Earthquakes and Structures, *Vol. 14, No. 2 (2018) 143-153*
DOI: https://doi.org/10.12989/eas.2018.14.2.143

Copula entropy and information diffusion theory-based new prediction method for high dam monitoring

Dongjian Zheng[1,2,3a], Xiaoqi Li[1,2,3], Meng Yang[*1,2,3], Huaizhi Su[1,2,3] and Chongshi Gu[1,2,3]

[1]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098, China
[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, Hohai University, Nanjing 210098, China

(Received June 6, 2017, Revised January 25, 2018, Accepted January 30, 2018)

Abstract. Correlation among different factors must be considered for selection of influencing factors in safety monitoring of high dam including positive correlation of variables. Therefore, a new factor selection method was constructed based on Copula entropy and mutual information theory, which was deduced and optimized. Considering the small sample size in high dam monitoring and distribution of daily monitoring samples, a computing method that avoids causality of structure as much as possible is needed. The two-dimensional normal information diffusion and fuzzy reasoning of pattern recognition field are based on the weight theory, which avoids complicated causes of the studying structure. Hence, it is used to dam safety monitoring field and simplified, which increases sample information appropriately. Next, a complete system integrating high dam monitoring and uncertainty prediction method was established by combining Copula entropy theory and information diffusion theory. Finally, the proposed method was applied in seepage monitoring of Nuozhadu clay core-wall rockfill dam. Its selection of influencing factors and processing of sample data were compared with different models. Results demonstrated that the proposed method increases the prediction accuracy to some extent.

Keywords: high dam monitoring; composite uncertain information; Copula entropy; information diffusion; fuzzy reasoning

1. Introduction

With the progress of hydraulic construction technology and booming of hydropower industry in developing countries, construction of high dam or super-high dam has become a trend. Since high dam is facing with more complicated stress environment, larger bearing load and even coupling effect of multiple factors, it proposes higher requirements on dam safety and has to make a more accurate and timely health diagnosis and analysis (Novak *et al*. 2007, Akpinar *et al*. 2014, Hassanvand Jamadi *et al*. 2017).

Moreover, since high dam generally has high water level and faces with higher hydraulic pressure and stress conditions of dam abutment and dam heel, the selection of model input factors is more sensitive to influences of prediction results, which has to consider more correlation factors and make correlation analysis more accurately. Meanwhile, many pre-buried monitors have been destroyed or re-installed in the long construction period of dam, resulting in non-continuous monitoring sequence and short sample sequence. Particularly, monitoring sample couldn't meet normal distribution completely in the storage period and conventional analysis method couldn't make effective analysis and prediction. Therefore, it is necessary to explore a prediction method of high-dam monitoring information which is applicable to coupling influences of multiple uncertain factors.

With respect to selection of correlation factors, the principal component analysis (PCA) has low index accuracy when there's poor colinearity of factors (Tao *et al*. 2011, Yitao *et al*. 2011). The fuzzy membership analysis couldn't select membership function and evaluation indexes under influences of multiple factors clearly (Opyrchał *et al*. 2003). The stepwise regression method deletes factors through the significance test and selection of deleted indexes will affect selection of correlation factors. Deleted indexes shall be chosen tentatively according to practical situations. Influences of multiple uncertain factors will cause big errors (Shen *et al*. 2014, Wen *et al*. 2013, Hu *et al*. 2011). Grey relational analysis considers grey uncertainty of factors, but has no definite indexes to select and delete factors (Liu *et al*. 2010, Liu *et al*. 2011). In this paper, the Copula function for correlation analysis in hydraulic field was applied to safety monitoring field of dam and its computing method was improved (Li *et al*. 2016, Allahdadi *et al*. 2017). Next, a high-accuracy factor selection method was established by combining partial mutual information theory and Hampel test.

For small sample problems under influences of multiple factors, the neural network method has big errors due to the small training sample size and irregular distribution (Maier *et al*. 2010, Valipour *et al*. 2013). The support vector

∗Corresponding author, Ph.D.
E-mail: ymym_059@126.com
[a]Professor
E-mail: zhengdj@hhu.edu.cn

ISSN: 2092-7614 (Print), 2092-7622 (Online)

machine can support nonlinear data and small sample operation, but it is only applicable to data classification and is inapplicable to accurate prediction of points (Hipni *et al.* 2013, Deka *et al.* 2014). Although fuzzy reasoning has simple computation and avoids causality of research factors, its selection of membership function influences the accuracy of results (Su *et al.* 2011). Rough set theory is novel and can process incomplete and uncertain information, but its theoretical process is imperfect and has to cooperate with fuzzy set and genetic algorithm (Liu *et al.* 2011). Based on advantages of above algorithms, the combination of grey correlation and fuzzy reasoning was proposed and the fuzzy reasoning was improved. Finally, a small sample processing method based on the combination of information diffusion theory and fuzzy reasoning theory was constructed (Zheng *et al.* 2013, Li *et al.* 2016, Yang *et al.* 2017).

Input factor set was determined by the improved Copula entropy. Samples of each input factor were expanded and processed by the improved information diffusion method, thus establishing a more efficient method for processing monitoring information of high dam. Finally, the proposed method was verified by the actual monitoring data of a ultrahigh earth-rock dam Nuozhadu Dam in the storage period.

2. Influence factor selection

2.1 Quantitative analysis

Factor selection significantly influences the accuracy of prediction for high dam monitoring. According to the in situ data of dam monitoring, the first-level factor set Z_1 contains most of the factors that need to be further removed. This set can be determined by feature analysis, structure analysis, and expert knowledge base. The selection of Z_1 during the water storage period in this paper mainly considers the influence of environmental factors, such as water level, temperature, aging, and the influence of structural factors such as the height of dam filling, the displacement of dam body and the seepage pressure, etc.

In the traditional correlation analysis, the correlation coefficient is used to determine the influence factors. However, this method lacks significant testing and correlation analysis between the factors. Rank correlation coefficients can be used to estimate the nonlinear correlation between the variables without requiring the distribution of the variables. Therefore, based on the factor set Z_1, Spearman correlation coefficient r and Kendall correlation coefficient τ are used to determine the second-level factor set Z_2 (Croux *et al.* 2010). r and τ are described as follows

$$r = 1 - 6\sum_{j=1}^{n} \alpha_i^2 \Big/ \left(N^3 - N\right) \tag{1}$$

Where N denotes the sample number and d_i denotes the difference between the numbers of variables sorted by time.

$$\tau = \frac{N_c - N_\alpha}{n\left(n-1\right)/2} \tag{2}$$

Where N_c+N_α≤n(n−1)/2, N_c denotes the number of pairs, N_α denotes the number of noncooperative pairs, and n is the number of data in a sample.

2.2 Quantitative analysis

The probability distribution function can be used to express the uncertainty of a random event. The dam monitoring information is usually a continuous random variable. The joint entropy of multivariate random variables can be expressed as follows (Da Silva *et al.* 2016)

$$\begin{gathered} H(x_1,..., x_n) = -\int_0^\infty \cdots \int_0^\infty f(x_1,..., x_n) \\ \log[f(x_1,..., x_n)]dx_1 dx_2 \ldots dx_n \end{gathered} \tag{3}$$

Where $f(x)$ denotes the probability density function of variable x, which can be expressed by the partial differential of the distribution function.

The Copula function is a class of functions that connect the joint distribution function with their respective edge distribution functions. The Copula function is defined as follows

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)), \tag{4}$$

Where F_i is the edge distribution function of each variable and x_n denotes the n-th random variable. The definition domain of x is [0,1].

