Rainfall-induced landslide susceptibility assessment using random forest weight at basin scale

Chengguang Lai, Xiaohong Chen, Zhaoli Wang, Chong-Yu Xu and Bing Yang

ABSTRACT

Rainfall-induced landslide susceptibility assessment is currently considered an effective tool for landslide hazard assessment as well as for appropriate warning and forecasting. As part of the assessment procedure, a credible index weight matrix can strongly increase the rationality of the assessment result. This study proposed a novel weight-determining method by using random forests (RFs) to find a suitable weight. Random forest weights (RFWs) and eight indexes were used to construct an assessment model of the Dongjiang River basin based on fuzzy comprehensive evaluation. The results show that RF identified the elevation (EL) and slope angle (SL) as the two most important indexes, and soil erodibility factor (SEF) and shear resistance capacity (SRC) as the two least important indexes. The assessment accuracy of RFW can be as high as 79.71%, which is higher than the entropy weight (EW) of 63.77%. Two experiments were conducted by respectively removing the most dominant and the weakest indexes to examine the rationality and feasibility of RFW; both precision validation and contrastive analysis indicated the assessment results of RFW to be reasonable and satisfactory. The initial application of RF for weight determination shows significant potential and the use of RFW is therefore recommended.

Key words | Dongjiang River basin, objective weight, rainfall-induced landslide, random forest, susceptibility assessment

Chengguang Lai
Zhaoli Wang
School of Civil Engineering and Transportation,
South China University of Technology,
Guangzhou 510641,
China

Xiaohong Chen (corresponding author)
Bing Yang
Center for Water Resource and Environment,
Sun Yat-Sen University,
Guangzhou 510275,
China
E-mail: *eescxh@mail.sysu.edu.cn*

Chong-Yu Xu
State Key Laboratory of Water Resources and Hydropower Engineering Science,
Wuhan University,
Wuhan 430072,
China
and
Department of Geosciences,
University of Oslo,
PO Box 1047, Blindern,
Oslo N-0316,
Norway

INTRODUCTION

Rainfall-induced landslides are one of the most common geological hazards, occurring over a wide range of spatial and temporal scales in mountainous landscapes (McKean & Roering 2001). Despite considerable efforts of hazard prevention and risk management, landslides continue to present an acute threat to life and property. It has been estimated that the worldwide damage of landslides that have occurred since the 20th century, have resulted in more than 62,000 deaths and a loss of at least US$9.7 billion (EM-DAT 2016). Effective measures to prevent landslides are urgently required to reduce loss of life, property, and infrastructure.

Landslide susceptibility assessment is one of the important steps for landslide risk analysis and has been widely applied to a variety of spatial scales due to its convenient application and compatibility with the geographical information systems (GIS) (Balteanu *et al.* 2010; Christian *et al.* 2015). The assessment requires evaluation of the possibility of a landslide occurring in certain areas based on multiple indexes such as precipitation, topography, morphology, lithology, and land-use type. Since landslide susceptibility assessment is a synthesis and involves several variables, multiplicity, complexity, uncertainty, and inaccuracy inevitably exists during the process, which is a worldwide multivariable

doi: 10.2166/nh.2017.044

evaluation problem (Jiang & Eastman 2000; Pistocchi *et al.* 2002; Lee & Evangelista 2006; Lee & Pradhan 2007; Su *et al.* 2015). Fuzzy comprehensive evaluation (FCE) is a fuzzy mathematics method that is convenient for expressing and processing random, fuzzy, insufficient, or inexact data and other distribution information (Giachetti & Young 1997a, 1997b), and has previously been applied to landslide susceptibility assessment (Yang *et al.* 2016). However, determining a reasonable index weight for FCE requires extensive research because differences in weighting may cause considerable differences in the results. Three main types of weightings are frequently used in assessment systems in general: the subjective weight (SW), the objective weight (OW), and the combination weight (CW), which combines both SW and OW. SW and OW are determined by the intentions of the decision-maker and are strongly affected by expert knowledge and biases, resulting in high subjectivity (Jiang *et al.* 2009; Zou *et al.* 2013). A suitable index weight should objectively reflect the index importance and should not be influenced by the intentions of the decision-maker. For these purposes, OW is more suitable than the SW and CW. However, shortcomings exist in the commonly used OW methods, e.g., entropy theory (Li *et al.* 2012; Jesmin & Sharif 2014; Yan *et al.* 2014), the criteria importance though inter-criteria correlation (CRITIC) method (Diakoulaki *et al.* 1995; Li & Mo 2015), the gray relational analysis (GRA) method (Chang *et al.* 2003; Jia *et al.* 2015), or the technique for order preference by similarity to ideal solution (TOPSIS) method (Opricovic & Tzeng 2004; Behzadian *et al.* 2012). The shortcomings of these methods include dependence on a significant amount of sample data, poor relevance, and complicated calculations, and can potentially lead to an unreasonable index weight. Therefore, new methods of OW are required to address the aforementioned problems.

Random forests (RFs) are one of the machine-learning algorithms and provide estimates regarding the importance of variables in classification (Breiman 2001). Significant theoretical and empirical studies in other fields, such as genomic ranking (Chen & Ishwaran 2012), neuroscience prediction (Smith *et al.* 2013), T-cell epitope classification (Huang *et al.* 2013), soil parent material mapping (Heung *et al.* 2014), vegetable oil analysis (Ai *et al.* 2014), and flood hazard risk assessment (Wang *et al.* 2015), have shown that RF may perform classification work effectively and quantitatively and will provide objective estimates of the important variables for classification. The quantitative estimate of variable importance is consistent with the idea of variable (index) weights, which implies that OW could, in theory, be computed via the importance of the variables. However, no study so far has focused on determining OW, utilizing RF in the field of landslide susceptibility assessment. This knowledge gap constituted the motivation for this study.

Taking the Dongjiang River basin as a study case, the main objectives of this study were: (1) to demonstrate that RF is able to calculate a reasonable OW for landslide susceptibility assessment; and (2) to construct a landslide susceptibility assessment model of basin scale, using FCE based on random forest weight (RFW). This study presents a novel methodology for the calculation of OW, and provides additional scientific references for landslide hazard analysis and risk management in the study basin.

