Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Information fusion diagnosis and early-warning method for monitoring the long-term service safety of high dams*

Xing LIU[1,2,3], Zhong-ru WU[†‡1,2,3], Yang YANG[1,2,3], Jiang HU[1,2,3], Bo XU[1,2,3]

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

([2]*National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Hohai University, Nanjing 210098, China*)

([3]*College of Water-Conservancy and Hydropower, Hohai University, Nanjing 210098, China*)

[†]E-mail: zrwu@hhu.edu.cn

Received May 7, 2012; Revision accepted Aug. 13, 2012; Crosschecked Aug. 17, 2012

Abstract: Analyzing the service behavior of high dams and establishing early-warning systems for them have become increasingly important in ensuring their long-term service. Current analysis methods used to obtain safety monitoring data are suited only to single survey point data. Unreliable or even paradoxical results are inevitably obtained when processing large amounts of monitoring data, thereby causing difficulty in acquiring precise conclusions. Therefore, we have developed a new method based on multi-source information fusion for conducting a comprehensive analysis of prototype monitoring data of high dams. In addition, we propose the use of decision information entropy analysis for building a diagnosis and early-warning system for the long-term service of high dams. Data metrics reduction is achieved using information fusion at the data level. A Bayesian information fusion is then conducted at the decision level to obtain a comprehensive diagnosis. Early-warning outcomes can be released after sorting analysis results from multi-positions in the dam according to importance. A case study indicates that the new method can effectively handle large amounts of monitoring data from numerous survey points. It can likewise obtain precise real-time results and export comprehensive early-warning outcomes from multi-positions of high dams.

Key words: Dam monitoring, Diagnosis, Early-warning, Multi-source information fusion, Information entropy
doi:10.1631/jzus.A1200122 **Document code:** A **CLC number:** TV3

1 Introduction

With the increasing number of dangerous and aged hydraulic projects in China, the long-term service safety of high dams on great rivers has attracted attention from all social sectors (Wu and Su, 2005; Wu and Gu, 2006). High dams and large reservoirs are often built on great rivers with millions of people and several important infrastructures downstream. Therefore, the failure of high dams has serious consequences for the lives, properties, and economic situation of residents downstream.

An integrated safety monitoring system is usually built according to standards, given the crucial importance of ensuring the safety of high dams and large reservoirs. The safety of high dams is ensured by evaluating their behaviors through prototype monitoring data analysis (Wu *et al*., 2007; Mata, 2011). With the development of technology, different advanced safety monitoring techniques (such as optical fiber monitoring and wireless sensing monitoring) and automatic monitoring systems have gradually been applied in monitoring the safety of high dams. Dense point groups (or sets) need to be laid at key positions to ensure the safety of high dams and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 51139001, 51179066, 51079046, and 50909041)

form a large-scale online monitoring data collection system. Due to the high frequency of intensive observation, the capacity of monitoring data increases as a geometric series. Current analysis methods can handle a local single data point repeatedly (De Sortis and Paoliani, 2007; Leger and Leclerc, 2007). However, the computing time surges when processing a large amount of data. The repeated treatment of data without considering their relationships hinders the assessment of data error and redundancy, leading to difficulties in obtaining accurate information regarding the behavior of high dams. Thus, methods for extracting effective information from numerous data and improvement of the speed and precision of analysis in acquiring comprehensive information about service behavior have become topics of interest in the field of safety monitoring of high dams.

Recently, various new theories (Friedman *et al*., 2000; Su, 2003; Yang *et al*., 2006; Su *et al*., 2007; Bao *et al*., 2008; Huang *et al*., 2010; Wu *et al*., 2010) have been introduced to solve these issues. The preliminary formation of a structural health monitoring and safety evaluation analysis system has been achieved. On the other hand, studies focused on the application of information fusion theory to the health diagnosis and early-warning methods of high dams are insufficient.

In the present study, a new method based on multi-source information fusion is proposed to conduct a comprehensive analysis of the prototype monitoring data of high dams. Information fusion at the data level is conducted by computing the distance among the monitoring data statistical characters (i.e., confidence distance measure) of multi-positions and multi-points to reduce data metrics. Bayesian fusion is conducted at the decision level to obtain a comprehensive diagnosis and analysis outcomes. Decision information entropy analysis is proposed for carrying out the sorting of analysis results in multi-positions according to importance. This procedure is done to diagnose the service behavior of high dams and release early-warnings of abnormity on time.

2 Data metrics reduction

The monitoring data of high dams are huge data sets based on multiple positions, sections and points, and contain data abnormity caused by monitoring error, data noise and equipment damage. For the purpose of actualizing effective filtration and fast reduction, we use inner relevancy of data at the same position to conduct input data fusion. Commonly, methods such as optional clustering ant colony and means clustering are chosen for processing data clustering and reduction. These methods identify the main data characters through eliminating irrelevant data and data noise. However, they are limited by their thousands of iterations and non-unique calculation results when facing numerous monitoring data. Relevancy exists in data measured at the same sections or positions of a dam (Fig. 1), and becomes more apparent when large numbers of measurements are being taken and their trends and statistical distribution characters are being obtained. In this study, the confidence distance measure (CDM) is proposed to research the relevancy of data statistical distribution characters and for processing and analyzing data.

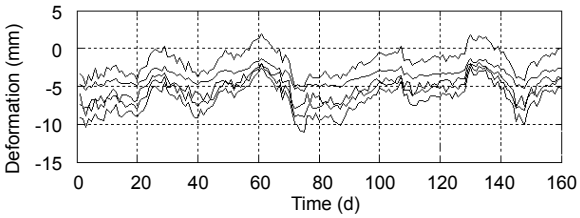

Fig. 1 Similarity of data measured at the same positions of a dam

CDM refers to a value that monitors the degree of deviation among measurement data of the same index parameter. Using m points to measure the same index parameter, the measurement sequence, mean values and deviation of points i and j are $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, x_i and x_j, σ_i and σ_j, respectively. Assuming these values obey a normal distribution, their probability density function curve as an Eigen function for points, can be written as $p_i(x|x_i)$ and $p_j(x|x_j)$. Then, CDM can reflect the deviation between two sequences:

$$d_{ij}=2\left|\int_{x_i}^{x_j}p_i(x\mid x_i)\mathrm{d}x\right|=2A, \tag{1}$$

$$d_{ji}=2\left|\int_{x_j}^{x_i}p_j(x\mid x_j)\mathrm{d}x\right|=2B. \tag{2}$$

where

$$p_i(x \mid x_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{1}{2}\left(\frac{x-x_i}{\sigma_i}\right)^2\right\}, \quad (3)$$

$$p_j(x \mid x_j)=\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{1}{2}\left(\frac{x-x_j}{\sigma_j}\right)^2\right\}, \quad (4)$$

A and B are integral areas surrounded by parts below the curves $p_i(x|x_i)$ or $p_j(x|x_j)$ and the intervals of (x_i, x_j) or (x_j, x_i), respectively (Fig. 2). The values of d_{ij} and d_{ji} vary between 0 and 1.