The Archimedean Copula function has the advantages of convenient construction and simple calculation and is widely used in the research on multidimensional variables (Hofert Silva *et al.* 2012). In this study, the Gumbel Copula function is used to describe the positive correlation between the increase of variables and the increase of effect quantities. To facilitate the calculation, the two-dimensional (2D) joint distribution function and density function are used. These functions can be expressed as follows

$$C(u,v) = \exp\left\{-\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{1/\theta}\right\}, \theta \in [1,\infty) \tag{5}$$

$$\begin{aligned} c(u,v) &= C(u,v) \\ &\times \frac{\left[(-\ln u)(-\ln v)\right]^{\theta-1}}{uv}\left[(-\ln u) + (-\ln v)^\theta\right]^{\frac{2}{\theta}-2} \\ &\times \left\{(\theta-1)\left[(-\ln u)(\ln v)\right]^{\theta-1} + 1\right\} \end{aligned} \tag{6}$$

Given that u=F(x) and v=F(y), Copula entropy can be expressed as follows

$$H_C(u,v) = -\int_0^1\int_0^1 c(u,v)\log c(u,v)dudv \tag{7}$$

The calculation of Copula entropy can be converted to calculate the parameter θ of the Copula joint distribution function and the edge distribution F of variables x and y. θ can be expressed as the following relationship with τ

$$\theta = 1/(1-\tau) \tag{8}$$

Cauchy distribution can truly reflect the distribution of random vectors and can easily avoid the phenomenon of local optimal value (GCordeiro Silva *et al.* 2011). Based on the previously presented analysis, Cauchy distribution function is used as the edge distribution function. The probability density function and distribution function of Cauchy distribution are calculated as follows

$$f(x\,;\,\mu,\gamma)=\frac{1}{\pi}\left[\frac{\gamma}{(x-\mu)+\gamma^{2}}\right] \tag{9}$$

$$F(x\,;\,\mu,\gamma)=\frac{1}{\pi}\arctan\frac{x-\mu}{\gamma}+\frac{1}{2} \tag{10}$$

Where μ denotes the location parameter for peak values of Cauchy distribution and γ is the scale parameter for the half-maximum value.

We define $F_x(\mu)=1/2$ and $F_x(\mu+\gamma)=3/4$. Cauchy distribution is a continuous distribution function with no expectation and variance. μ and γ can be estimated using the median and the quantile, respectively (Pekasiewicz *et al.* 2014).

The Gumbel distribution function and density function can be obtained by using parameter θ, and the joint distribution function $C(\gamma,\mu)$ can be calculated by using parameters μ and γ of the edge distribution function $F(x)$. Finally, the combined distribution density function is brought into Eq. (7) to obtain the value of Copula entropy.

2.3 Partial mutual information (PMI) calculation

2.3.1 PMI calculation based on Copula entropy

After setting a certain filter standard, the final input factor set Z_3 can be obtained by calculating the PMI value C_{PMI} of variables in Z_2. We assume that the 2D variables are x and y; thus, C_{PMI} can be calculated as follows (Chen *et al.* 2014)

$$C_{PMI}=H(x')+H(y')-H(x',y') \tag{11}$$

Where ln-log$_e$, $x'=x-E\left[x\middle|Z_2\right]$, $y'=y-E\left[y\middle|Z_2\right]$, $E[\cdot]$denotes the expectation value, x', y' represents the residual information of x and y under the consideration of factor set Z_2. The relationship between C_{PMI} and the 2D joint entropy is shown in Fig. 1.

According to the definition, the joint entropy of multidimensional variables can be expressed by the summation of Copula entropy and edge entropy of n-dimension variables. Given that $du=dxf(x)$, $dv=dyf(y)$, and $f(x,y)=c(u,v)f(x)f(y)$, the expression of 2D joint entropy can be described as follows

$$H(X,Y)=-\int_0^{\infty}\int_0^{\infty}f(x,y)\ln f(x,y)dxdy \tag{12}$$

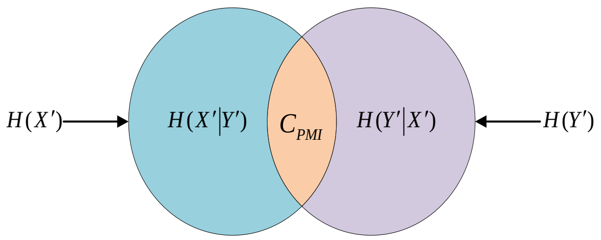


Fig. 1 Relationship between C_{PMI} and entropy of variables x and y

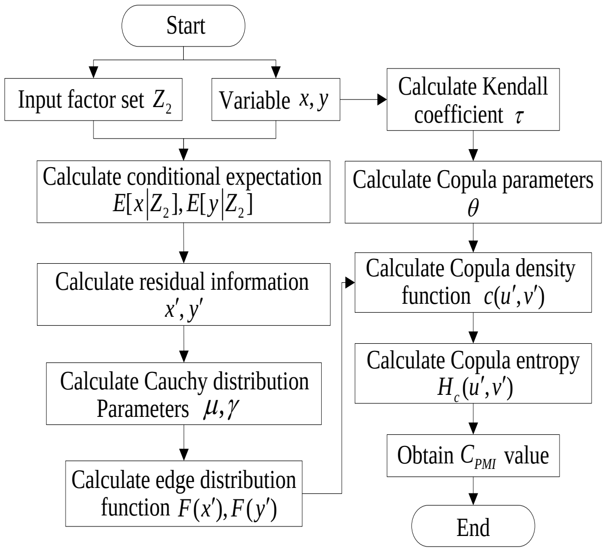

Start
Input factor set Z_2
Variable x, y
Calculate Kendall coefficient τ
Calculate conditional expectation $E[x|Z_2], E[y|Z_2]$
Calculate Copula parameters θ
Calculate residual information x', y'
Calculate Copula density function $c(u',v')$
Calculate Cauchy distribution Parameters μ,γ
Calculate Copula entropy $H_c(u',v')$
Obtain C_{PMI} value
Calculate edge distribution function $F(x'),F(y')$
End

Fig. 2 Calculation flowchart of C_{PMI}

For the case of only two variables x' and y', (12) can be simplified as follows

$$H(x',y')=H(x')+H(y')+H_C(u',v') \tag{13}$$

Through Eq. (11) and Eq. (13), we determine that $C_{PMI}=-H_C(u',v')$, indicating that C_{PMI} is the negative entropy of the Copula function. The process of C_{PMI} calculation with Copula entropy is shown in Fig. 2.

2.3.2 Factor selection criteria based on PMI

The PMI algorithm requires a criterion to determine how large the C_{PMI} will be when the variable x can be incorporated into Z_3. The Hampel test is used as the stopping criterion of the algorithm. This criterion can measure C_{PMI} of a variable in a set of variables and can determine whether the value is significantly higher than that of the other variables or not. The Hampel test can be expressed as follows

$$\begin{aligned} H_j &= \frac{d_j}{1.4826 d_j^{(50)}} \\ d_j &= \left|C_{PMI}-C_{PMI}{}^{(50)}\right| \end{aligned} \tag{14}$$

Where 1.4826 denotes the normalized constant, which makes H_j equal to σ; $d_j^{(50)}$ denotes the median of d_j; and $C_{PMI}{}^{(50)}$ denotes the median of the C_{PMI} values in a set of data. According to the 3σ criteria of the standard deviation, factors are integrated into Z_3 when $H_j>3$.

Therefore, combined with Copula entropy and PMI algorithm, an influence factor selection method is established.