METHODOLOGY

Fuzzy comprehensive evaluation

Dealing with the relationship between multiple indexes and susceptibility levels is a difficult aspect of susceptibility assessment as it involves the problems of multi-index and multi-level fuzzy synthetic evaluation (Jiang *et al.* 2009; Zhao *et al.* 2012; Lai *et al.* 2015). FCE is one of the fuzzy mathematics methods which is widely used for risk evaluation due to its ability of expressing and processing multi-variable and multi-level data or information (Giachetti & Young 1997a, 1997b; Feng & Luo 2009; Li 2013). Therefore, FCE was chosen to construct a landslide susceptibility assessment model.

Supposing there are n indexes in the assessment system, a domain of discourse of index variables $\boldsymbol{U}=\{\boldsymbol{u_1},\boldsymbol{u_2},\ldots,\boldsymbol{u_n}\}$ is constructed. Then, a domain of discourse of m levels, $\boldsymbol{V}=\{\boldsymbol{v_1},\boldsymbol{v_2},\ldots,\boldsymbol{v_m}\}$ is established. A fuzzy relation matrix $\boldsymbol{R}$ can be constructed on the basis of a single factor evaluation between $\boldsymbol{U}$ and $\boldsymbol{V}$. $\boldsymbol{R}$ can be expressed via Equation (1):

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & r_{ij} & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \quad (1)$$

where r_{ij} is the membership degree of index variable u_i and level v_j. Generally, it can be calculated via the following equations (Lai *et al.* 2015):

$$r_{i,1}(x_i) = \begin{cases} 1 & x_i \le V_{i,1} \\ \dfrac{V_{i,2} - x_i}{V_{i,2} - V_{i,1}} & V_{i,1} < x_i < V_{i,2} \\ 0 & x_i \ge V_{i,2} \end{cases} \quad (2)$$

$$r_{i,j}(x_i) = \begin{cases} 0 & x_i \le V_{i,j-1}, x_i \ge V_{i,j+1} \\ \dfrac{x_i - V_{i,j-1}}{V_{i,j} - V_{i,j-1}} & V_{i,j-1} < x_i \le V_{i,j} \\ \dfrac{V_{i,j+1} - x_i}{V_{i,j+1} - V_{i,j}} & V_{i,j} < x_i < V_{i,j+1} \end{cases} \quad (3)$$

$$r_{i,m}(x_i) = \begin{cases} 0 & x_i \le V_{i,m-1} \\ \dfrac{x_i - V_{i,m-1}}{V_{i,m} - V_{i,m-1}} & V_{i,m-1} < x_i < V_{i,m} \\ 1 & x_i \ge V_{i,m} \end{cases} \quad (4)$$

where x_i is the actual value of index variable u_i. $V_{i,j}$ represents the grading standard of index variable u_i and level v_j and it is determined according to the actual situation. The comprehensive membership can be computed as:

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{W} \times \boldsymbol{R} \\ &= [w_1, w_2, \ldots, w_n] \times \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & r_{ij} & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \\ &= [c_1, c_2, \ldots, c_n] \end{aligned} \quad (5)$$

where $\boldsymbol{W}$ is the index weight vector and was calculated by the RF method in this study, and $c_1, c_2, \ldots, c_n$ are the comprehensive membership degrees. Then, the final susceptibility for landslide can be computed according to the maximum membership degree law (Xue & Yang 2014). In this study, the maximum comprehensive membership degree was selected as the representative value of the susceptibility level. Five levels were determined in this study, corresponding to very low, low, medium, high, and very high susceptibility levels, respectively. For example, if c_2 were the maximum among the five comprehensive memberships, the susceptibility level would be classified as low; if c_5 were the maximum, the susceptibility level would be classified as very high.

Random forest weight

In a RF, multiple samples are drawn using the resampling bootstrap method, and classification and regression trees (CART) are built corresponding to each bootstrap sample (Breiman 2001). RF uses CARTs to classify data and to measure variable importance and the general operating principle, as shown in Figure 1. One of the advantages of RF is that it provides information on the statistical importance of each single variable within the overall result. This function helps the decision-maker to appreciate the contribution an index adds to the total risk. In the forest, approximately one-third of the data instances are not used to grow a tree, and these instances are termed the out-of-bag (OOB) data and can be used to estimate error (Yeh *et al.* 2014; Li *et al.* 2016). The OOB error rate can be obtained by the following steps (Breiman 2001). First, each case that was left out in the construction of the k^{th} tree is put down the k^{th} tree to get a classification; in this way, a test set classification can be obtained for each case for about one-third of the trees.

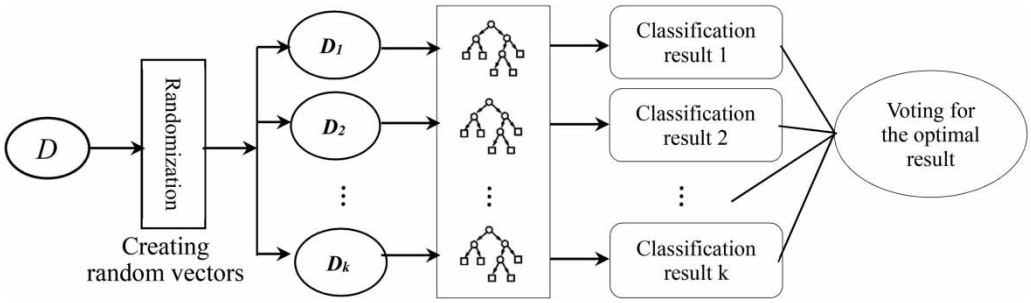

Figure 1 | Random forest (RF) operating principle.

Second, at the end of the run, j is the class that got most of the votes every time case n was OOB. Finally, the proportion of times when j is not equal to the true class of n averaged over all cases is the OOB error rate.

In RF, if there are M input variables, and a number $m << M$ is specified so that at each node m variables are selected at random from M; then, the best split of these m is used to split the node. The Gini value is used to measure the purity of the node in RF, and the bigger the Gini value, the lower the purity. The minimum Gini value is the split standard of the node, and the corresponding variable is considered the optimal variable. The Gini value can be computed as:

$$Gini(t) = 1 - \sum_{j=1}^{k} [p(j|t)]^2 \tag{6}$$

where $p(j|t)$ is the probability of class j at node t. Each time a node split is made on variable i, the Gini impurity criterion for both descendent nodes is less than that of the parent node, providing a Gini decrease after each split. Summing up the Gini decrease for each individual variable over all trees in the forest provides a fast variable importance that is often very consistent with the permutation importance measure (Breiman 2001). Generally, the mean Gini decrease (MGD) for each individual variable over all trees in the forest is frequently used as an estimate regarding the importance of variables. Thus, this study proposes the RFW as:

$$w_i = \frac{D_i}{\sum_{i=1}^{M} D_i} \tag{7}$$

where w_i and D_i are the ith variable weight and the MGD value, respectively. The RFW equation is therefore based on MGD without involving subjective factors. The RFW equation measures the importance of variables and is available to provide reasonable weights for the FCE.