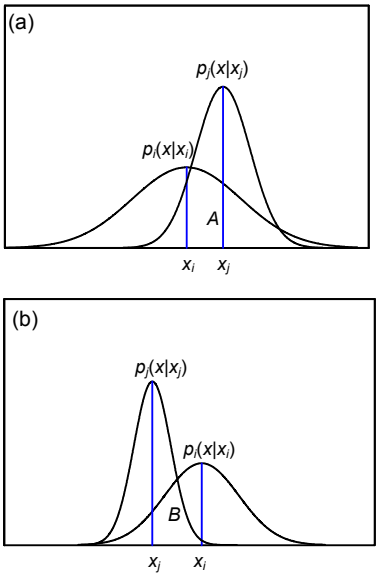

Fig. 2 Probability distribution and CDM for *A* (a) and *B* (b)

Usually, the value of d_{ij} can be calculated by the error function erf(θ):

$$\mathrm{erf}(\theta)=\frac{2}{\pi}\int_0^{\theta}\mathrm{e}^{-u^2}\mathrm{d}u. \quad (5)$$

Assuming $u=\dfrac{x-x_i}{\sqrt{2}\sigma_i}$, $\mathrm{d}u=\dfrac{\mathrm{d}x}{\sqrt{2}\sigma_i}$, then Eq. (5) can be rewritten as follows:

$$\mathrm{erf}(\theta)=\frac{2}{\pi\sigma_i}\int_{x_i}^{x_i+\sqrt{2}\theta\sigma_i}\exp\left\{-\frac{1}{2}\left(\frac{x-x_i}{\sigma_i}\right)^2\right\}\mathrm{d}x. \quad (6)$$

If $x_j=x_i+\sqrt{2}\theta\sigma_i$, $\theta=\dfrac{x_j-x_i}{\sqrt{2}\sigma_i}>0$, then

$$\mathrm{erf}\left(\frac{x_j-x_i}{\sqrt{2}\sigma_i}\right)=\frac{\sqrt{2}}{\sqrt{\pi}}\int_{x_i}^{x_j}\exp\left\{-\frac{1}{2}\left(\frac{x-x_i}{\sigma_i}\right)^2\right\}\mathrm{d}x =2\int_{x_i}^{x_j}p_i(x \mid x_i)\mathrm{d}x. \quad (7)$$

Comparing with Eq. (1), d_{ij} can be expressed as

$$d_{ij}=\left|\mathrm{erf}\left(\frac{x_j-x_i}{\sqrt{2}\sigma_i}\right)\right|. \quad (8)$$

Similarly,

$$d_{ji}=\left|\mathrm{erf}\left(\frac{x_i-x_j}{\sqrt{2}\sigma_j}\right)\right|. \quad (9)$$

For m points, a confidence distance matrix $\boldsymbol{D}_m$ is formed by calculating the CDM d_{ij} (i, j =1, 2, …, m) between each pair. $\boldsymbol{D}_m$ is given as

$$\boldsymbol{D}_m=\begin{pmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & & \vdots \\ d_{m1} & \cdots & d_{mm} \end{pmatrix}. \quad (10)$$

Obviously, the smaller the CDM value, the smaller is the distance of measurement sequences between two points, as well as the difference. To distinguish the difference, ε is set as a threshold. If $d_{ij}<\varepsilon$, the points of two sides have a good coherence; called point i supports point j. Conversely, if the points of two sides have a bad coherence; called point i does not support point j. A hard threshold is often chosen to directly partition the data difference. However, given the fact that the value of d_{ij} is between 0 and 1, which is a relatively fuzzy range and varies with the change of monitoring data, the Otsu (1979) method has been used to conduct data partitioning to achieve a better dynamic partition of numerous data. The Otsu method parts CDM quickly by analyzing the maximum between-cluster variance of data and converts CDM to degree of support to form a relationship matrix of degree of support between points. The matrix is

$$\boldsymbol{R}_m=\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{m1} & \cdots & r_{mm} \end{pmatrix}. \quad (11)$$

Based on $\boldsymbol{R}_m$, the comprehensive degree of support (CDS) can be calculated to reflect the support of one certain point by other data. The CDS for each point is computed by calculating the maximum eigenvalues λ and corresponding eigenvector $\boldsymbol{Y}$ of $\boldsymbol{R}_m$:

$$\lambda \boldsymbol{Y} = (\lambda y_1, \lambda y_2, \cdots, \lambda y_m) = \boldsymbol{R}_m^{\mathrm{T}} \boldsymbol{Y}, \tag{12}$$

which is expanded to

$$\lambda y_k = y_1 r_{1k} + y_2 r_{2k} + \cdots + y_m r_{mk}, k = 1, 2, ..., m, \tag{13}$$

where λy_k represents the support of m points to point k. A larger value of the degree of support means the point is more reliable and important. Otherwise, the point is supported only by a few points and is of low reliability, even invalid, and should be eliminated. Finally, data conducted by metrics reduction form an effective point data set (PDS) $\boldsymbol{X}$=$\{x_1, x_2, ..., x_n\}$. Subsequent analysis and diagnosis can then be applied to these data.

3 Bayesian fusion and analysis

If we focus only on single point analysis in processing large amounts of data, it is hard to acquire the complete behavior of the dam and contradictory analysis results could be obtained due to the effect of data error invalidation. Bayesian fusion method fuses analysis results to build an effective method that can draw a unified conclusion. It forms a decision fusion set (DFS) including analysis decision variables by aggregating analysis results at each point; it then conducts fusion to the analysis results in realizing the unity of analysis and decision. The following section will describe decision variables fusion based on the Bayesian fusion method.