3. Influence factor selection

Information diffusion is a fuzzy mathematical algorithm, which transforms traditional sample points into fuzzy sets and optimally uses the hidden information of data. First, the

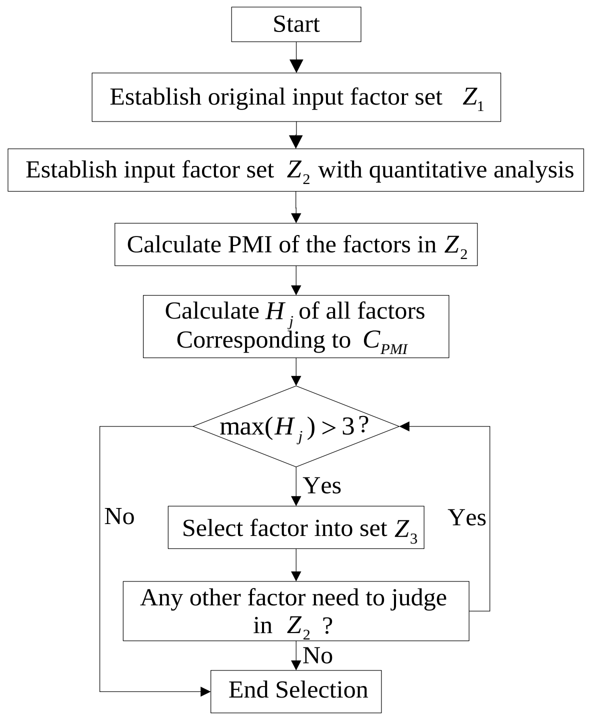

Start
Establish original input factor set Z_1
Establish input factor set Z_2 with quantitative analysis
Calculate PMI of the factors in Z_2
Calculate H_j of all factors Corresponding to C_{PMI}
$\max(H_j) > 3$?
Yes
No
Select factor into set Z_3
Yes
Any other factor need to judge in Z_2 ?
No
End Selection

Fig. 3 Calculation flowchart of C_{PMI}

original information is processed. We assume that the discrete domain $\omega \underline{\underline{\Delta}} \{\omega_1, \omega_2, \ldots, \omega_n\}$ and $\varepsilon \underline{\underline{\Delta}} \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$, respectively.

ω_i and ε_j (i=1,2,…,n and j=1,2,…,m) are the discrete control points of ω and ε, respectively.

The selection of control points should meet the following three principles:

(1) Minimize loss of information;

(2) The partition of the domain can cover all sample points;

(3) As far as possible, the division of the domain is equally divided. If the sample points of a certain area are dense, it can be widened appropriately.

The original information of the variables spreads to all control points in the domain according to certain rules. Every point in the domain will obtain the diffusion information from the original information. The fuzzy relation matrix is established using discrete regression algorithm of 2D normal information diffusion (Huang *et al.* 2012). We assume that

$$\begin{cases} a_1 = \min\limits_{1 \le j \le n} \{\varepsilon_j\} \text{ and } b_1 = \max\limits_{1 \le j \le n} \{\varepsilon_j\} \\ a_2 = \min\limits_{1 \le j \le n} \{\omega_j\} \text{ and } b_2 = \max\limits_{1 \le j \le n} \{\omega_j\} \end{cases} \tag{15}$$

The 2D normal information diffusion discrete regression equation can be written as follows

$$f_p(\varepsilon, \omega) = \frac{1}{2\pi p h_1 h_2} \sum_{j=1}^{n} \exp\left[-\frac{(\varepsilon - \varepsilon_j)^2}{2h_1^2} - \frac{(\omega - \omega_j)^2}{2h_2^2} \right] \tag{16}$$

Where $h_1 = 1.4208(b_1 - a_1)/(p-1)$,

$h_2 = 1.4208(b_2 - a_2)/(p-1)$, and $n \ge 10$.

After normalizing h_1 and h_2, Eq. (16) can be simplified as follows

$$\begin{aligned} f_p(\varepsilon, \omega) &= \frac{1}{2\pi p h^2 \varphi} \sum_{j=1}^{n} \exp\left[-\frac{(\varepsilon' - \varepsilon'_j)^2 + (\omega' - \omega'_j)^2}{2h^2} \right]. \\ &= \tilde{f}_p(\varepsilon', \omega') \end{aligned} \tag{17}$$

The parameters in Eq. (17) should satisfy the following expression

$$\begin{cases} \varepsilon' = (\varepsilon - a_1)/(b_1 - a_1), \omega' = (\omega - a_2)/(b_2 - a_2) \\ \varepsilon', \omega' \in [0,1] \\ \varepsilon'_j = (\varepsilon_j - a_1)/(b_1 - a_1), \omega_j' = (\omega_j - a_2)/(b_2 - a_2) \\ \varepsilon'_j, \omega_j' \in [0,1] \\ \varphi = (b_1 - a_1)/(b_2 - a_2) \\ h = 1.4208/(p-1) \end{cases} \tag{18}$$

We will obtain the $n \times m$ -order matrix $R' = \{r'_{ij}\} = \{\tilde{f}_p(\varepsilon'_i, \omega'_j)\}$ based on the sample data by Eq. (17). After normalizing R', the fuzzy relation matrix $R^{(1)}$ can be obtained, which represents the original information membership degree.

4. Fuzzy reasoning

After fuzzy diffusion of the sample information, fuzzy reasoning is performed. The process of fuzzy reasoning is divided into two stages: the first-order fuzzy reasoning calculates the membership degree of the original data and obtains the weight of each control point in the domain of effect quantity; the second-order fuzzy reasoning calculates the weight of a single influence factor and derives the final weight matrix.

4.1 First-order reasoning

After calculating the membership degree matrix $R^{(1)}$, the weight of each control point in the domain is calculated. If we assume that A and B are fuzzy membership degree sets of domain ω and ε, respectively (Zimmermann *et al.* 2012), then

$$B = A \cdot R \tag{19}$$

Where R denotes the fuzzy relation matrix and "·" represents the rule of operation that often uses simple matrix multiplication or max–min algorithm.

Matrix multiplication or Max-Min operation is suitable for cases where the diagnostic requirements for impact and evaluation quantities are not high. Lattice closeness degree can be used to express the closeness of fuzzy sets. The first-order fuzzy reasoning is mainly used to calculate the membership degree of the original data, and to get the influence weight of each factor in the field of effect quantity. Considering that the monitoring data of dam body vary regularly (normal) with the influence factors, the method of lattice closeness reasoning is used to calculate membership degree. Given that the monitoring data of the

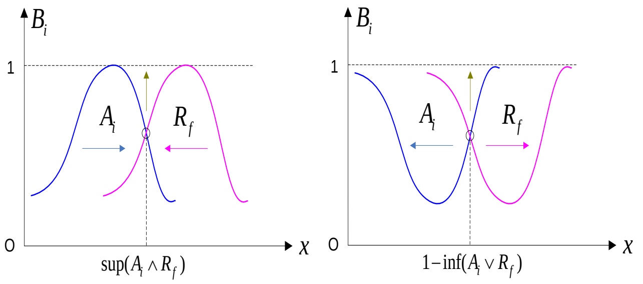


Fig. 4 Fuzzy membership degree

dam have a certain regularity (normal) change with the influence factors, the membership degree *B* can be calculated by the lattice closeness degree (Zimmermann *et al.* 2011), as follows

$$B_i^j = 0.5\left\{\underset{x\in X}{\vee}(A_i^j \wedge R_f) + \left[1 - \underset{x\in X}{\wedge}(A_i^j \vee R_f)\right]\right\} \tag{20}$$

Where *i* denotes the number of bits in matrix *B*, *j* denotes the order of fuzzy reasoning, *f* is the column heading of matrix *R*, $\wedge$ denotes the small product operator, and $\vee$ represents the large product operator. A smaller exterior product means a smaller distance between the fuzzy sets, as shown in Fig. 4.

