlement of the training dataset, $\overline{\delta}$ denotes the average of the monitoring settlement of the training dataset, $\hat{\delta}_i$ denotes the modeled settlement of the training dataset, δ_j denotes the monitoring settlement of the testing dataset, $\hat{\delta}_j$ denotes the modeled settlement of the testing dataset, and N and M denote the number of data in the training and testing dataset, respectively.

Figure 12 shows the R^2 and $RMSE$ of each monitoring point. The blue dots and green bars represent the results of R^2 and $RMSE$, respectively. The results of R^2 and $RMSE$ vary from 0.800 to 0.995 and from 1.439 to 2.702, respectively, and the average of R^2 and $RMSE$

are 0.945 and 2.055. R^2 of all monitoring points are above 0.8 (as shown in the red area), which indicates that the proposed model has a high fitting precision.
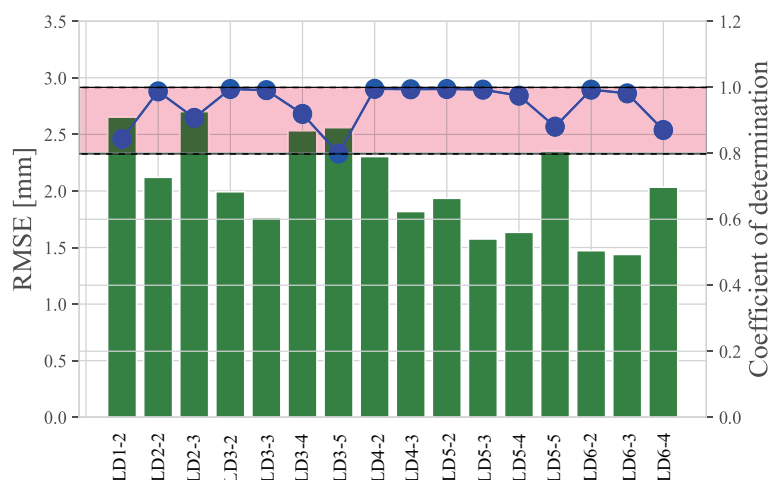

Figure 12. Coefficient of determination and *RMSE* of the proposed model.

5.3.3. Comparison between Different Models

In order to verify the superiority of the proposed model in fitting and forecasting, we compare our model with three widely used single monitoring point statistical models. Similarly, we apply 80% of the data as the training dataset, using LSTM, back propagation neural network (BPNN), and hydrostatic–seasonal–time (HST), to fit the monitoring settlement and predict the remaining 20% of the data. Figure 13 exhibits the fitting and prediction results of four models for one monitoring point (LD3-3). In the figure, fitting curves of all four models can fit the monitoring settlement reasonably well, and they can reach most monitoring data closely. In settlement monitoring data, there exists noisy data which fluctuates in several time periods, and these data are not overfitted in all four models. Compared with the other three models, the proposed model has the highest R^2 (0.991) and the lowest *RMSE* (1.763), demonstrating the proposed model has good fitting ability and prediction performance among these four models.
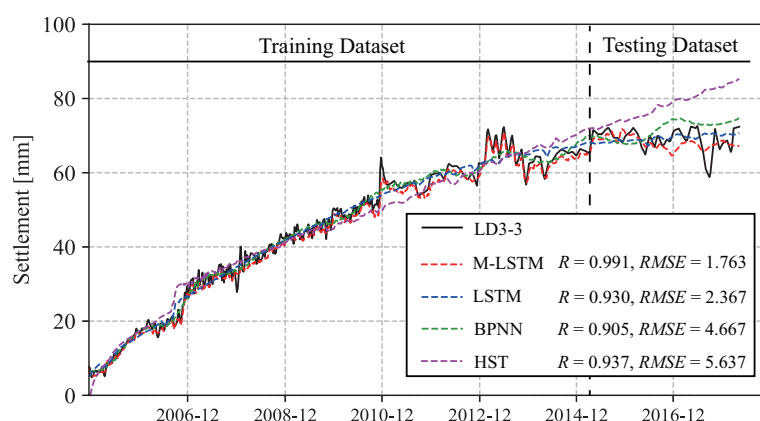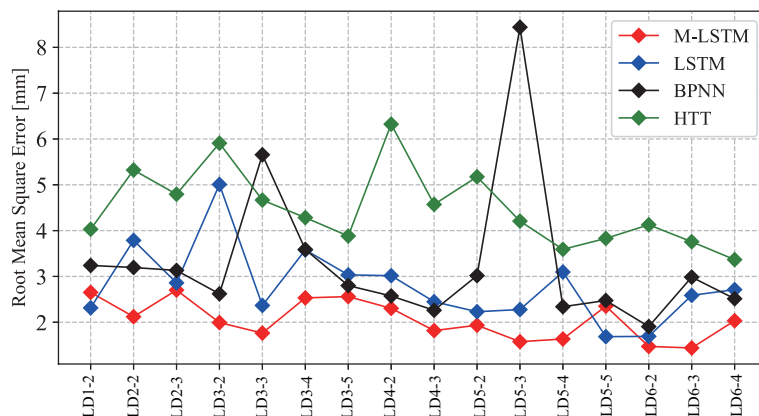