3.1 Bayesian estimation and dynamic learning of decision variables

In moment algorithm, maximum total-probability algorithm and maximum likelihood algorithm, the unknown parameter θ is defined as a non-random variable to estimate parameters. Sequentially, the basic idea of Bayesian estimation is to obtain information additional to θ in advance, so that the precision of estimating θ can be improved. The specific theory of Bayesian estimation is described as follows: assuming state θ is $\{\theta_1, \theta_2, \cdots, \theta_k\}$, $p(\theta_i)$ stands for the prior probability and $p(x|\theta_i)$ stands for probability of event x under the condition of state θ_i. The posterior probability $p(\theta_i|x)$ then can be calculated as

$$p(\theta_i \mid x) = \frac{p(x \mid \theta_i) p(\theta_i)}{\sum_{i=1}^{k} p(x \mid \theta_i) p(\theta_i)}. \tag{14}$$

This means that prior probability $p(\theta_i)$ can be converted to $p(\theta_i|x)$ by observing event x, where $p(\theta_i)$ is treated as additional information of an unknown parameter, so that the estimation precision is improved.

According to Eq. (14), we can develop the specific calculation processes based on Bayesian theory. Assuming DFS $\boldsymbol{X}$ obeys $N(\mu, \sigma^2)$ and parameter μ obeys $N(\mu_0, {\sigma_0}^2)$:

$$\begin{aligned} p(\mu \mid \boldsymbol{X}) &= \frac{p(\boldsymbol{X} \mid \mu) p(\mu)}{\int p(\boldsymbol{X} \mid \mu) p(\mu) \mathrm{d}\mu} \\ &= \alpha \prod_{k=1}^{n} p(x_k \mid \mu) p(\mu) \\ &= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right\}}^{p(x_k|\mu)} \\ &\quad \cdot \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right\}}^{p(\mu)} \\ &= \alpha' \exp\left\{-\frac{1}{2}\left[\left(\frac{l}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}, \end{aligned} \tag{15}$$

where $\alpha = \dfrac{1}{p(\boldsymbol{X})} = \dfrac{1}{P(x_1, x_2, \ldots x_n)}$, α and α' are factors depending on fusion set $\boldsymbol{X}$ and irrelevant to μ. This is an exponential function and its exponential part is a quadratic function about μ. Thus, $p(\mu|\boldsymbol{X})$ obeys a normal distribution and keeps its distribution as the number n of fusion sets increases. It also can be written as

$$p(\mu \mid \boldsymbol{X}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]. \quad (16)$$

Comparing parameters of Eqs. (15) and (16) and according to the principle that the corresponding parts are equal, we can obtain μ_n and σ_n^2:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0, \quad (17)$$

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}, \quad (18)$$

where $\bar{\mu}_n$ is the mean value of fusion set $\boldsymbol{X}$. Then, the Bayesian estimation $\hat{\mu}$ of μ can be computed after acquiring the distribution of μ:

$$\hat{\mu} = \int_{\Omega} \mu \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{ -\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right\} \mathrm{d}\mu, \quad (19)$$

where $\hat{\mu}$ is the optimal fusion value in the fusion set and is more accurate and objective. Given that $p(\mu|\boldsymbol{X})$ obeys a normal distribution, $\hat{\mu} = \mu_n$. Commonly, σ_n^2 stands for the degree of estimation uncertainty. Eqs. (17) and (18) show that as the number n of fusion variables keeps on increasing, μ_n gradually approaches $\bar{\mu}_n$, σ_n^2 to δ^2/n. Meanwhile, the distribution of $p(\mu|\boldsymbol{X})$ becomes sharper, with its uncertainty degree decreasing continually, and reaches its peak around the real value of the estimating parameters. Its distribution then approaches a Dirac function, thereby forming a dynamic learning process (Fig. 3).

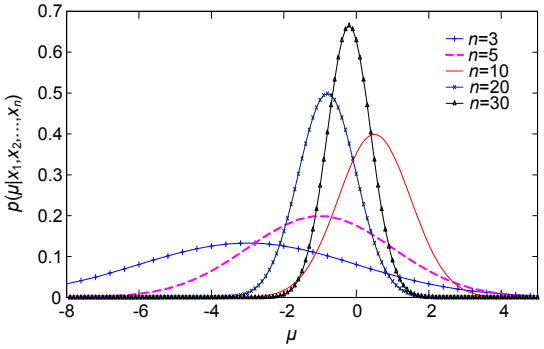

Fig. 3 Changing process of posterior probability

Information, including monitoring data and decision variables, is not provided during the real time monitoring of dams, whereas it is gradually deepening and changing in the course of time. The Bayesian learning method can not only realize the optimal analysis of existing decision variables, but also fuse with new decision information with the passage of time in order to obtain more precise and real-time results. Consequently, this method is more suited to conduct the online analysis for safety monitoring of dams. Considering that each new analysis updates based on the previous analysis, the first analysis should assume its estimating parameters obey a uniform distribution or use maximum likelihood algorithm to estimate because of the lack of prior information. The following analysis chooses the posterior probability distribution fusing results to be the prior analysis for the next step, and employs new decision variables to fuse and update. Since the behavior of dams changes dynamically, it is necessary to analyze data from the latest period and gradually abandon part of the old information when fusing new information. Moreover, when external environmental factors (such as weather and water level) change a lot, the time period should be relatively short in order to reflect the fluctuation of a dam's behavior. Conversely, the time period could be relatively long if a dam's behavior is stable, in order to improve the precision of the analysis results.

3.2 Selection of diagnosis algorithm

Diagnosis algorithm is used to analyze observed data, acquire a dam's behavior reflected by a single point and provide decision variables to check for abnormal behavior. Commonly used diagnosis algorithms are gray model, time series model, ANN model, wavelet analysis, and combination forecasting models (Hecht-Nielsen, 1989; Kim and Melhem, 2004; Trivedi and Singh, 2005; Yuen and Lam, 2006; Chen, 2009). These methods are seldom applied in practical projects because of their low calculation efficiency. Apart from these algorithms, statistical models have been studied in detail and applied to varied types of dams and monitoring data analysis (De Sortis and Paoliani, 2007; Leger and Leclerc, 2007; Su *et al*., 2007). Consequently, a statistical model was chosen as the main diagnosis algorithm. Other models can be considered when a dam's behavior has a hysteretic

nature and uncertainty, the monitoring data are limited or the error is too large.