The fuzzy relation matrix $R^{(1)}$ is formed by original information distribution and normalization. $A^{(1)}$ can be calculated by the following equation (Wei *et al.* 2011):

$$\begin{cases} A_i^{(1)} = \max\left(0, 1 - \dfrac{|a - a_i|}{\Delta}\right), & a_{\min} < a < a_{\max} \\ A_i^{(1)} = [1, 0, \mathrm{L}, 0], & a \le a_{\min} \ \mathit{and}\ a_{\min} \in A_i \\ A_i^{(1)} = [0, 0, \mathrm{L}, 1], & a \ge a_{\max} \ \mathit{and}\ a_{\max} \in A_i \end{cases} \tag{21}$$

Where $\Delta = a_{i+1} - a_i$, *i* denotes the *i*-th control point of the domain. After obtaining $A^{(1)}$ and $R^{(1)}$, the first-order reasoning $B^{(1)} = A^{(1)} \cdot R^{(1)}$ can be conducted and the membership degree $B^{(1)}$ of each control point in ω can be obtained.

4.2 Second-order reasoning

The weight of each influencing factor is calculated by the second-level fuzzy reasoning. Because the single factor reasoning is not required high-accuracy for the sample set, matrix multiplication is used to calculate the weight of the sample set.

Given that effect quantity is affected by many factors, the weight of each factor should be calculated and the second-order reasoning $B^{(2)} = A^{(2)} \cdot R^{(2)}$ should be performed to obtain the comprehensive membership degree $B^{(2)}$ of the measured values. The weight of a single influence factor can be calculated by a variety of algorithms, such as analytic hierarchy process, principal component analysis, and gray correlation degree. The sample sequence of dam monitoring in storage period is short and has gray uncertainty; thus, gray correlation degree algorithm can be used to calculate $A^{(2)}$ (Wei *et al.* 2011). Gray correlation degree κ can be calculated as follows

$$\kappa_{ij}(x_i, x_j) = \frac{1}{p}\sum_{k=1}^{n} \frac{\min_i \min_j \Delta_{ij}(k) + \rho \max_i \max_j \Delta_{ij}(k)}{\Delta_{ij}(k) + \rho \max_i \max_j \Delta_{ij}(k)} \tag{22}$$

Where $\Delta_{ij} = |\varepsilon_i - \omega_j|$, $\min_i \min_j \Delta_{ij}$ denotes the absolute difference when *k* takes the minimum value.

ρ should take a small value when the factor observation sequence has an abnormal value. ρ subjectively reflects the attention degree that the researchers focused on $\max_i \max_j |\varepsilon_i - \omega_j|$ and objectively reflects the impact degree the factors are applied to κ. To select factors more accurately, we use the mean absolute value of the difference between the control points to determine ρ (Li *et al.* 1998; Lotfi *et al.* 2012), as follows

$$\begin{cases} \dfrac{\Delta_\omega}{\Delta_{\max}} \le \rho \le 1.5\dfrac{\Delta_\omega}{\Delta_{\max}} & \mathit{if} \quad \Delta_{\max} > 3\Delta_\omega \\ 1.5\dfrac{\Delta_\omega}{\Delta_{\max}} \le \rho \le 2\dfrac{\Delta_\omega}{\Delta_{\max}} & \mathit{if} \quad \Delta_{\max} \le 3\Delta_\omega \\ \Delta_\omega = \dfrac{1}{n \cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}|\omega_i - \varepsilon_j| \end{cases} \tag{23}$$

We first normalize the domain of influence factor and effect quantity when calculating the weight matrix $A^{(2)}$ of the single factor. Afterward, ρ is obtained by Eq. (23) according to Δ_ω. After calculating ρ, we obtained κ by Eq. (22). Given the different dimensions between influence factor and effect quantity, performing the non-dimensional treatment to the correlation coefficient matrix is necessary.

4.3 Information integration

We use the maximum value of B_i instead of the fuzzy quantity of the membership information with the aid of the maximum likelihood criterion after obtaining the membership information of B_i. Non-fuzzy data are converted into fuzzy information after diffusion. To eliminate the influence of

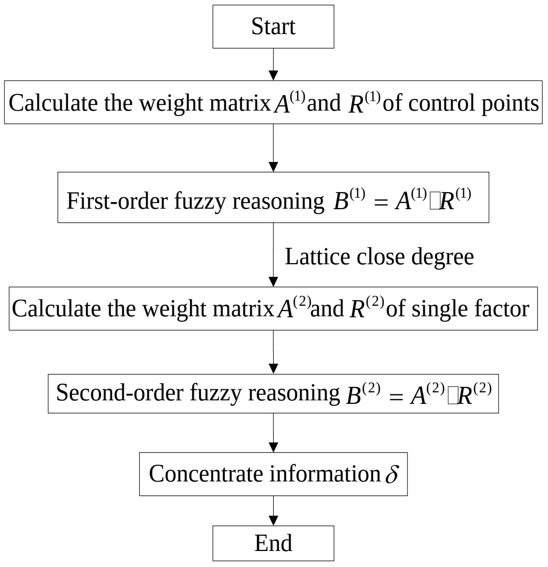

Start
Calculate the weight matrix $A^{(1)}$and $R^{(1)}$of control points
First-order fuzzy reasoning $B^{(1)} = A^{(1)}□R^{(1)}$
Lattice close degree
Calculate the weight matrix $A^{(2)}$and $R^{(2)}$of single factor
Second-order fuzzy reasoning $B^{(2)} = A^{(2)}□R^{(2)}$
Concentrate information δ
End

Fig. 5 Calculation process of effect quantity

information diffusion and obtain the best predictive value, centralizing the information is necessary; this procedure can be expressed as follows (Wei *et al.* 2011, Papaleontiou *et al.* 2012)

$$\delta = \sum_{i=1}^{n} B^{k}\left(\delta_i\right)\omega_i \Big/ \sum_{i=1}^{n} B_i^{k} \qquad (24)$$

Where ω_i denotes the value of ω at the i-th point, and k is the constant that generally takes the value of 2.

Therefore, combined with fuzzy reasoning and gray correlation degree, compound uncertain information forecasting method is established.

5. Comprehensive effective degree

A higher mean value of the prediction accuracy means a higher accuracy of the prediction approach. By contrast, a higher discrete degree of prediction accuracy distribution means a lower accuracy of the prediction approach. The iterative calculation is used in prediction based on information diffusion. Thus, the fitting result significantly influences the prediction. Therefore, we use the comprehensive availability index to verify the accuracy of the prediction results at each point. The comprehensive availability index is expressed as follows

$$m = \zeta m_1 + (1-\zeta) m_2 \qquad (25)$$

Where m denotes the comprehensive availability degree, with the values in the range of [0,1]; m_1 denotes the fitting accuracy; m_2 denotes the prediction accuracy; and α is a parameter that is generally 0.5.

We assume that A_i is a random variable. The availability degree function is the integral of the forecasting precision in the interval. Setting $[0,T_0]$ as the sample interval and $[T_0+1,T_0+L]$ as the forecasting interval. m_1 and m_2 can be calculated as follows

$$m_1 = \left\{1-\left[\int_0^{T_0} Q_1(\mathrm{t})A_t^2\,dt-\left(\int_0^{T_0} Q_1(\mathrm{t})A_t\,dt\right)^2\right]^{\frac{1}{2}}\right\} \times \int_0^{T_0} Q_1(\mathrm{t})A_t\,dt \qquad (26)$$

$$m_2 = \left\{1-\left[\int_{T_0+1}^{T_0+T} Q_2(\mathrm{t})A_t^2\,dt-\left(\int_{T_0+1}^{T_0+T} Q_2(\mathrm{t})A_t\,dt\right)^2\right]^{\frac{1}{2}}\right\} \times \int_{T_0+1}^{T_0+T} Q_2(\mathrm{t})A_t\,dt \qquad (27)$$

A_i can be calculated by using residual e_i and measured value ω_i, with the following formulas

$$A_t = 1-\left|\frac{e_i}{v_i}\right| \quad and \quad 0 \le \left|\frac{e_i}{v_i}\right| \le 1 \qquad (28)$$

Density functions $Q_1(t)$ and $Q_2(t)$ can be obtained according to the actual situation.