Figure 13. Fitting and prediction results of four models for LD3-3.

We establish the four models on the training dataset of each monitoring point. Table 2 lists the results of R^2 calculated for each monitoring point. R^2 of M–LSTM, LSTM, BPNN, and HST varies from 0.995 to 0.800, from 0.991 to 0.741, from 0.994 to 0.753, and from 0.962 to 0.409, respectively. For all sixteen monitoring points, the averages of R^2 are 0.945 (M–LSTM), 0.910 (LSTM), 0.895 (BPNN), and 0.807 (HST). Results indicate that all four models can fit the training dataset of these 16 monitoring points well.

Table 2. R^2 of M–LSTM, LSTM, BPNN, and HST models.

Monitoring Point	M–LSTM	LSTM	BPNN	HST
LD1-2	0.843	0.985	0.754	0.576
LD2-2	0.987	0.807	0.971	0.918
LD2-3	0.907	0.989	0.873	0.665
LD3-2	0.995	0.926	0.991	0.951
LD3-3	0.991	0.93	0.905	0.937
LD3-4	0.919	0.541	0.829	0.745
LD3-5	0.799	0.991	0.753	0.409
LD4-2	0.995	0.983	0.994	0.962
LD4-3	0.994	0.991	0.99	0.96
LD5-2	0.995	0.985	0.987	0.961
LD5-3	0.993	0.95	0.761	0.946
LD5-4	0.974	0.781	0.947	0.87
LD5-5	0.88	0.99	0.867	0.635
LD6-2	0.993	0.974	0.988	0.941
LD6-3	0.981	0.779	0.917	0.865
LD6-4	0.87	0.754	0.793	0.579

We then calculate the *RMSE* of each testing dataset of the monitoring point, and the results are exhibited in Figure 14. In terms of the prediction performance, the *RMSE* of M–LSTM, LSTM, BPNN, and HST model varies from 1.431 to 2.683, from 1.853 to 3.185, and from 1.759 to 7.370, respectively. Among all 16 monitoring points, the proposed model has the smallest *RMSE* in 14 monitoring points, which means the proposed method has the best prediction performance. With the consideration of previous monitoring data, the prediction accuracy of the model could be increased.

Figure 14. *RMSE* of each monitoring point based on M–LSTM, LSTM, BPNN, and HST.

In the following, we verify the prediction performance of each model with 70% of data as the training dataset, and 20%, 25%, and 30% of the data as the testing dataset, respectively. In order to reflect the predictive ability of the model comprehensively, we propose the average of MSE, SMAPE, and MAPE for all monitoring points as the evaluation indicators for prediction. The expressions are exhibited as follows:

$$AMSE = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} (\delta_{ij} - \hat{\delta}_{ij})^2 \tag{30}$$

$$ASMAPE = \frac{100\%}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{|\hat{\delta}_{ij} - \delta_{ij}|}{(|\hat{\delta}_{ij}| + |\delta_{ij}|)/2} \tag{31}$$

$$AMAPE = \frac{100\%}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{|\hat{\delta}_{ij} - \delta_{ij}|}{|\delta_{ij}|} \tag{32}$$

where *AMSE* denotes the average of mean square error, *ASMAPE* denotes the average of symmetric mean absolute percentage error, and *AMAPE* denotes the average of mean absolute percentage error; other variables in the Equations (30)–(32) denote the same meaning as the variables in Equations (28) and (29).

Figure 15a–c illustrate the results of prediction evaluation indicators (*AMSE*, *ASMAPE*, and *AMAPE*) of the four models under different prediction testing datasets (20%, 25%, and 30% of the data) after modeling the sixteen monitoring points in the example.