Entropy weight

As a commonly used method of OW, EW is compared with RFW in this study. As a measure of disorder, the concept of entropy originates in thermodynamics and measures the heat energy that fails to generate work (Li *et al.* 2012). Entropy was first applied to information theory by Shannon (1948) and became the method of measurement of order in a system. EW is based on the information entropy theory and reflects the useful information content, offered by each variable (Jesmin & Sharif 2014; Yan *et al.* 2014; Lai *et al.* 2015).

For the calculation of EW, a judgment matrix $\boldsymbol{Y}$ with m evaluation objects and n variables was constructed:

$$\boldsymbol{Y} = (y_{ij})_{m\times n} \ (i = 1, 2, \ \ldots, n; j = 1, 2, \ \ldots, m) \tag{8}$$

To eliminate the influence of variable dimension and numerical range, $\boldsymbol{Y}$ needs to be normalized to a standard matrix $\boldsymbol{B}$ so that:

$$b_{ij} = \frac{y_{ij} - y_{min}}{y_{max} - y_{min}} \tag{9a}$$

$$b_{ij} = \frac{y_{max} - y_{ij}}{y_{max} - y_{min}} \tag{9b}$$

where y_{ij} is the actual value of the variable, and y_{min} and y_{max} are the minimum and maximum values, respectively. Equation (9a) is available for the positive variable so that a large attribute value relates to a higher susceptibility level, while Equation (9b) is available for the negative variable so that a large attribute value relates to a lower susceptibility level. According to information theory, the variable's entropy value $\boldsymbol{H_i}$ is calculated as:

$$H_i = -\frac{1}{\mathrm{Ln}\, m} \sum_{j=1}^{m} f_{ij} \, \mathrm{Ln} f_{ij} (i = 1, 2, \ \ldots, n; j = 1, 2, \ \ldots, m) \tag{10}$$

where $f_{ij} = \dfrac{b_{ij}}{\sum_{j=1}^{m} b_{ij}}$ and $0 \le H_i \le 1$. Then, the EW can be computed as:

$$w_i = \frac{1 - H_i}{n - \sum_{i=1}^{n} H_i} \tag{11}$$

The EW should meet the condition $\sum_{i=1}^{n} w_i = 1$. Apparently, a smaller entropy value relates to a larger EW, indicating that the considered variable is more important.

STUDY AREA AND DATA

Study area

The Dongjiang River basin, located in south China, has a drainage area of approximately 27,000 km^2, accounting for approximately 5.96% of the Pearl River basin (Figure 2). Six cities, Ganzhou, Heyuan, Huizhou, Dongguan, Guangzhou, and Shenzhen, are located in the basin. The basin comprises rolling country with low-relief surfaces in the downstream and mainly mountainous regions in the middle and upper reaches. Frontal and typhoon-type rainfall are predominant and annual rainfall ranges from 1,500 mm to 2,400 mm (Liu *et al.* 2010). Latosolic red soil, krasnozem, and paddy soil are the main soil types found in the basin and become soft when infiltrated by rain. The Paleozoic erathem, mainly consisting of feldspathic quartz sandstone, siltstone, schist, and shale, is well developed in the upstream of the basin, while the mid- and downstream is covered by mesozoic erathem (68.68%) mainly consisting of conglomerate and sandshale. A large number of small faults are distributed mainly in the middle and upper reaches and most of them are fractures within the bedrock, which may cause potential geological instability in the area. The basin is an economically advanced area with a dense population and valuable property. Therefore, once natural disasters occur, significant economic, ecological, and social losses can be expected in the basin (Liu & Li 2008; Wang *et al.* 2015).

The Dongjiang River basin is in a high-occurrence landslide disaster zone due to the high volume of precipitation, the soft soil, and adverse terrain (Lai *et al.* 2016a, 2016b). Examples of occurrences in the area include: a severe landslide in Zijin County as a result of continuously heavy rain in June 2007, causing two deaths; six people of Daba Town of Heping County were buried by a rainstorm-triggered landslide in May 2014; a landslide in Heping County caused by continuous heavy rain killed three villagers and injured two children in March 2016. According to incomplete statistics, thousands of landslides induced by rainfall and artificial action were reported during 2000–2010 and landslide-debris flows caused casualties and severe economic losses. For this reason, a systematic landslide susceptibility assessment of the basin is vital. The Dongjiang River basin

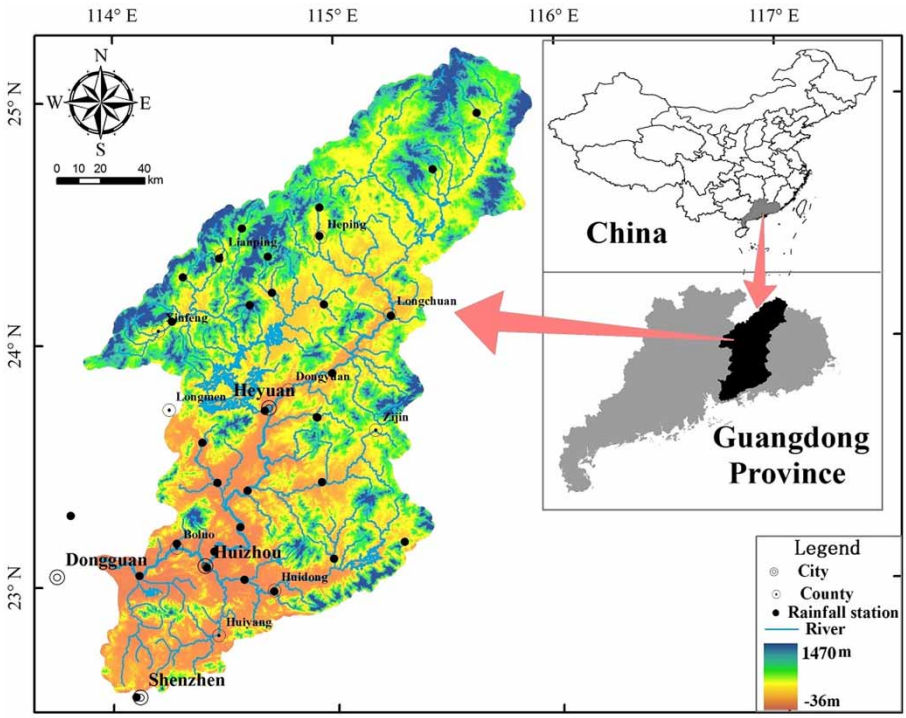

Figure 2 | Location map of the study area.

provides a typical case for the exploration of the topic of this study.