Assuming that β is the weight coefficient of $Q_1(t)$ and $Q_1(t)$, it satisfies

$$\begin{cases} Q(t) = \begin{cases} \beta Q_1(t) & t \in \left(0, T_0\right] \\ (1-\beta)Q_2(t) & t \in \left[T_0, T_0+T\right] \end{cases} \\ Q(t) \ge 0 \quad and \quad \int_0^{T_0+T} Q(t)dt = 1 \end{cases} \qquad (29)$$

6. Case study

6.1 Project profile

To verify the compound uncertain information forecasting method for the monitoring of high dam, we take a high earth-rock dam located in Yunnan, China, for analysis. This dam is clay core wall rockfill dam, with the maximum dam height, length of dam crest, width of dam crest and normal water level being 261.5 m, 627.87 m, 18 m and 812 m, respectively.

The reservoir adopts the method of storing water by stages, and the measurement of the seepage gauge data of the dam is affected by many uncertain factors. Because of the small number of samples, the results of the conventional prediction models and methods are not satisfactory. The algorithm of combining Copulas function and information diffusion theory presented in this paper can not only solve this problem, but also verify the validity of the method.

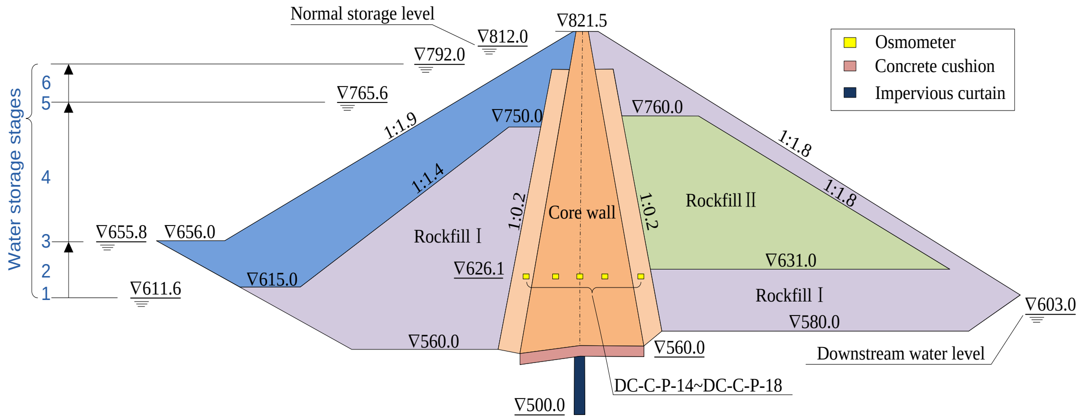


Fig. 6 Cross-section of the dam

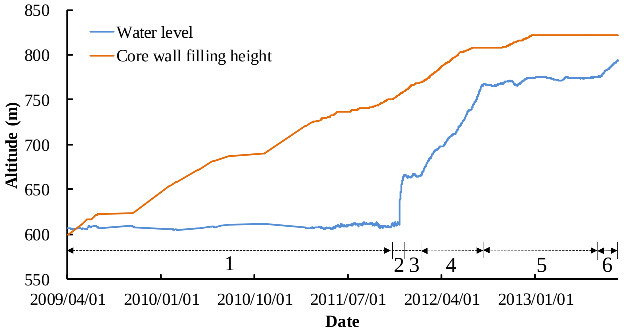


Fig. 7 The filling process of the dam and the division of the water storage stage

Table 1 Correlation calculation of Z_1

Component	Symbol	Formula	r	τ	Select?
1	$\sin\eta$	$\eta = 2\pi it/365$	-0.2368	-0.2271	-
2	$\cos\eta$		0.1732	0.2468	-
3	$\sin^2\eta$		-0.1944	-0.2483	-
4	$\cos^2\eta$		0.1805	0.2216	-
5	H_{sh}	$H_{sh}^i = (H_{sh})^i$	0.8048	0.7531	√
6	H_{sh}^2		0.7876	0.7531	√
7	H_{xh}		0.1780	0.0897	-
8	P		0.0635	0.0438	-
9	ς	$\varsigma = \dfrac{d_i - d_0}{100}$	-0.4463	-0.2216	-
10	ς^2		-0.3704	-0.2313	-
11	$\ln\varsigma$		0.5175	0.4365	√
12	t		0.9873	0.9670	√
13	H_{ch}		0.9913	0.9699	√
14	H_{sd}		0.9707	0.9692	√

In the actual monitoring, the settlement of the core wall of the dam is the most serious. In addition, in the period of water storage, the external factors and variables of the seepage gauge in the core wall are the largest. The Copula combined with PMI algorithm is the most suitable for multivariable correlation calculation, and it can more accurately filter the secondary alternative factor set Z_2.

The water level in the storage period is a stage uplift, from 611.6 m to 792.0 m. The core wall pressure gauge with 626.1 m height meets the monitoring requirements of six storage periods, so the osmometers with 626.1 m elevation are used to calculate the pressure.

The structure of the dam and the division of the water storage stage are shown in Figs. 6-7.

6.2 Calculation of stages 1-2

6.2.1 Input factor selection

We use the calculation process of point DB-C-P-15 as an example for analysis. On the basis of factor set Z_1 determined by quantitative analysis, the correlation of inputting factors can be analyzed by Spearman coefficient r and Kendall rank correlation coefficient τ. Inputting factor set Z_2 can be determined through the significant testing of the factors in Z_1, as shown in Table 1.

Table 2 Domain division of the factors

Factor	Domain	Segment number	Control points	Dimension
H_ω	ω	6	[627,664,701,738,775,812,849]	m
H_{sh}	ε	6	[605.0,506.2,607.4,608.6, 609.8 611 612 2]	m
H_{ch}	ε	6	[633,654,675,696,717,738,759]	m
H_{sd}	ε	6	[670,678,686,694,702,710,718]	m
t	ε	3	[0,1,2,3]	year

Table 3 Result of fuzzy relation matrix

	$H_{\omega1}$ (627)	$H_{\omega2}$ (664)	$H_{\omega3}$ (701)	$H_{\omega4}$ (738)	$H_{\omega5}$ (775)	$H_{\omega6}$ (812)	$H_{\omega7}$ (849)
H_{sh1} (605.0)	0.036	0.322	1.000	0.206	0.026	0.003	0.000
H_{sh2} (603.2)	0.032	0.303	1.000	0.222	0.023	0.003	0.000
H_{sh3} (607.4)	0.017	0.011	0.519	1.000	0.015	0.003	0.000
H_{sh4} (608.6)	0.000	0.002	0.226	1.000	0.013	0.002	0.000
H_{sh5} (609.8)	0.000	0.001	0.117	1.000	0.019	0.011	0.007
H_{sh6} (611.0)	0.000	0.001	0.002	0.586	1.000	0.971	0.043
H_{sh7} (612.2)	0.000	0.001	0.002	0.035	0.523	1.000	0.150

After the significant testing, we obtain the factor set $Z_2 = \{H_{sh}, H_{sh}^2, t, H_{ch}, H_{sd}, \ln\varsigma\}$. The final input factor set Z_3 can be obtained by using the Copula entropy algorithm, which calculates the C_{PMI} value and Hampel distance of the factors in Z_2. After several Hampel tests, Z_3 is obtained as $\{H_{sh},H_{ch},H_{sd},t\}$.