The method for calculating decision variables for the statistical model is described below:

After acquiring survey points data at a certain position in the dam, we conduct a diagnosis at each survey point. Taking deformation data as an example, the model can be demonstrated as

$$\hat{\delta} = a_0 + \hat{\delta}(H) + \hat{\delta}(T) + \hat{\delta}(t), \tag{20}$$

where $\hat{\delta}$ is the amount of deformation; $\hat{\delta}(H)$, $\hat{\delta}(T)$, and $\hat{\delta}(t)$ are components of hydraulic pressure, temperature, and time effects, respectively. The effect of external environmental loads on dam structure is shown by varied combinations of hydraulic pressure, temperature, and time effects.

The predictive value $\hat{\delta}$ of the deformation amount can be calculated from Eq. (20), and then to calculate the redundancy of relative δ (this is an observed value): $\Delta\delta = |\delta - \hat{\delta}|$. Comparing $\Delta\delta$ with the standard deviation S of the model, the behavior of a dam is divided into normal behavior, suspected-abnormal behavior and abnormal behavior. Assuming threshold values for suspected-abnormal and abnormal behaviors are E_a=2S and E_b=3S, respectively, then the single point diagnosis equation for the i survey point is

$$\begin{cases} 0 \le \Delta\delta_i \le E_{ai}, \text{ normal}, \\ E_{ai} < \Delta\delta_i \le E_{bi}, \text{ suspected-abnormal}, \\ \Delta\delta_i > E_{bi}, \text{ abnormal}. \end{cases} \tag{21}$$

Based on Eq. (21), decision variables y_a, y_b and DFS $\boldsymbol{Y}_a$, $\boldsymbol{Y}_b$ are obtained when the behavior of a dam is between the suspected-abnormal and abnormal intervals, as follows:

$$\begin{aligned} \boldsymbol{Y}_a &= \{y_{a1}, y_{a2}, ..., y_{an}\} \\ &= \{\Delta\delta_1 - E_{a1}, \Delta\delta_2 - E_{a2}, ..., \Delta\delta_n - E_{an}\}, \end{aligned} \tag{22}$$

$$\begin{aligned} \boldsymbol{Y}_b &= \{y_{b1}, y_{b2}, ..., y_{bn}\} \\ &= \{\Delta\delta_1 - E_{b1}, \Delta\delta_2 - E_{b2}, ..., \Delta\delta_n - E_{bn}\}. \end{aligned} \tag{23}$$

Then, the optimal fusion values for decision variables can be obtained by Bayesian fusion, subsequently corresponding suspected-abnormal and abnormal values $\hat{\mu}_a$ and $\hat{\mu}_b$, respectively.

3.3 Sorting of comprehensive diagnosis results at multi-positions according to their importance

Decision variables and the unknown uncertainty factors of decision and fusion results vary in the course of gradual decision fusing. The facticity and degree of importance of diagnosis decision results at multi-positions differ from each other. Hence, we introduce information entropy theory (IET) to estimate and compare the diagnosis and decision results to evaluate their relative importance, which will actualize the decision information entropy analysis (DIE).

If the possible probability distribution of a decision variable is $p(\mu)$, then scatter (i.e., uncertainty) of its distribution can be calculated as

$$d = -\sum_{\mu} p(\mu) \ln p(\mu). \tag{24}$$

When two possible probability distributions $p^{i-1}(\mu)$ and $p^i(\mu)$ of a decision variable at the same position are obtained, in order to weigh the relative value of uncertainty between them, the relative uncertainty is defined as

$$D_{i-1,i} = \frac{d_{i-1}}{d_i} = \frac{\sum_{\mu} p^{i-1}(\mu) \ln p^{i-1}(\mu)}{\sum_{\mu} p^i(\mu) \ln p^i(\mu)}. \tag{25}$$

where $D_{i\text{-}1,i}$ stands for the ratio of the degree of scatter between the former and latter decisions. A larger $D_{i\text{-}1,i}$ represents a relatively high degree of scatter of the former decision and its uncertainty. Conversely, it represents a decrease in the degree of scatter of the latter compared to the former decision and an increase in certainty. Therefore, $D_{i\text{-}1,i}$ indicates the relative increase trend of certainty for the integer decision. With the degree of scatter being gradually reduced and the high uncertainty, the decision process could be more true and important. If this decision process is neglected, there will be a high risk of causing loss (Fig. 4).

Since the fusion value $\hat{\mu}$ of the decision result represents the overall trend of the decision distribution,

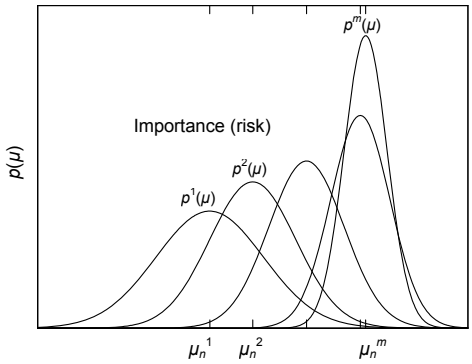

Fig. 4 Changing process of decision importance (risk)

a bigger $\hat{\mu}$ indicates a more obvious decision trend. Consequently, the relative importance can be defined by the combination of the fusion value $\hat{\mu}$ of the decision variables and relative uncertainty.

The relative importance of a single-step is

$$H = \frac{\hat{\mu}^i}{\hat{\mu}^{i-1}} D_{i-1,i}. \tag{26}$$

The cumulative relative importance of m-step is

$$H = \prod_{i=1}^{m} \frac{\hat{\mu}^i}{\hat{\mu}^{i-1}} D_{i-1,i} = \frac{\hat{\mu}^m}{\hat{\mu}^0} D_{0,m}. \tag{27}$$

When diagnosing different positions, different degrees of importance of diagnostic decision outcomes at multi-positions can be obtained by calculating the cumulative importance of each decision change. If behaviors at multi-positions are determined as abnormal, the most importance position should be assigned the most serious early-warning, to emphasize its importance and risk.