Index introduction and data sources

Index selection varies among study areas according to the specific characteristics of each location. One index can have significant impacts on landslide susceptibility in a specific area, but may have a limited influence in another area. In the landslide susceptibility assessment of this study, the proposed factors of rainfall, topography, geology, and ecology were considered (Miller *et al.* 2009; Pradhan 2011; Avtar *et al.* 2011; Jimenez-Peralvarez *et al.* 2011; Roodposhti *et al.* 2014; Lei *et al.* 2014; Su *et al.* 2015). Additionally, historical landslide events indicated that most of the landslides that had occurred in the study basin were induced by extreme rainfall, and consequently, eight indexes were finally selected to quantitatively represent regional characteristics. The eight indexes are as follows:

- Maximum 1-day precipitation (M1DP, mm): rainfall has a major influence as a landslide 'trigger' (Miller *et al.* 2009). M1DP was selected among the maximum 6 h, 12 h, 1-day, and 3-day precipitations, considering historical rainstorms that cause landslides in the basin. Thirty-four rainfall stations in the basin were used to calculate this index.
- Elevation (EL, m): the elevation of a geographic location denotes its height above or below a fixed reference point. As one of several important topographic parameters, this index was selected to measure the relative elevation of the basin. The database of EL is extracted from the digital elevation model (DEM) with a resolution of 30 m.
- Slope angle (SL, degree): similar to elevation, SL is also one of the key topographic parameters. SL is frequently applied as an index, reflecting the degree of topographic change in landslide susceptibility studies since landslides are directly related to SL. A large SL provides a large potential energy to cause earth-body sliding.
- Distance to fault (DF, m): this index is based on geological conditions. Areas near faults potentially suffer from geological instability, i.e., the closer to the faults, the more dangerous it will be.
- Soil erodibility factor (SEF): this factor provides a quantitative description of the inherent erodibility of a particular soil; it is a measure of susceptibility of soil particles to detachment and transport by rainfall and runoff. Therefore, the SEF index is based on surface soil conditions and reflects the ease with which soil can be eroded. The SEF was calculated via the erosion-productivity impact calculator (EPIC) model (Williams *et al.* 1990). A large SEF value denotes soil that is much more susceptible to erosion.
- Shear resistance capacity (SRC, MPa): this factor characterizes the capability of rock resistance to shear. Therefore, this index is also based on geological conditions and reflects the lithological character. A large SRC value shows that this type of lithology is able to withstand a considerable collapsing force.
- Runoff coefficient (RC): this index reflects the macro influence of human activity. It is based on the land-cover type of corresponding RC values (Table 1, Lai *et al.* 2016a, 2016b). A large RC value means that more rainwater is converted into surface runoff and less water can infiltrate the underground, thus greatly reducing the probability of breaking the soil structure.
- Normalized difference vegetation index (NDVI): this index represents vegetation cover. A large NDVI value means that the area has dense vegetation, providing a well-developed root system that can both maintain and stabilize soils. Areas with high vegetation cover are generally safer than bare areas.

The spatial distribution of each index is shown in Figure 3. Among the eight indexes, positive indexes (variables) include M1DP, EL, SL, and SEF. The SRC, DF, RC, and NDVI are the negative indexes (variables).

Table 1 | Land-cover type and runoff coefficient (RC)

Land-cover type	RC	Land-cover type	RC
Paddy field	0.98	Water body	1
Nonirrigated farmland	0.6	Intertidal zone	0.4
Open forest land	0.15	Mudflat	0.5
Shrubbery	0.18	Urban land	0.9
Closed forest land	0.22	Rural residential area	0.8
High coverage grassland	0.2	Construction land	0.85
Moderate coverage grassland	0.25	Sand	0.1
Low coverage grassland	0.3	Bare land	0.7