First, domains ε and ω of the DB-C-P-15 measuring point are segmented, as shown in Table 2.

6.2.2 Fuzzy reasoning

The first-order reasoning is performed on (H_{sh},H_ω), (H_{ch},H_ω), (H_{sd},H_ω), and (t,H_ω). According to Eq. (17), fuzzy relation matrix $R^{(1)}$ is obtained after the normalization, as shown in Table 3.

$B^{(1)}$ can be obtained through the first-order reasoning $B^{(1)}$=$A^{(1)}$·$R^{(1)}$ with the help of the lattice closeness degree after obtaining $A^{(1)}$. We obtain the distinguishing coefficients according to the coefficient interval calculated by Eq. (23).

The weight matrix can be obtained using the gray correlation degree. After normalizing the weight matrix for each factor, the weight matrix of the single factor is obtained as $A^{(2)}$=[0.21,0.23,0.25,0.31].

Finally, the fuzzy relation matrix of second-order reasoning is obtained.

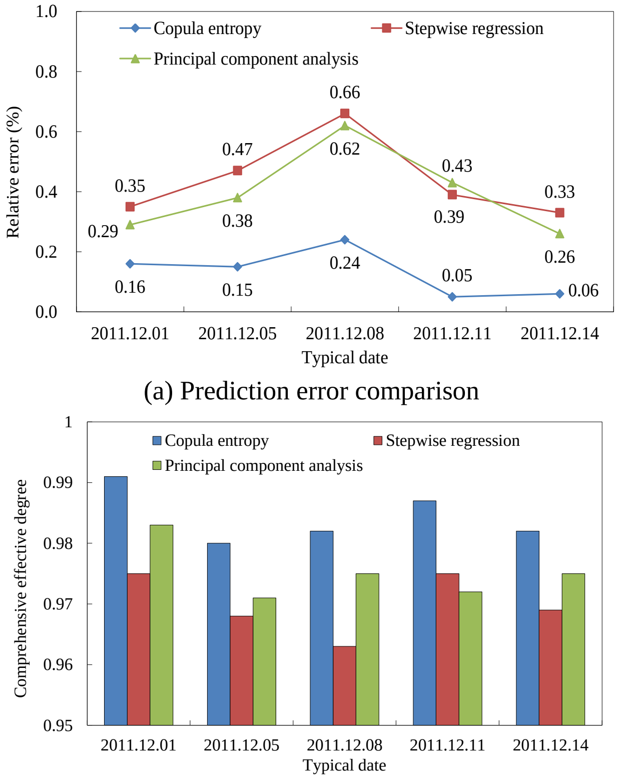
Matrix multiplication is adopted as the operation rule "·" without considering the correlation between the sample information and the influence factors. The result of reasoning is expressed as follows:

$$B^{(2)} = [0, 0.012, 0.039, 0.435, 0.527, 0.878, 0.106]$$

After the information concentration using Eq. (24), we

Table 4 Comparison of the three factor selection methods

Method	Z_3
Copula entropy + PMI	$H_{sh}, H_{ch}, H_{sd}, t$
Stepwise regression	$\sin\eta, \cos\eta, H_{sh}, H_{sh}^2, H_{ch}, H_{sd}, \varsigma, t$
Principal component analysis	$\sin\eta, \cos\eta, H_{sh}, H_{ch}, H_{sd}, \ln\varsigma, t$



(a) Prediction error comparison

(b) Comprehensive effectiveness comparison

Fig. 8 Comparison of different factor selection methods

obtain the forecasting value of δ. Finally, iterative calculation is performed to obtain the other prediction values.

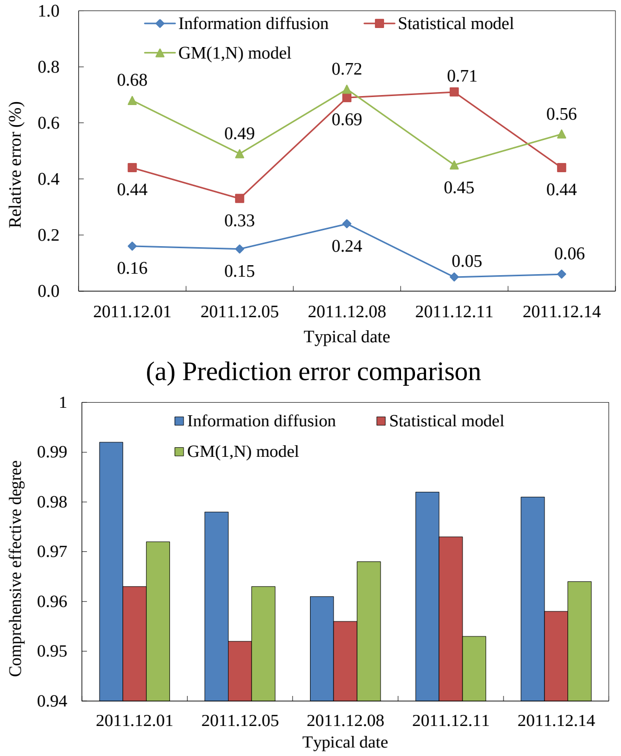
6.2.3 Comprehensive effective degree

The comprehensive effective degree is calculated to verify the accuracy of the prediction results. A total of 181 groups of fitting data are used in stage 1, 15 groups of forecasting data are used in stage 2. The fitting interval of the sample is [0,181], and the prediction interval is [182,196]. $Q(t)$ is calculated according to the actual distribution $1/n$.

Finally, the comprehensive effective degree can be obtained as m=0.992.

6.2.4 Model comparison and analysis

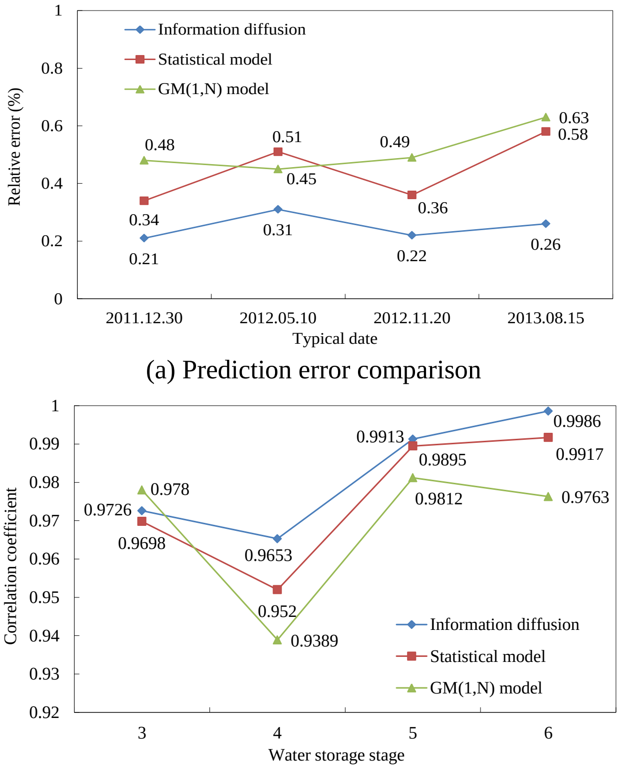
The stepwise regression analysis and principal component analysis take into account the structural factors of the dam body, which are different from the simple correlation analysis, and are comparable to the methods proposed in this paper. The stepwise regression method excludes the factors according to the significance of the variables. The magnitude of significant index F determines the number of culling factors. Principal component analysis (PCA) selects principal components according to the