Figure 15. Results of: (**a**) *AMSE* of M–LSTM, LSTM, BPNN, and HST models under different prediction lengths, (**b**) *ASMAPE* of M–LSTM, LSTM, BPNN, and HST models under different prediction lengths, (**c**) *AMAPE* of M–LSTM, LSTM, BPNN, and HST models under different prediction lengths.

With the increase of the length of the testing dataset, the values of *AMSE*, *ASMAPE*, and *AMAPE* grow, which indicates that the prediction performance of the four models will become weaker with the extension of the prediction length. Specifically, for *AMSE* and *ASMAPE*, with increasing prediction length (from 20% to 30%), the results of M–LSTM are 9.10, 20.00, and 38.96 and 0.15, 0.20, and 0.34, respectively; the results of HST are 48.77, 57.48, and 88.51 and 0.24, 0.25, and 0.67, respectively. The results of other models are located between the results of M–LSTM and HST. For *AMAPE* under different prediction lengths, the minimum values in the four models are the results of M–LSTM, which are 0.08, 0.12, and 0.30, respectively, and the maximum values are 0.20 (20%) of HST, 0.25 (25%) of BPNN, and 0.60 (30%) of HST.

The M–LSTM model obtains a more accurate result than LSTM, BPNN, and HST models with a single monitoring point with the expansion of prediction lengths. Different from conventional neural networks, the nodes of hidden layers in LSTM models are related to others at the previous time step. The links make the hidden layer contain history of the input information, which means the LSTM model can utilize the previous information fully. On the other hand, the LSTM model with multiple correlated settlement data is less prone to overfitting local noisy data. Vulnerability to noisy data is a common issue in the single monitoring model. M–LSTM utilize settlement data series with a correlated relationship simultaneously; the influence from local noisy data in certain settlement data series will be weakened by most correct data samples when modeling, therefore the M–LSTM model can be more robust.

6. Conclusions

Accurate prediction results of settlement is significant to the evaluation of the structural health status of CFRDs during operation. In this research, a method based on the LSTM model is proposed to predict the settlement of the CFRD. Firstly, we classified the settlement points into several categories according to the correlation relationship among the data series. Then, the LSTM model was utilized to train all monitoring settlement datasets in the same category simultaneously; in this way, the lag effect from input variables, settlement condition, and the correlation relationship among monitoring data series can be considered. In addition, the CS algorithm was adopted to optimize the parameters of the neural network structure in LSTM models in advance.

We used the monitoring settlement data series of CFRD at Langyashan PSPS to validate the proposed model. The coefficients of determination R^2 of the proposed model are above 0.85 for the training dataset of all monitoring points. Compared with the LSTM, BPNN, and HST models of a single monitoring point, the proposed model has the most precise prediction results, which has the smallest *RMSE* in most (14 out of 16) monitoring points. Furthermore, the proposed model can provide the most stable and accurate results with the expansion of the prediction length. In future research, emphasis will be given on the exploration of an optimized LSTM model which is capable of obtaining a higher prediction accuracy, and the proposed method will have promising applications in the prediction of data series for other types of civil structures.

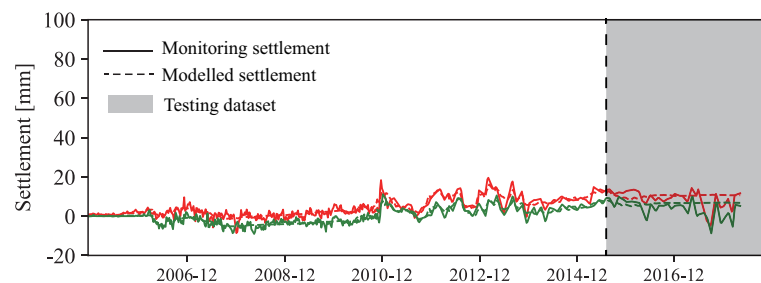## Appendix A. Fitting and Predicting Results

Figure A1. Fitting and prediction results of M–LSTM model based on the training datasets in Cluster 1 (LD1-2 in red lines and LD3-5 in green lines).
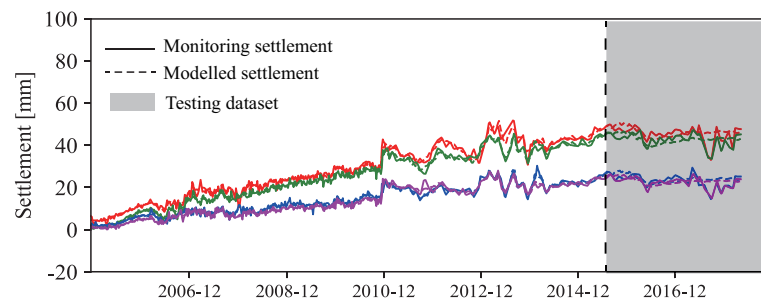