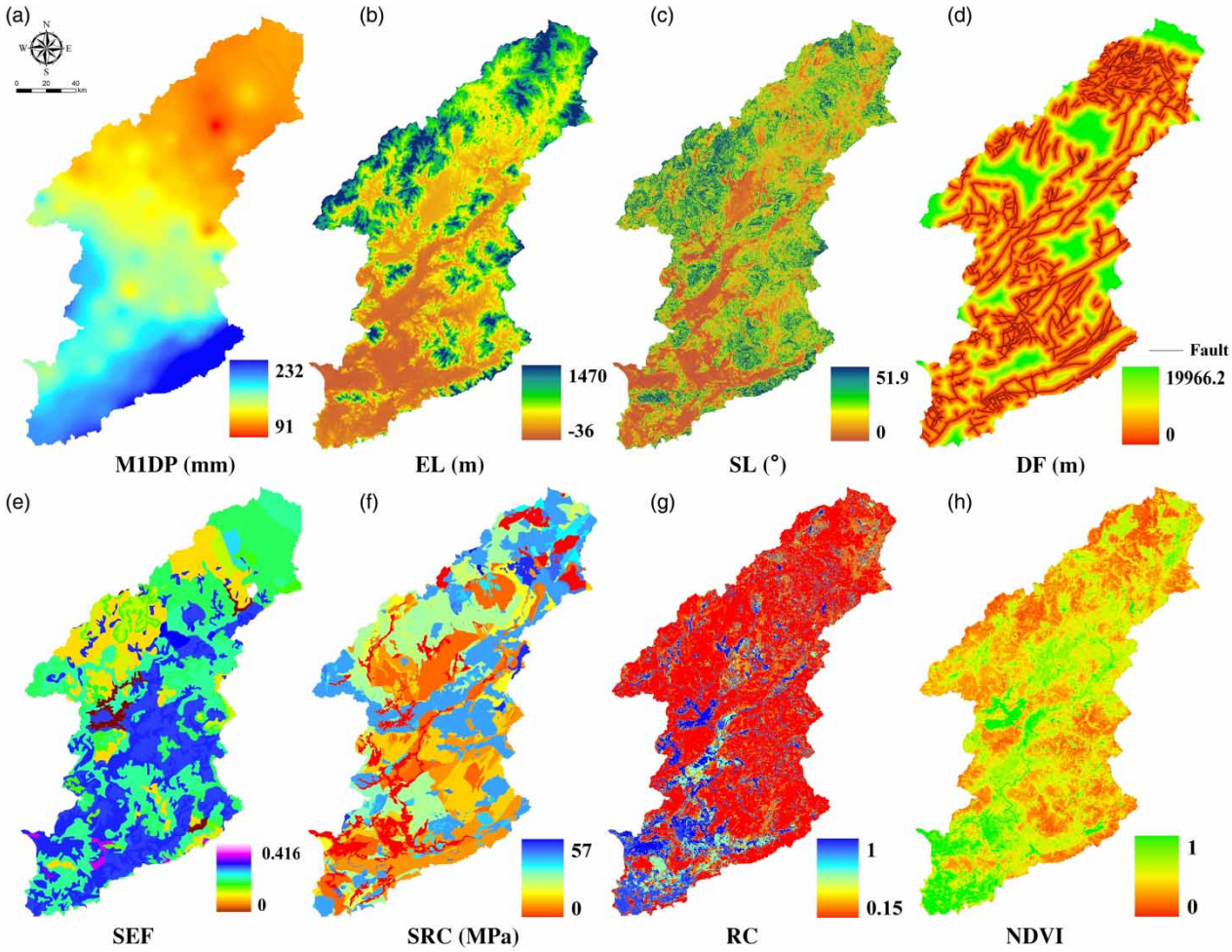

Figure 3 | Spatial distribution of landslide indexes: (a) maximum 1-day precipitation (M1DP), (b) elevation (EL), (c) slope angle (SL), (d) distance to fault (DF), (e) soil erodibility factor (SEF), (f) shear resistance capacity (SRC), (g) runoff coefficient (RC), and (h) normalized difference vegetation index (NDVI).

Data sources: Precipitation data (1960–2010) were accessed from the Hydrology Bureau of the Guangdong Province (http://www.gdsw.gov.cn/wcm/gdsw/index.html). DEM (SRTM30) was available from the CGIAR-CSI SRTM 30-m Database, which had been corrected by SRTM 90-m and local elevation data. SL data were extracted according to DEM. Soil-type data were obtained from the Food and Agriculture Organization of the United Nations (http://www.fao.org/home/en/). Land-cover type data (2010) were provided by the Resources and Environment Science Data Center of the Chinese Academy of Sciences (http://www.resdc.cn/Default.aspx). Lithology and fault data (1:250,000) were available from the National Geological Archives of China (http://www.ngac.org.cn). NDVI (30 m) was provided by the National Aeronautics and Space Administration (NASA) (http://www.nasa.gov/). All of these indexes were converted into a grid format with a cell size of 30 m × 30 m, using the GIS technique. Data-processing tools, such as the R-project, Arc.GIS 9.3, and MS Excel were used.

Landslide dataset used in the analysis

This study aims to construct a landslide susceptibility assessment model, utilizing FCE based on RFW. According to the assessment flow chart (Figure 4), a proper training and validation dataset had to be created prior to computing RFW. A database of historical landslide sites was created during

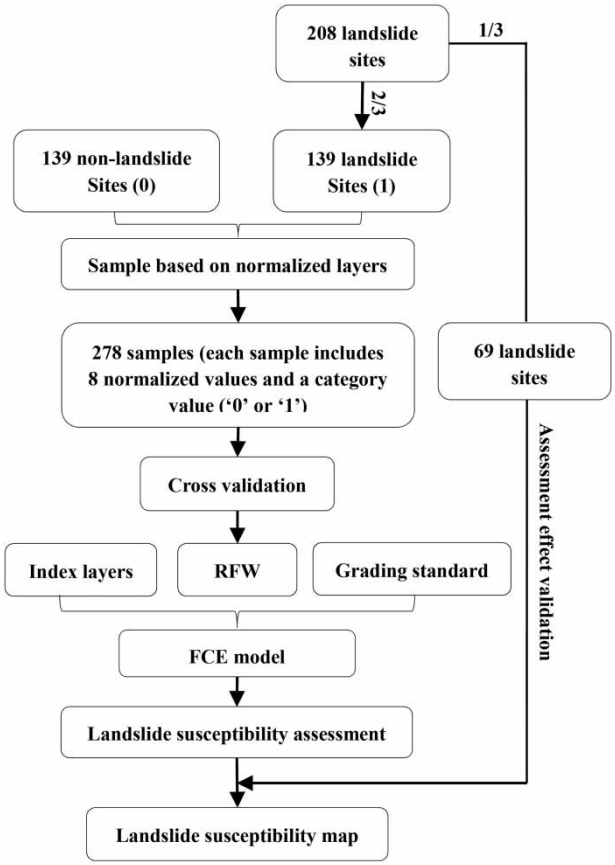

Figure 4 | Flow chart of landslide susceptibility assessment. RFW and FCE represent random forest weight and fuzzy comprehensive evaluation, respectively.

the study via field survey, air photo/satellite image interpretation, and literature and news research. Generally, only rainfall-induced landslides that occurred after extreme rainfalls were considered; landslides caused by artificial actions, including slope excavation, mine excavation, and reservoir construction were not taken into account in this study. Data for a total of 208 landslide sites were collected during this study. From these 208 landslide sites, a random sample of two-thirds (139) was used to create a training dataset, and the remaining sites (69) were used as data for validation. In other words, the remaining 69 sites were used to validate the final assessment effect, i.e., the more landslides located in high and very susceptibility level zones, the better the assessment effect will be. This study used the classification function of RF, which required at least two categories to perform the calculation (Breiman 2001). The 139 sites were classified as the first category and identified as '1′. An equal sample size to the first category (139 non-landslide sites) were randomly and uniformly drawn as the second category and identified as '0′. According to Equation (9), the eight indexes were then normalized to eight raster layers, using GIS techniques. Thereafter, samples from 278 sites (including 139 landslides sites and 139 non-landslide sites) were created by extracting the normalized values of the eight layers via the 'Sample' tool coupled in Arc.GIS 9.3. These 278 samples, including eight normalized values with a category value ('0′ or '1′), constituted a complete training dataset (Figure 5). The dataset was then included into the RF package of the R-project to train and construct the model. The common model-checking algorithm of cross-validation was used to estimate the accuracy and to reduce effects of calculation occasionality (Heung *et al.* 2014; Lai *et al.* 2016a, 2016b).