4 Service behavior fusion diagnosis and early-warning system for high dams

A flowchart of a high dam service behavior fusion diagnosis (BFD) scheme is shown in Fig. 5. The changing process of a dam's behavior is categorized as three states: normal behavior, suspected-abnormal behavior and abnormal behavior. When fusing diagnosis and analysis, CDM is used to estimate similar changes of multi-points at the same position (or section) on the dam under the same environmental loads. This synergistic effect analysis can eliminate invalid survey points. In order to avoid the effect of different testing values within the range of survey points on the synergistic effect, testing values are first normalized. The confidence distance matrix $\boldsymbol{D}_m$ and support matrix $\boldsymbol{R}_m$ are then calculated and the CDS of the survey points is acquired to conduct deletion and choice to obtain a fusion set of survey point data, which is $\boldsymbol{X}=\{x_1, x_2, ..., x_n\}$. On that basis, the fusion set of data for the decision set of diagnosis $\boldsymbol{Y}=\{y_1, y_2, ..., y_n\}$ are converted using the current diagnosis model. Judgment of whether this position behaves suspected-abnormally can be made according to single-step fusion diagnosis. If certain positions behave suspected-abnormally, multi-step fusion diagnosis is applied and their cumulative importance calculated. When abnormal behavior is checked, early-warning should be undertaken. When abnormity at multi-positions appears simultaneously, we export the early-warning outcomes whose cumulative importance has been sorted. The specific methods of fusion diagnosis for single-step and multi-step are explained below.

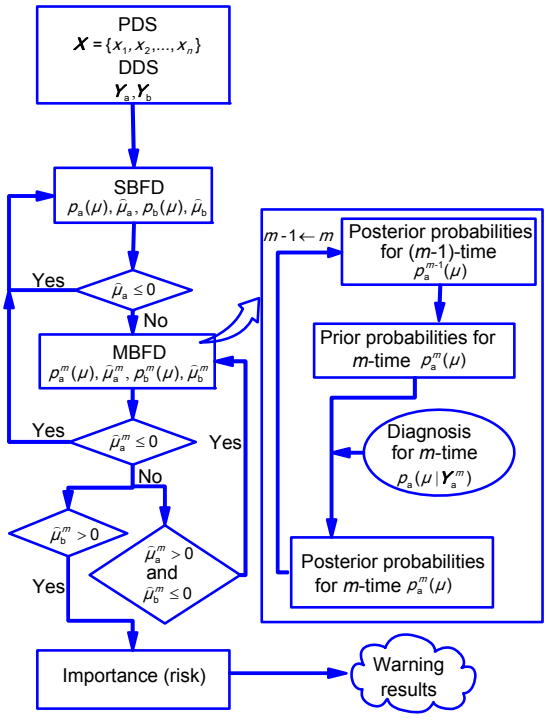

Fig. 5 Flowchart of service behavior fusion diagnosis (BFD)

4.1 Single-step Bayesian fusion diagnosis (SBFD)

The first step of SBFD is to conduct single point diagnosis and calculation of data to obtain DFS $\boldsymbol{Y}_a$ including decision variables. After Bayesian fusion, the overall suspected-abnormal decision distribution $p_a(\mu)$ and fusion value $\hat{\mu}_a$ of n survey points at this position are calculated as

$$p_a(\mu)=p_a(\mu\,|\,\boldsymbol{Y}_a)=p_a(\mu\,|\,y_{a1},y_{a2},...,y_{an})=\frac{p_a(\mu;y_{a1},y_{a2},...,y_{an})}{p_a(y_{a1},y_{a2},...,y_{an})}, \tag{28}$$

$$\hat{\mu}_a=\int p_a(\mu)\mu \mathrm{d}\mu. \tag{29}$$

Equations for calculating the abnormal decision distribution $p_b(\mu)$ and fusion value $\hat{\mu}_b$ are similar to Eqs. (28) and (29). In daily monitoring, if the fusion value $\hat{\mu}_a>0$, it means that there is a suspected-abnormity at this position, and therefore multi-step bayesian fusion diagnoses (MBFD) are applied to this position.

4.2 Multi-step Bayesian fusion diagnoses (MBFD)

When conducting MBFD to suspected-abnormal positions, the real-time decision distribution and fusion values are calculated by upgrading diagnosis decision variables. It is a Bayesian learning process. Assuming $p_a^m(\mu)$ and $\hat{\mu}_a^m$ are the suspected-abnormal decision distribution and fusion value respectively of m fusion, and $\boldsymbol{Y}_a^m$ is its suspected-abnormal decision set, then

$$p_a^m(\mu)=p_a(\mu\,|\,\boldsymbol{Y}_a^1,\boldsymbol{Y}_a^2,\cdots,\boldsymbol{Y}_a^m)=\frac{p_a(\boldsymbol{Y}_a^m\,|\,\mu)p_a^{m-1}(\mu)}{p_a(\boldsymbol{Y}_a^m)}=\frac{p_a(\boldsymbol{Y}_a^m\,|\,\mu)p_a(\mu\,|\,\boldsymbol{Y}_a^1,\boldsymbol{Y}_a^2,\cdots,\boldsymbol{Y}_a^{m-1})}{p_a(y_{a1}^m,y_{a2}^m,...,y_{an}^m)}, \tag{30}$$

$$\hat{\mu}_a^m=\int p_a^m(\mu)\mu \mathrm{d}\mu, \tag{31}$$

where $p_a^{m-1}(\mu)$ is the posterior probability distribution of the former fusion diagnosis and prior distribution of this time. Corresponding equations for abnormal diagnosis decision distribution $p_b(\mu)$ and fusion value $\hat{\mu}_b^m$ are defined by substituting b for a in Eqs. (30) and (31).

In m diagnosis process, MBFD can stop and return to SBFD of this position when the primary suspected-abnormity falls into the normal range, i.e., when $\hat{\mu}_a^m\le 0$. When the primary suspected-abnormity falls into the abnormal state, $\hat{\mu}_b^m>0$, its cumulative importance is calculated. If k positions appear abnormal simultaneously, the relative sizes of their cumulative importance are compared and sorted early-warning results at the positions are exported:

$$H_i=\frac{\hat{\mu}_a^m}{\hat{\mu}_a^0}\frac{\sum_\mu p_a^0(\mu)\ln p_a^0(\mu)}{\sum_\mu p_a^m(\mu)\ln p_a^m(\mu)},\quad i=1,2,\cdots,k,$$
$$\text{Alarm}\rightarrow\text{Sort descending }\{H_1,H_2,\cdots,H_k\}. \tag{32}$$

5 Case study

The case study aims to employ the information fusion diagnosis and early-warning method to diagnose and analyze the observed data from multi-positions of an arch dam in order to obtain their diagnosis results.