(a) Prediction error comparison

(b) Comprehensive effectiveness comparison

Fig. 9 Comparison of the same input factors and different calculation models



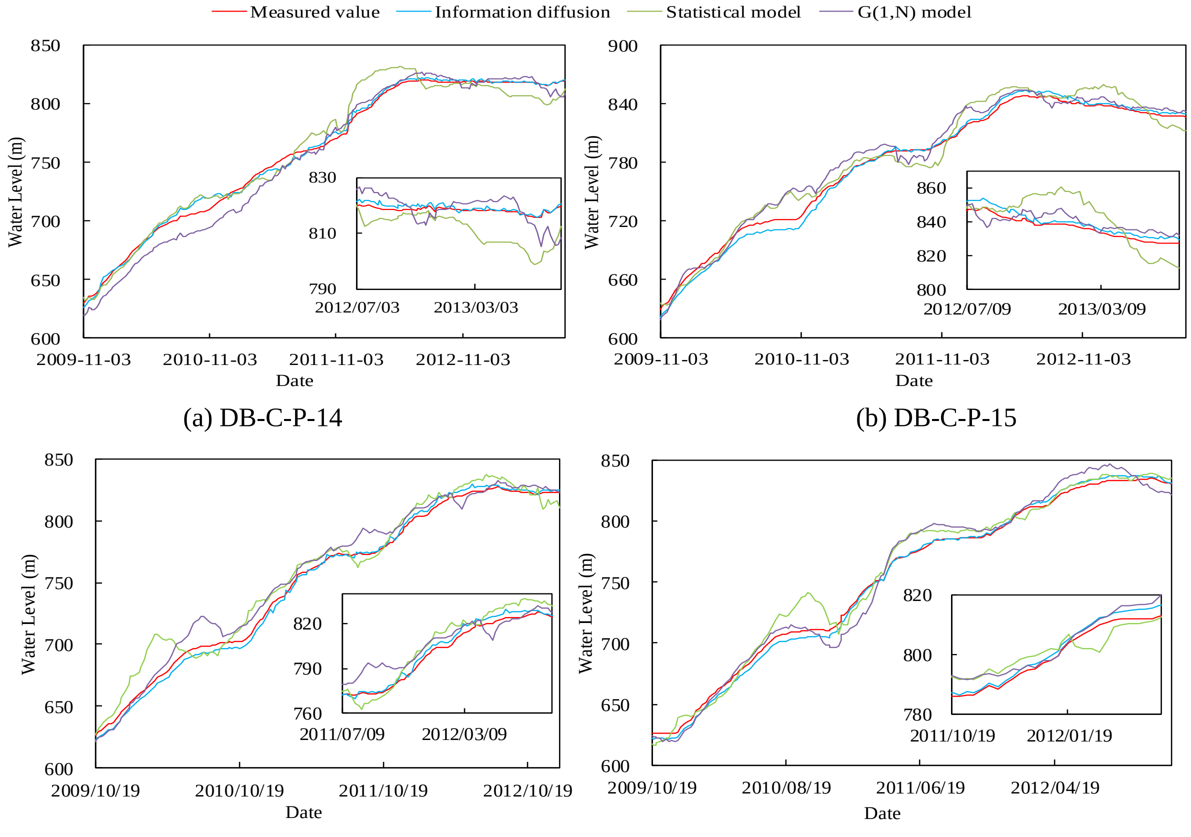
(a) Prediction error comparison

(b) Comparison of the correlation coefficients

Fig. 10 Comparison of different models in stages 3-6

Table 5 Effective degree comparison of stages 3-6 of water storage

Fitting stage	Prediction stage	Information diffusion			GM (1, *N*) model			Statistical model		
		m_1	m_2	m	m_1	m_2	m	m_1	m_2	m
1~2	3	0.962	0.951	0.957	0.951	0.953	0.952	0.969	0.951	0.960
1~3	4	0.984	0.979	0.982	0.988	0.962	0.975	0.972	0.969	0.971
1~4	5	0.986	0.975	0.981	0.982	0.971	0.977	0.976	0.962	0.969
1~5	6	0.991	0.982	0.987	0.986	0.981	0.984	0.981	0.971	0.976



(a) DB-C-P-14

(b) DB-C-P-15

(c) DB-C-P-16

(d) DB-C-P-17

Fig. 11 Fitting process lines

contribution rate of variables. They are similar to the Copula entropy proposed in this paper based on the PMI value. Therefore, the comparison between the three methods is more practical.

Table 4 shows the two factor selection methods, which are used for comparison with the Copula entropy algorithm to verify the prediction effect of stage 2. Fig. 8(a) illustrates the comparison of the three methods, and we can find that copula entropy algorithm has less relative error than stepwise regression and principal component analysis. Fig. 8(b) shows the comprehensive effective degree of Copula entropy algorithm is higher than the other two methods.

In Fig. 9, two different models with the same input factors are used to compare with the information diffusion theory. Fig. 9(a) shows the relative error of information diffusion method is the smallest among the three models. Fig. 9(b) shows the comprehensive effective degree of different models in typical dates. As shown in Fig. 9(b), we can clearly see the information diffusion method has the overall highest effective degree.

6.3 Calculation of stages 3-6

The prediction of all stages is needed to verify the forecasting accuracy of the compound uncertain information forecasting method proposed in this study. Table 5 shows the effective degree comparison of the three methods in stages 3-6. Figs. 10(a)-10(b) illustrate the relative error and correlation coefficients of the three models in typical dates. As shown in Table 5, Figs. 10(a)-10(b), information diffusion method has more obvious advantages than statistical model and grey model.

6.4 Verification of other measuring points

The measuring samples of DB-C-P-14, DB-C-P-16, and DB-C-P-17 are calculated to further verify the accuracy of the method based on Copula entropy and information diffusion theory. Fig. 11 shows the fitting process lines of these points, and we can see the fitting effect of the approach proposed in this study is better than other models.

7. Conclusions

This paper mainly improved and integrated Copula entropy and information diffusion theory, and established the composite uncertain information prediction method integrating factor selection and small sample analysis. The main content of this paper was as follows:

- Monitoring factors of high dam are chosen by improved Copula entropy method proposed in this paper. The Copula function uses the two-dimensional Archimedean function and replaces normal distribution by the Cauchy distribution. Later, the Copula entropy is deduced. The final input factors are determined by combining the partial mutual information theory and Hample test.
- For small sample analysis, sample information is expanded firstly by information diffusion method, followed by fuzzy reasoning. During the fuzzy reasoning, lattice closeness degree method is used instead of the Max-Min method and the grey correlation theory is introduced in. Samples are predicted by combining the fuzzy reasoning and grey correlation.
- To verify the proposed method, Copula entropy is compared with stepwise regression method and PCA which are often used in factor selection field, while the information diffusion method is compared with the statistical model and grey model which are commonly used in high dam monitoring. Results confirm the superiority of the proposed method under various uncertain factors.
- To meet both fitting accuracy and prediction accuracy of samples, results are evaluated by comprehensive validity. The comparative analysis reveals that the comprehensive validity of Copula entropy and information diffusion method is in accordance with relative error, which further proves validity of the proposed method.

Acknowledgments

The research described in this paper was financially supported by National Natural Science Foundation of China (SN: 51579085, 51579083, 41323001, 51279052, 51379068, 51479054, 51139001, 51579086), the National Key Research and Development Program of China (SN: 2016YFC0401601), A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Province “333 High-Level Personnel Training Project(Grant No. 2016-B1307101), Open Foundation of State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (SN: 20165042112, 20145027612), the Fundamental Research Funds for the Central Universities (Grant No. 2015B20714, 2016B04114, 2015B25414, 2014B37114, 2014B37414, 2015B25414, 2014B1605336), Jiangsu Natural Science Foundation (Grant No. BK20140039), the Doctoral Program of Higher Education of China (SN: 20130094110010), Jiangsu Basic Research Program (Grant No. BK20160872). Jiang Su’s “Six Talents Summit” Project (JY-008, JY003).