A grading standard had to be set prior to performing the FCE method. However, there is currently no unified standard for index grading, implying that the standard is determined according to the specific and actual situation of the studied basin. As shown in Table 2, critical values of the grading standard corresponding to five levels were

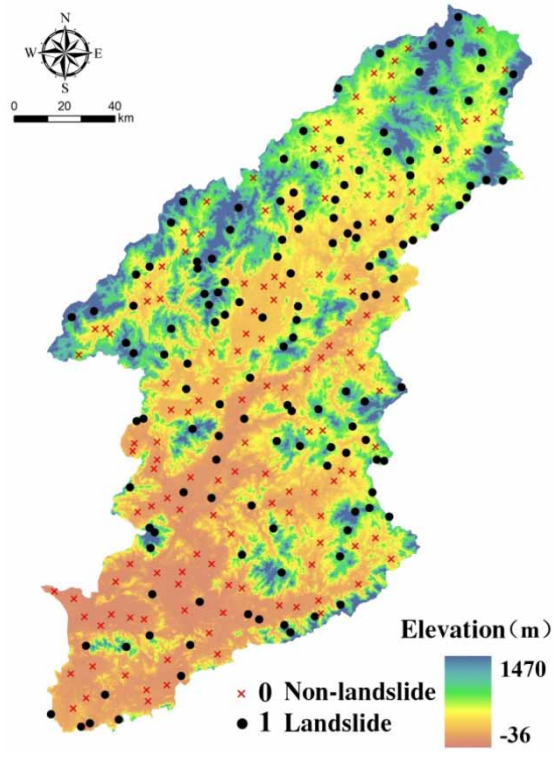

Figure 5 | Distribution of samples in the Dongjiang River basin.

Table 2 | Grading standard and index characteristic values in the Dongjiang River basin

Index	V_1	V_2	V_3	V_4	V_5	Mean	Cv
M1DP	110	121	137	151	167	137.60	0.20
EL	82	182	293	446	1470	277.87	0.77
SL	2.6	6.5	10.8	16.2	51.9	9.72	0.77
DF	546	1,404	2,574	4,602	19,966	2,797.73	1.03
SEF	0.2486	0.2584	0.2649	0.3006	0.4160	0.26	0.17
SRC	5.79	14.92	29.84	43.86	57.00	23.69	0.71
RC	0.15	0.22	0.60	0.90	1.00	0.34	0.93
NDVI	0.40	0.46	0.52	0.59	1.00	0.50	0.24

Note: Cv presents the coefficient of variation measure of the degree of uneven spatial distribution. The larger the value, the larger the spatial difference.
M1DP, maximum 1-day precipitation; EL, elevation; SL, slope angle; DF, distance to fault; SEF, soil erodibility factor; SRC, shear resistance capacity; RC, runoff coefficient (RC); NDVI, normalized difference vegetation index.

obtained via the quantile method, a typical method for classification of GIS technique where each class contains an equal number of features. Membership degrees were calculated according to Equations (2)–(4), using the critical values of the grading standard. Then, the data of eight index layers, RFW, and membership degrees were entered into the FCE method according to Equation (5). Finally, the landslide susceptibility map was generated based on GIS techniques.

RESULTS AND ANALYSIS

OOB error rate estimation

The number of classification trees k and the number of variables that have been tried at each split m were the two main parameters with significant impact on error rate. Liaw & Wiener (2002) recommended that $m = \text{sqrt}(M)$ and thus, m should be set to 3 in this study. Therefore, only the sensitivity of the parameter k was tested. The OOB error rate, a common coefficient of measuring RF's testing performance, was utilized in the sensitivity analysis. The OOB error rate varies strongly in different fields and study cases (Zhang *et al.* 2008; Ok *et al.* 2012; Huang *et al.* 2013; Heung *et al.* 2014), and a low value means superior performance and general effectivity. Five-fold cross-validation, including the training and testing process, were applied to reduce accidental error. The 278 samples in the cross-validation had classification tree settings ranging from 10 to 20,000, respectively. Five series of training and testing OOB error rates were calculated after cross-validation. Figure 6 shows average OOB error rates of both training and testing, based on the five series of error rates.

Figure 6 indicates that the training error rate decreased from 21.62% to 16.07% with a classification trees setting ranging from 10 to 20,000, but leveled off at over 200 trees. Similar to the training error rate, the testing error rate decreased from 18.55% to 13.82%, but leveled off at approximately 50 trees. Both training and testing error rates decreased with increasing number of classification trees, suggesting that both error rates can be reduced, to a certain extent, by increasing the number of classification trees. However, limitless increase of trees does not significantly contribute to increasing accuracy.

Table 3 shows the OOB error rate of a five-fold cross-validation with 3,500 classification trees. The error rate of training ranges from 14.29% to 17.41% with an average

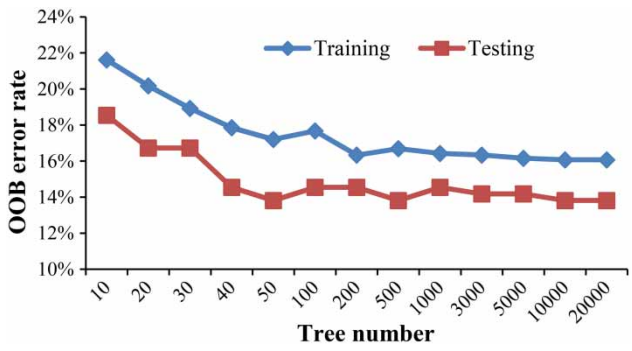

Figure 6 | Out-of-bag (OOB) error rate of training and testing.