5.1 Description of the project

A hydropower station is located in the upstream of the Yellow River in Qinghai Province, China. It serves mainly power generation, with consideration of water supply, flood control, and irrigation. The water retaining structure is composed of a concrete hyperbolic arch dam and a gravity buttress on the left bank. The height of the arch dam is 155 m and its installed capacity is 1200 MW from 20 dam sections. The hydropower was used for impoundment in 1996 and for power generation the following year. The project investigation indicated that the geological structure of the dam foundation developed well. There are a number of weak crushed zones along the river, which result in a crashing structure and low deformation modulus of rock body in this region. The rock body of the riverbed foundation in this region is deformed. After ten years of operation, the behavior of the dam has suffered a great change. For this study, we have chosen the radial displacement data from November, 2005 to November, 2009 of a

dam section of riverbed to diagnose and analyze the service behavior of the dam.

5.2 Acquisition of a fusion set of the survey points

We chose the radial displacement data of the dam body, which are observed from the vertical elevation survey points of dam sections No. 6, No. 9 and No. 11 located in the riverbed. Twelve survey points were analyzed and diagnosed with four survey points at each dam section. According to the differences among dam sections, the survey points were categorized to three groups, recorded as: p6-1–p6-4, p9-1–p9-4, and p11-1–p11-4. After normalizing the data between 0 and 1, mean values x and the degree of scatter σ were obtained (Table 1).

CDM of survey points in each group were calculated using Eqs. (8) and (9), and then their support relationship matrix $\boldsymbol{R}_m$ was obtained. Relationship matrixes of sections No. 6, No. 9 and No. 11 were denoted $\boldsymbol{R}_4^1$, $\boldsymbol{R}_4^2$, and $\boldsymbol{R}_4^3$, respectively.

$$\boldsymbol{R}_4^1=\begin{bmatrix}1.00 & 0.19 & 0.30 & 0.19\\ 0.03 & 1.00 & 1.00 & 1.00\\ 0.08 & 1.00 & 1.00 & 1.00\\ 0.03 & 1.00 & 1.00 & 1.00\end{bmatrix},$$

$$\boldsymbol{R}_4^2=\begin{bmatrix}1.00 & 0.19 & 0.30 & 0.19\\ 0.03 & 1.00 & 1.00 & 1.00\\ 0.08 & 1.00 & 1.00 & 1.00\\ 0.03 & 1.00 & 1.00 & 1.00\end{bmatrix},$$

$$\boldsymbol{R}_4^3=\begin{bmatrix}1.00 & 0.79 & 1.00 & 1.00\\ 0.85 & 1.00 & 0.53 & 0.96\\ 1.00 & 0.35 & 1.00 & 0.99\\ 1.00 & 0.93 & 0.98 & 1.00\end{bmatrix}.$$

Subsequently, CDS (Table 2) for each survey point was calculated using the maximum modulus eigenvalue λ and the corresponding eigenvector $\boldsymbol{Y}$. Comparison of the CDS values for survey points in each group (Fig. 6) showed that the degree of support for most survey points was below 1.6. The degrees of support for p6-1 and p9-2 in their own groups were below 0.6, and thus they were eliminated. The remaining 10 points formed three groups of survey points DFS: {p6-2, p6-3, p6-4}, {p9-1, p9-3, p9-4} and {p11-1, p11-2, p11-3, p11-4}.

5.3 Single-step diagnosis

The statistical models were established for the 10 points using data from November, 2005 to October, 2009 based on Eq. (20). For an arch dam, the hydraulic pressure component is multinomial with a maximum power of 4. The multiple correlation coefficient R and standard deviation S of the survey points data were then computed (Table 3).

The SBFD began in November of the same year for these three dam sections. Relatively large radial displacements were found when analyzing data on Nov. 10, 2009. Observed and simulated values for each point were obtained and the difference between

Table 1 Mean values and degree of scatter of survey points

Point No.	x (mm)	σ	Point No.	x (mm)	σ
p6-1	0.449	0.072	p9-3	0.405	0.058
p6-2	0.389	0.055	p9-4	0.428	0.068
p6-3	0.396	0.054	p11-1	0.396	0.057
p6-4	0.389	0.056	p11-2	0.389	0.054
p9-1	0.400	0.055	p11-3	0.402	0.059
p9-2	0.365	0.052	p11-4	0.386	0.056

Table 2 Comprehensive degree of support of survey points

Point No.	CDS	Point No.	CDS
p6-1	0.578	p9-3	1.746
p6-2	1.705	p9-4	1.676
p6-3	1.715	p11-1	1.898
p6-4	1.707	p11-2	1.679
p9-1	1.735	p11-3	1.690
p9-2	0.596	p11-4	1.951

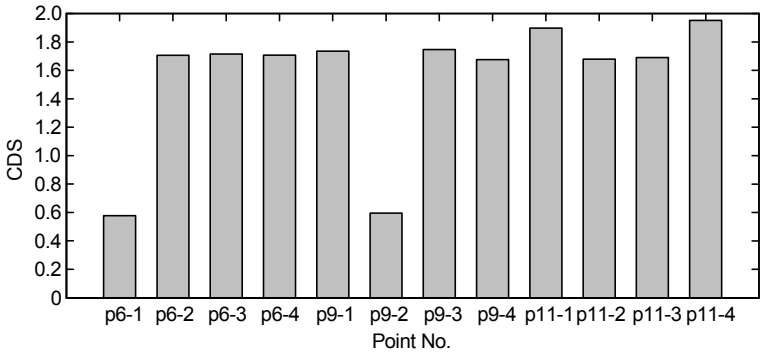

Fig. 6 Comparison of CDS for each survey point

Table 3 Multiple correlation coefficient and standard deviation of survey point data