Figure A2. Fitting and prediction results of M–LSTM model based on the training datasets in Cluster 2 (LD2-2 in red, LD3-3 in green, LD5-4 in blue, and LD6-3 in purple lines, respectively).
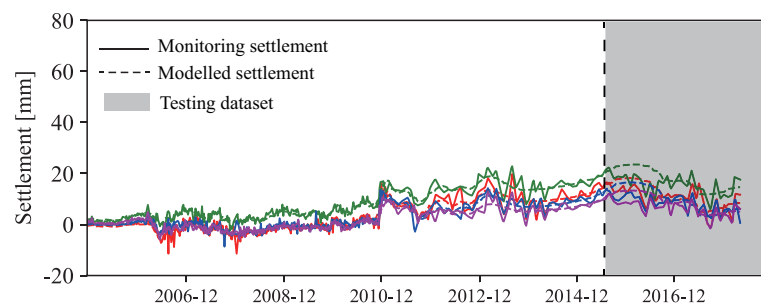
Figure A3. Fitting and prediction results of M–LSTM model based on the training datasets in Cluster 3 (LD2-3 in red, LD3-4 in green, LD5-5 in blue, and LD6-4 in purple lines, respectively).
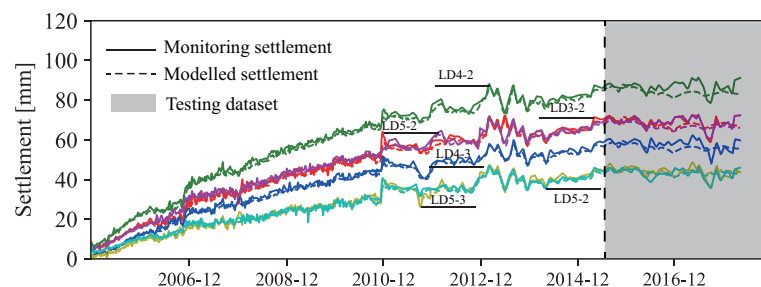

Figure A4. Fitting and prediction results of M–LSTM model based on the training datasets in Cluster 4 (LD3-2, LD4-2, LD4-3, LD5-2, LD5-3, and LD6-2).

References

1. Xu, B.; Zou, D.; Liu, H. Three-dimensional simulation of the construction process of the Zipingpu concrete face rockfill dam based on a generalized plasticity model. *Comput. Geotech.* **2012**, *43*, 143–154. [CrossRef]
2. Wen, L.; Chai, J.Z.; Qin, Y. A statistical review of the behaviour of concrete-face rockfill dams based on case histories. *Geotechnique* **2018**, *68*, 749–771. [CrossRef]
3. Yuan, D.; Wei, B.; Xie, B.; Zhong, Z. Modified dam deformation monitoring model considering periodic component contained in residual sequence. *Struct. Control Health Monit.* **2020**, *27*, e2633. [CrossRef]
4. Yuan, D.; Gu, C.; Qin, X.; Shao, C.; He, J. Performance-improved TSVR-based DHM model of super high arch dams using measured air temperature. *Eng. Struct.* **2022**, *250*, 113400. [CrossRef]
5. Ma, H.; Chi, F. Technical Progress on Researches for the Safety of High Concrete-Faced Rockfill Dams. In *Engineering*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 332–339.
6. Behnia, D.; Ahangari, K.; Noorzad, A.; Moeinossadat, S.R. Predicting crest settlement in concrete face rockfill dams using adaptive neuro-fuzzy inference system and gene expression programming intelligent methods. *J. Zhejiang Univ.–Sci. A Appl. Phys. Eng.* **2013**, *14*, 58–71. [CrossRef]
7. Leger, P.; Leclerc, M. Hydrostatic, Temperature, Time-Displacement Model for Concrete Dams. *J. Eng. Mech.* **2007**, *133*, 267–277. [CrossRef]
8. Mata, J.; Castro, A.T.D.; Costa, J.S.D. Constructing statistical models for arch dam deformation. *Struct. Control Health Monit.* **2014**, *21*, 423–437. [CrossRef]
9. Leger, P.; Cote, P.; Tinawi, R. Finite element analysis of concrete swelling due to alkali-aggregate reactions in dams. *Comput. Struct.* **1996**, *60*, 601–611. [CrossRef]
10. Hu, Y.; Gu, C.; Meng, Z.; Shao, C. Improve the Model Stability of Dam's Displacement Prediction Using a Numerical-Statistical Combined Model. *IEEE Access.* **2020**, *8*, 147482–147493. [CrossRef]
11. Sun, P.M.; Bao, T.F.; Gu, C.S.; Jiang, M.; Wang, T.; Shi, Z.W. Parameter sensitivity and inversion analysis of a concrete faced rock-fill dam based on HS-BPNN algorithm. *Sci. China (Technol. Sci.)* **2016**, *59*, 1442. [CrossRef]
12. Meng, Z.; Hu, Y.; Ancey, C. Using a Data Driven Approach to Predict Waves Generated by Gravity Driven Mass Flows. *Water* **2020**, *12*, 600. [CrossRef]
13. Hariri-Ardebili, M.A.; Pourkamali-Anaraki, F. Support vector machine based reliability analysis of concrete dams. *Soil Dyn. Earthq. Eng.* **2018**, *104*, 276–295. [CrossRef]
14. Su, H.; Li, X.; Yang, B.; Wen, Z. Wavelet support vector machine-based prediction model of dam deformation. *Mech. Syst. Signal Process.* **2018**, *110*, 412–427. [CrossRef]
15. Tabari, M.; Sanayei, H. Prediction of the intermediate block displacement of the dam crest using artificial neural network and support vector regression models. In *Soft Computing—A Fusion of Foundations, Methodologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2019.