Table 3 | Out-of-bag (OOB) error rate (%) of five-fold cross-validation

Fold	1	2	3	4	5	Average
Training	16.52	15.18	14.29	17.41	17.41	16.16
Testing	14.55	12.73	18.18	14.55	9.09	13.82

value of 16.16%, and the range of testing was 9.09%–18.18% with an average rate of 13.82%. Both these error rates are considered acceptable since each fold error rate is below 20%, providing reliability and rationality for the next step. Therefore, the number of classification trees was finally set to 3,500 for this study, considering accuracy and operation time.

Weight analysis

Five series of MGD were generated by the five-fold cross-validation. Thereafter, RFW was calculated via average MGD based on Equation (7). As shown in Figure 6 and Table 4, the EL and SL, representing topography conditions, were the top two most important of the eight indexes, occupying 46.31% of the total average MGD. NDVI was regarded as the third most important index, with a percentage of 14.07%. M1DP, DF, and RC ranked fourth, fifth, and sixth with percentages of 11.56%, 10.46%, and 6.40%, respectively. However, the indexes SEF and SRC were less consequential, with an average MGD of only 11.2% of the total.

According to Figure 3(b), many mountains are located in the upstream and eastern part of the basin. A mountain provides a sufficient sliding earth body for a landslide event and greatly increases the probability of a landslide event to occur. Figure 5 shows that most landslide sites are located in mountainous regions, confirming that the high impact index EL identified by RF plays a vital role in causing a landslide. Similar to EL, SL is considered as the second most important index by RF. A large SL provides significant potential energy to cause earth body sliding. Figure 3(c) shows that most landslide sites are located in areas with a large SL, verifying that SL also plays a significant role in landslide development. NDVI ranks third and has a similar spatial distribution as EL and SL. The remaining indexes (i.e., M1DP, DF, RC, SEF, and SRC) play relatively insignificant roles since the landslide site locations have no clear regularity correlated with the spatial distribution of these indexes. Additionally, the Dongjiang River basin has dense vegetation cover with a well-developed root system, where the vegetation and root system could greatly conserve soil and decrease top-soil erosion. In this case, the NDVI greatly offsets the impact of SEF and consequently has a larger weight than the SEF, which means that the NDVI has a larger contribution to the landslide susceptibility level than the SEF, but does not mean that the SEF is unimportant for the formation of a landslide.

The EW was calculated for comparison in this study. With normalized values of the eight indexes, the 139 sites that classified as first category and identified as '1′ were applied to calculating the EW; however, the 139 non-landslide sites that were classified as second category and identified as '0′ did not add to the calculation and, consequently, the entropy method failed to differentiate the two sample categories (i.e., '1′ and '0′). Table 4 indicates that the index weights of the two methods differ considerably. For example, the entropy method considers RC as the most vital index, while the RFW regards EL as the most important index. The SEF's EW is only 0.0049, which implies that the effects of soil type on the landslide should be negligible. However, the SEF reached 0.0571 in RFW despite ranking second from last. The other indexes, such as M1DP, DF, RC, and NDVI, also exhibited considerable differences. The main cause for the difference between

Table 4 | Index weight based on random forest (RF) and entropy method

Index	M1DP	EL	SL	DF	SEF	SRC	RC	NDVI
RFW	0.1156	0.3044	0.1587	0.1046	0.0571	0.0549	0.0640	0.1407
EW	0.1719	0.2127	0.1333	0.0152	0.0049	0.1186	0.3187	0.0247

M1DP, maximum 1-day precipitation; EL, elevation; SL, slope angle; DF, distance to fault; SEF, soil erodibility factor; SRC, shear resistance capacity; RC, runoff coefficient (RC); NDVI, normalized difference vegetation index; RFW, random forest weight; EW, entropy weight.

both methods may be that the RFW is based on historical landslide sites that can provide effective and accurate information on landslide attributes and location. The RF is a machine-learning algorithm with strong data mining ability that is able to mine the internal laws between the indexes and the susceptibility categories that include the historical landslide information; then, the RFW is generated using MGD based on these internal laws. However, the EW is only determined by one category (mark '1′) without considering available information on non-landslide sites (mark '0′), thus fails to reflect the relationship between multiple indexes and susceptibility levels.

Landslide susceptibility analysis

The RFW, shown in Figure 7, was used as the final weight. The landslide susceptibility was calculated via Equation (5) and Figure 8(a) shows the landslide susceptibility map of the Dongjiang River basin based on RFW. The zones that featured very high and high susceptibility levels are mainly located in the western and southern parts of the basin, which are mainly mountainous areas. The very low and low susceptibility level zones are distributed in the central and upstream areas of the basin and are predominantly flat. Medium susceptibility level zones are mainly located in the transition areas between high and low susceptibility level zones. From very low to very high, the zone proportions of each susceptibility level were 18.93%, 20.96%, 20.30%, 20.08%, and 19.73%, respectively. The dangerous zones, including very high and high susceptibility level zones, occupy approximately 39.89%, indicating that the landslide susceptibility situation in the Dongjiang River basin is quite significant.

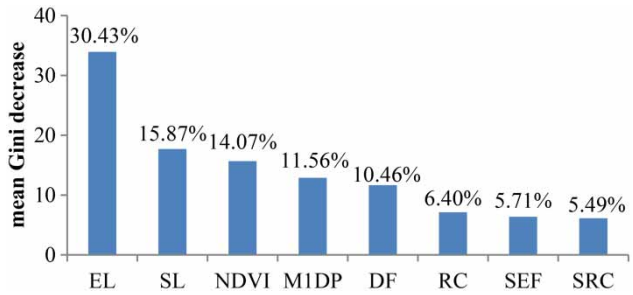

Figure 7 | Average mean Gini decrease (MGD) and random forest weight (RFW) of index. The RFW is displayed as a percentage.

To evaluate the landslide susceptibility results, other data were used to validate the reliability of high and very high susceptibility level zones. A total of 69 samples mentioned in the section 'Landslide dataset used in the analysis' (approximately one-third of the 208 landslide sites) were utilized to validate the reliability of this result. Figure 8(a) shows the verification sites randomly distributed across the basin, implying the rationality of these sites. Only the validation landslide sites located exactly in dangerous zones (high and very high susceptibility level zones) were selected to achieve estimating accuracy, meaning that the more sites in the dangerous zones, the higher the accuracy. Table 5 shows that the numbers of sites located in very high and high susceptibility level zones were 11 and 44, respectively, with a total of 55 sites located in dangerous zones. However, the numbers of landslide sites in medium, low, and very low susceptibility level zones were only 7, 2, and 5, respectively, occupying 20.29% of the total.