No.	Point No.	R	S	No.	Point No.	R	S
1	p6-2	0.904	0.628	6	p9-4	0.954	0.531
2	p6-3	0.905	0.857	7	p11-1	0.977	0.379
3	p6-4	0.909	0.653	8	p11-2	0.949	0.447
4	p9-1	0.914	0.525	9	p11-3	0.940	0.461
5	p9-3	0.958	0.284	10	p11-4	0.956	0.576

Table 4 Suspected-abnormal data diagnosed by SBFD

Point No.	δ (mm)	$\hat{\delta}$ (mm)	$\Delta\delta$ (mm)	E_a (mm)	$\boldsymbol{Y}_a$ (mm)
p6-2	20.053	18.933	1.120	1.256	−0.136
p6-3	16.321	14.951	1.370	1.714	−0.344
p6-4	12.244	11.199	1.045	1.306	−0.261
p9-1	22.205	20.965	1.240	1.050	0.900
p9-3	18.237	17.783	0.454	0.568	−0.114
p9-4	15.189	13.969	1.220	1.062	0.158
p11-1	25.064	24.114	0.950	0.758	0.192
p11-2	19.454	18.454	1.000	0.894	0.106
p11-3	18.521	17.431	1.090	0.922	0.168
p11-4	13.260	14.282	1.022	1.152	−0.130

them and the suspected-abnormity limitation E_a were calculated. Then the suspected-abnormity decision set $\boldsymbol{Y}_a$ was defined to conduct fusion diagnosis. The analysis results are listed in Table 4. The suspected-abnormity fusion values $\hat{\mu}_a$ for dam sections No. 6, No. 9 and No. 11 were −0.247 mm, 0.078 mm and 0.084 mm, respectively, and corresponding variances σ_a were 0.114, 0.167 and 0.157, respectively. Eq. (21) and Fig. 7 indicate that the fusion diagnosis results of dam section No. 6 were normal, and those of dam sections No. 9 and No. 11 were

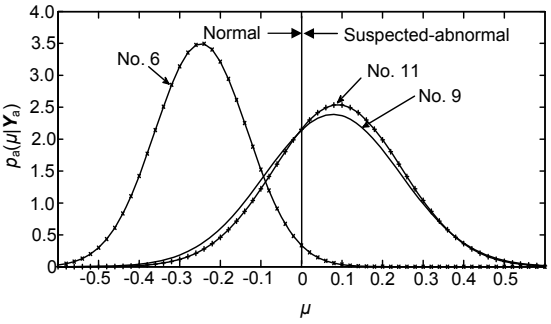

Fig. 7 SBFD distribution of suspected-abnormity

suspected-abnormal with their fusion values $\hat{\mu}_a>0$, meaning that these positions are suspected-abnormal.

5.4 Multi-step diagnosis

The fusion diagnosis of Nov. 10, 2009 is regarded as the first step in conducting MBFD for dam sections No. 9 and No. 11. The fusion values $\hat{\mu}_a$ $\hat{\mu}_b$ and variances σ_a, σ_b of suspected-abnormal and abnormal diagnosis distributions are listed in Table 5, and their development trends are depicted in Fig. 8. The fusion values $\hat{\mu}_a$ peak at step 3. The degrees of scatter σ_a are relatively small and the development trends come to a maximum. Fusion diagnosis results of the following steps 4 and 5 gradually return to the normal range. In the whole diagnosis process, abnormal fusion values $\hat{\mu}_b$ remain lower than 0, which indicates that these two sections do not show abnormal behavior. Consequently, it is unnecessary to release an early-warning. These sections can return to SBFD.

In MBFD, steps from 1 to 4 fall within the suspected-abnormity range. Calculating the single-step and multi-step cumulative relative importance of sections No. 9 and No. 11 for the first step of No. 9 is convenient for comparison. Table 6 demonstrates

Table 5 Suspected-abnormal data of MBFD

Step	Date	No. 9 section				No. 11 section			
		$\hat{\mu}_a$ (mm)	σ_a	$\hat{\mu}_b$ (mm)	σ_b	$\hat{\mu}_a$ (mm)	σ_a	$\hat{\mu}_b$ (mm)	σ_b
Step 1	Dec. 11, 2009	0.187	0.163	−0.260	0.258	0.220	0.151	−0.245	0.239
Step 2	Dec. 12, 2009	0.271	0.151	−0.176	0.239	0.310	0.135	−0.151	0.213
Step 3	Dec. 13, 2009	0.410	0.114	−0.037	0.180	0.453	0.108	−0.013	0.177
Step 4	Dec. 14, 2009	0.276	0.173	−0.171	0.274	0.313	0.148	−0.152	0.234
Step 5	Dec. 15, 2009	−0.037	0.164	−0.483	0.259	−0.013	0.162	−0.478	0.256

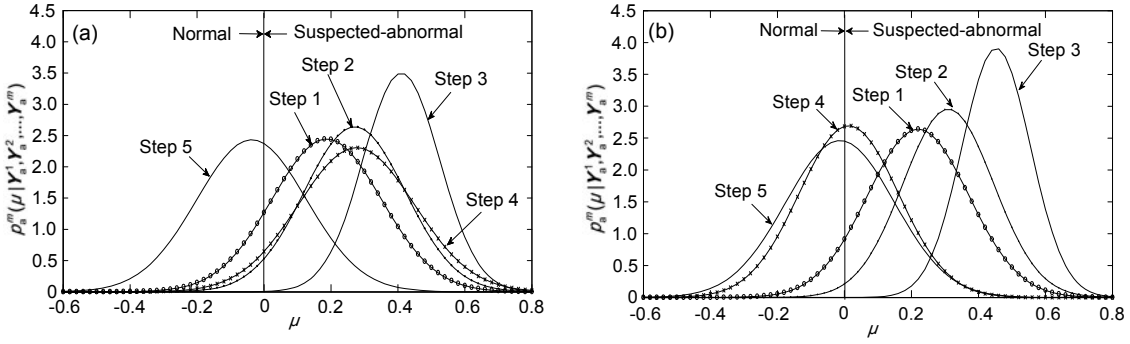

Fig. 8 Distribution results of MBFD
Development process of possible abnormal distribution for No. 9 dam part (a) and No. 11 dam part (b)

Table 6 Relative importance of single-step and multi-step

Step	Date	No. 9 section		No. 11 section	
		H (single-step)[a]	H (multi-step)[b]	H (single-step)[a]	H (multi-step)[b]
Step 1	Dec. 11, 2009	2.33	2.33	2.71	2.71
Step 2	Dec. 12, 2009	1.34	3.11	1.36	3.42
Step 3	Dec. 13, 2009	1.17	3.65	1.34	4.26
Step 4	Dec. 14, 2009	0.98	3.58	0.95	3.77

[a] According to Eq. (26); [b] According to Eq. (27)

that the cumulative relative importance of each dam section peaks at step 3 and the importance of section No. 11 is bigger than that of No. 9. Examination of Fig. 8 leads to the same conclusion, that the diagnosis distribution of No. 11 has a relatively clear trend and a small degree of scatter.