16. Salazar, F.; Moran, R.; Toledo, M.A.; Onate, E. Data-Based Models for the Prediction of Dam Behaviour: A Review and Some Methodological Considerations. *Arch. Comput. Methods Eng.* **2017**, *24*, 1–21. [CrossRef]
17. Wang, S.; Xu, C.; Liu, Y.; Wu, B. A spatial association-coupled double objective support vector machine prediction model for diagnosing the deformation behaviour of high arch dams. *Struct. Health Monit.* **2021**, *21*, 14759217211017030. [CrossRef]
18. Hu, Y.; Shao, C.; Gu, C.; Meng, Z. Concrete Dam Displacement Prediction Based on an ISODATA-GMM Clustering and Random Coefficient Model. *Water* **2019**, *11*, 714. [CrossRef]
19. Adnan, R.M.; Khosravinia, P.; Karimi, B.; Kisi, O. Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline. *Appl. Soft Comput.* **2020**, *100*, 107008. [CrossRef]
20. Khaledian, E.; Pandey, S.; Kundu, P.; Srivastava, A.K. Real-Time Synchrophasor Data Anomaly Detection and Classification Using Isolation Forest, KMeans, and LoOP. *IEEE Trans. Smart Grid* **2020**, *12*, 1949–3053. [CrossRef]
21. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Abdel-Nasser, M.; Mahmoud, K. Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Comput. Appl.* **2019**, *31*, 2727–2740. [CrossRef]
23. Altan, A.; Karasu, S.; Zio, E. A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. *Appl. Soft Comput.* **2021**, *100*, 106996. [CrossRef]
24. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
25. Min, H.; Xi, J.; Xu, S.; Yin, F.L. Prediction of chaotic time series based on the recurrent predictor neural network. *IEEE Trans. Signal Process.* **2004**, *52*, 3409–3416.
26. Yang, X.S.; Deb, S. Cuckoo Search via Levy Flights. In Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), Coimbatore, India, 9–11 December 2009.
27. Singh, H.; Kommuri, S.; Kumar, A.; Bajaj, V. A new technique for Guided Filter based Image Denoising using Modified Cuckoo Search Optimization. *Expert Syst. Appl.* **2021**, *176*, 114884. [CrossRef]
28. Haifanga, L.; Yinqi, Z. Creep rate and creep model of rockfill. *Procedia Eng.* **2012**, *28*, 796–802. [CrossRef]
29. Janbu, N. Soil models in offshore engineering. *Géotechnique* **1985**, *35*, 241–281. [CrossRef]
30. Oyen, M.L.; Cook, R.F. Load–displacement behavior during sharp indentation of viscous–elastic–plastic materials. *J. Mater. Res.* **2003**, *18*, 139–150. [CrossRef]
31. Sigtryggsdóttir, F.G.; Snæbjörnsson, J.T.; Grande, L. Statistical model for dam-settlement prediction and structural-health assessment. *J. Geotech. Geoenviron. Eng.* **2018**, *144*, 04018059. [CrossRef]