References

Akpinar, U., Binici, B. and Arici, Y. (2014), “Earthquake stresses and effective damping in concrete gravity dams”, *Earthq. Struct.*, **6**(3), 251-266.

Allahdadi, M.N. and Li, C. (2017), “Effect of stratification on current hydrodynamics over Louisiana shelf during Hurricane Katrina”, *Water Sci. Eng.*, **10**(2), 154-165.

Awan, J.A. and Bae, D. (2013), “Application of adaptive neuro-fuzzy inference system for dam inflow prediction using long-range weather forecast”, *Digital Information Management (ICDIM), 2013 Eighth International Conference on. IEEE*, 247-251.

Chen, L., Singh, V.P., Guo, S., Zhou, J. and Ye, L. (2014), “Copula entropy coupled with artificial neural network for rainfall-runoff simulation”, *Stoch. Environ. Res. Risk Assess.*, **28**(7), 1755-1767.

Cordeiro, G.M. and Lemonte, A.J. (2011), “The beta-half-Cauchy distribution”, *J. Prob. Statist.*, **2011**, Article ID 904705, 18.

Croux, C. and Dehon, C. (2010), “Influence functions of the Spearman and Kendall correlation measures”, *Statist. Meth. Appl.*, **19**(4), 497-515.

da Silva, V.D.P.R., Belo Filho, A.F., Almeida, R.S.R., de Holanda, R.M. and da Cunha Campos, J.H.B. (2016), “Shannon information entropy for assessing space-time variability of rainfall and streamflow in semiarid region”, *Sci. Total Environ.*, **544**, 330-338.

Deka, P.C. (2014), “Support vector machine applications in the field of hydrology: a review”, *Appl. Soft Comput.*, **19**, 372-386.

Foster, M., Fell, R. and Spannagle, M. (2011), “The statistics of embankment dam failures and accidents”, *Can. Geotech. J.*, **37**(5), 1000-1024.

Hassanvand Jamadi, M. and Alighardashi, A. (2017), “Application of Froude dynamic similitude in anaerobic baffled reactors to prediction of hydrodynamic characteristics of a prototype reactor using a model reactor”, *Water Sci. Eng.*, **10**(1), 53-58.

Hipni, A., El-shafie, A., Najah, A., Karim, O.A., Hussain, A. and Mukhlisin, M. (2013), “Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS)”, *Water Res. Manage.*, **27**(10), 3803-3823.

Hofert, M., Mächler, M. and Mcneil, A.J. (2012), “Likelihood inference for Archimedean copulas in high dimensions under known margins”, *J. Multiv. Anal.*, **110**(5), 133-150.

Hu, D.X., Zhou, Z.Q., Li, Y. and Wu, X.L. (2011), “Dam safety analysis based on stepwise regression model”, *Appl. Mech. Mater.*, **204**, 2158-2161.

Huang, C. and Shi, Y. (2012), *Towards Efficient Fuzzy Information Processing: Using the Principle of Information Diffusion*, Vol. 99, Physica.

Li, W. (1998), “Applied example of availableness calculating and selecting in the forecast method”, *Math. Pract. Theory*, **28**(4), 366-370.

Li, X.Q., Zheng, D.J. and Ju, Y.P. (2016), “Input factor optimization study of dam seepage statistical model based on copula entropy theory”, *J. Hohai Univ. (Nat. Sci.)*, **44**(3), 370-376.

Li, X.Q., Zheng, D.J., Cao, L.P. *et al*. (2016), “Application of information diffusion approach to effect quantity prediction in dam monitoring”, *J. Hohai Univ. (Nat. Sci.)*, **44**(6), 536-343.

Liu, H.C., Liu, L., Bian, Q.H., Lin, Q.L., Dong, N. and Xu, P.C. (2011), “Failure mode and effects analysis using fuzzy evidential reasoning approach and grey theory”, *Exp. Syst. Appl.*, **38**(4), 4403-4415.

Liu, J. and Li, D. (2011), “Research on loss of life of dam failure based on rough set theory”, *Electric Technology and Civil Engineering (ICETCE*), 2011 International Conference on. IEEE, 601-604.

Liu, S. and Forrest, J.Y.L. (2010), *Grey Systems: Theory and Applications*, Springer.

Lotfi, V. and Samii, A. (2012), “Dynamic analysis of concrete gravity dam-reservoir systems by wavenumber approach in the frequency domain”, *Earthq. Struct.*, **3**(3-4), 533-548.

Maier, H.R., Jain, A., Dandy, G.C. and Sudheer, K.P. (2010), “Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions”, *Environ. Model. Softw.*, **25**(8), 891-909.

Novak, P., Moffat, A.I.B., Nalluri, C. and Narayanan, R. (2007), *Hydraulic Structures*, CRC Press.

Opyrchał, L. (2003), “Application of fuzzy sets method to identify seepage path through dams”, *J. Hydra. Eng.*, **129**(7), 546-548.

Papaleontiou, C.G. and Tassoulas, J.L. (2012), “Evaluation of dam strength by finite element analysis”, *Earthq. Struct.*, **3**(3-4), 457-471.

Pekasiewicz, D. (2014), “Application of quantile methods to estimation of cauchy distribution parameters”, *Statist. Tran. New Ser.*, **15**(1), 133-144.

Shen, W.W. and Ren, J.M. (2014), “Multiple stepwise regression analysis crack open degree data in gravity dam”, *Appl. Mech. Mater.*, **477-478**, 888-891.

Su, H., Wen, Z. and Wu, Z. (2011), “Study on an intelligent inference engine in early-warning system of dam health”, *Water Res. Manage.*, **25**(6), 1545-1563.

Tao, J., Zhang, B., Li, S. and Chao, W. (2011), “Principal component analysis of collinearity of dam safety monitoring data”, *Water Res. Power*, **39**(2), 50-52.

Valipour, M., Banihabib, M.E. and Behbahani, S.M.R. (2013), “Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir”, *J. Hydrol.*, **476**, 433-441.

Wei, G.W. (2011), “Gray relational analysis method for intuitionistic fuzzy multiple attribute decision making”, *Exp. Syst. Appl.*, **38**(9), 11671-11677.

Wen, H.Y., Zhou, L., Chen, G.Y., Hu, J.Y. and He, M.L. (2013), “Research on dam deformation analysis model considering multi-factors”, *Appl. Mech. Mater.*, **351**, 1318-1324.

Yang, M., Su, H. and Wen, Z. (2017), “An approach of evaluation and mechanism study on the high and steep rock slope in water conservancy project”, *Comput. Concrete*, **19**(5), 527-535.

Yitao, M., Houlei, X., Feng, W., Bangbin, W. and Lulin, W. (2011), “Time-varying prediction model of dam monitoring data based on principal component analysis”, *Water Power*, **37**(10), 100-103.

Zheng, D., Cheng, L. and Xu, Y. (2013), “Evaluate the impact of cold wave on face slab cracking using fuzzy finite element method”. *Math. Prob. Eng.*, **82026**, 1-11.

Zhu, H.H., Yin, J.H., Dong, J.H. and Zhang, L. (2010), “Physical modelling of sliding failure of concrete gravity dam under overloading condition”, *Geomech. Eng.*, **2**(2), 89-106.

Zimmermann, H.J. (2011), *Fuzzy Set Theory-and its Applications*, Springer Science & Business Media.

Zimmermann, H.J. (2012), *Fuzzy Sets, Decision Making, and Expert Systems*, Vol. 10, Springer Science & Business Media.

CC