6 Conclusions

In this paper, we propose a new method to conduct a comprehensive analysis of the dense prototype monitoring data of high dams based on a combination of information fusion theory and existing analysis models. It has been applied to a project case and obtained satisfactory results:

1. Considering the cooperating effect of changes among multi-points at the same position, CDM is employed to measure and select the survey points. It can eliminate invalid points, reduce data metric and acquire DFS. This process can realize the acquisition of effective information.

2. For the DFS of multi-points at the same position, BFD is undertaken to capture the overall trend and acquire the overall fusion diagnosis results of that position. Thus, comprehensive fusion of varied results is actualized.

3. SBFD is used to distinguish suspected-abnormal behavior of multi-positions at the dam body. For the few positions diagnosed as abnormal, MBFD is conducted expressly. This procedure gradually narrows the monitoring diagnosis range and improves the speed and efficiency of diagnosis.

4. In MBFD, the latest diagnosis information is fused into the primary decision continually when suspected-abnormal positions are being monitored and analyzed. This can greatly improve the real-time character and precision of diagnosis.

5. The analysis of the development trend and the degree of scatter of the diagnosis decision distribution imports uncertainty into the final diagnosis results. Thus, different degrees of importance of diagnosis and early-warning at multi-positions can be acquired. These results are helpful to the function and maintenance of subsequent projects.

References

Bao, T.F., Wu, Z.R., Gu, C.S., 2008. Influence of fractality of fracture surfaces on stress and displacement fields at crack tips. *Science China Serial E-Technological Sciences*, **51**(Supp II):95-100. [doi:10.1007/s11431-008-6004-3]

Chen, S.M., 2009. A New Method to Forecast Enrollments Using Fuzzy Time Series and Clustering Techniques. International Conference on Machine Learning and Cybernetics, Baoding, China, p.3026-3029. [doi:10.1109/ICMLC.2009.5212604]

De Sortis, A., Paoliani, P., 2007. Statistical analysis and structural identification in concrete dam monitoring. *Engineering Structures*, **29**(1):110-120. [doi:10.1016/j.engstruct.2006.04.022]

Friedman, N., Linial, M., Nachman, I., 2000. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3-4):601-620. [doi:10.1089/106652700750050961]

Hecht-Nielsen, R., 1989. Theory of the Back Propagation Neural Network. International Joint Conference on Neural Networks, Washington DC, USA, p.593-605. [doi:10.1109/IJCNN.1989.118638]

Huang, H.W., Yang, J.N., Zhou, L., 2010. Comparison of various structural damage tracking techniques based on experimental data. *Smart Structure and Systems*, **6**(9):1057-1077. [doi:10.1117/12.774621]

Kim, H.S., Melhem, H., 2004. Damage detection of structures by wavelet analysis. *Engineering Structures*, **26**(3):347-362. [doi:10.1016/j.engstruct.2003.10.008]

Leger, P., Leclerc, M., 2007. Hydrostatic, temperature, time-displacement model for concrete dams. *Journal of Engineering Mechanics*, **133**(3):267-277. [doi:10.1061/(ASCE)0733-9399(2007)133:3(267)]

Mata, J., 2011. Methods of analysis for the prediction and the verification of dam behavior. *Engineering Structures*, **33**(3):903-910. [doi:10.1016/j.engstruct.2010.12.011]

Otsu, N., 1979. A threshold selection method from gray level histograms. *IEEE Transactions on Systems Man and Cybernetics*, **9**(1):62-66. [doi:10.1109/TSMC.1979.4310076]

Su, H.Z., 2003. Intelligent Sensing and Fusion System and Its Application to Dam Safety Monitoring. MS Thesis, Hohai University, Nanjing, China (in Chinese).

Su, H.Z., Wu, Z.R., Wen, Z.P., 2007. Identification model for dam behavior based on wavelet network. *Computer-Aided Civil and Infrastructure Engineering*, **22**(6):438-448. [doi:10.1111/j.1467-8667.2007.00499.x]

Trivedi, H.V., Singh, J.K., 2005. Application of grey system theory in the development of a runoff prediction model. *Biosystems Engineering*, **92**(4):521-526. [doi:10.1016/j.biosystemseng.2005.09.005]

Wu, Z.R., Su, H.Z., 2005. Dam health diagnosis and evaluation. *Smart Materials and Structures*, **14**(3):130-136. [doi:10.1088/0964-1726/14/3/016]

Wu, Z.R., Gu, C.S., 2006. Safety Diagnosis and Hidden Defects Detection of Major Hydraulic Concrete Structures. Higher Education Press, Beijing, China, p.1-4 (in Chinese).

Wu, Z.R., Li, J., Gu, C.S., 2007. Review on hidden trouble detection and health diagnosis of hydraulic concrete structures. *Science China Serial E-Technological Sciences*, **50**(1):34-50. [doi:10.1007/s11431-007-6003-9]

Wu, H.Y., Zhou, Z.Y., Xiong, S.S., Wang, X.H., Lan, J.H., 2010. A review of detection techniques for dam hidden defects. *Journal of Yangtze River Scientific Research Institute*, **17**(3):38-40 (in Chinese).

Yang, J., Hu, D.X., Wu, Z.R., 2006. Bayesian uncertainty inverse analysis method based on pome. *Journal of Zhejiang University (Engineering Science)*, **40**(5):801-808 (in Chinese).

Yuen, K.V., Lam, H.F., 2006. On the complexity of artificial neural networks for smart structures monitoring. *Engineering Structures*, **28**(7):977-984. [doi:10.1016/j.engstruct.2005.11.002]