A landslide susceptibility map based on EW (Table 4) was also created for comparison. Figure 8(b) shows that very high susceptibility level zones are mainly located in the western and southeastern parts of the basin. Most of these areas clearly belong to mountainous areas. High susceptibility level zones include Longgang, Xinfeng, east Huizhou, and central Dongyuan. The low and very low susceptibility level zones were mainly located in the central parts of the upstream. The EW medium susceptibility level zones were similar to those of RFW and were mainly located in the transition areas between the high and low susceptibility level zones. The spatial distribution between the landslide susceptibility maps varied considerably. This variation clearly results from large differences in the index weights. The 69 landslide sites were also utilized to validate the accuracy of EW. Table 5 shows a total of 44 sites that lie within dangerous zones with a percentage of approximately 63.77%, which is lower than the result for RFW.

DISCUSSION

Figure 8 indicates that 14 sites (20.29%) that were located in the non-hazardous areas may partly be explained by data errors existing in the assessment system, including the indexes data error and validation data error. For

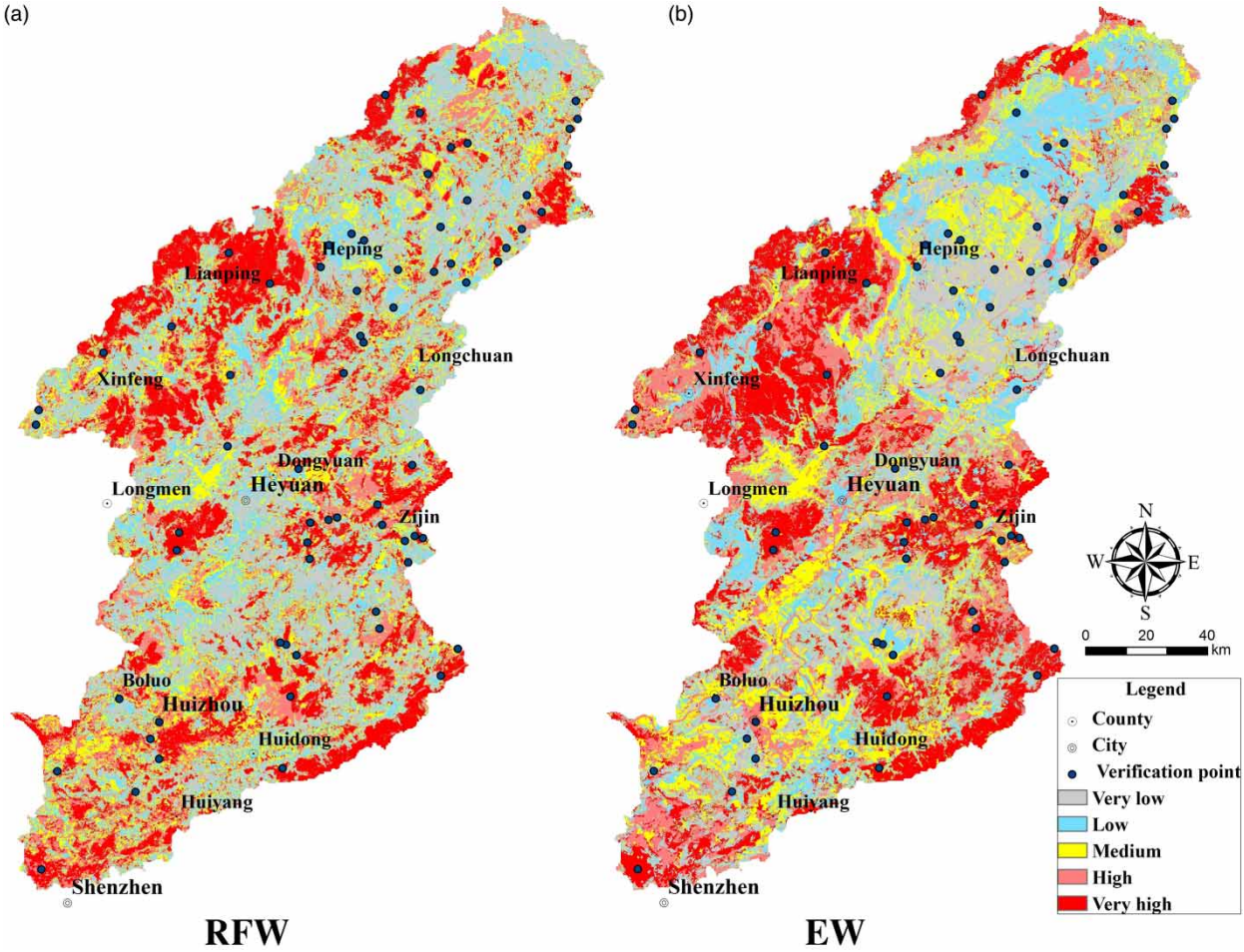

Figure 8 | Landslide susceptibility maps based on (a) the random forest weight (RFW) and (b) entropy weight (EW).

Table 5 | Validation accuracy and landslide sites amount in different susceptibility level areas

	Very high	High	Medium	Low	Very low	Dangerous	Accuracy
RFW	11	44	7	2	5	55	79.71%
EW	5	39	17	3	8	44	63.77%
RFWL	9	44	8	3	5	53	76.81%
RFWH	6	15	16	12	20	21	30.43%

Note: RFWL and RFWH represent the landslide susceptibility maps that remove the least important indexes (SEF and SRC) and the most important indexes (EL and SL), respectively; the validation site amount of 'Dangerous' includes the total of 'Very high' and 'High'.

example, the NDVI and RC were interpreted and generated based on remote-sensing imagery; however, considerable uncertainty and subjectivity exist in the process, thus significantly decreasing index precision (Liu *et al.* 2008). Moreover, the locations of historical landslide events used in this analysis were not always exact because information of accurate locations of a few landslides was incomplete and uncertain, which may also decrease the validation precision. Nevertheless, the percentage of sites located in the dangerous zones reached 79.71%, which is

a satisfactory accuracy for a landslide susceptibility assessment in general.

Figure 7 reveals that RF ranks EL and SL as the two most important indexes, while SEF and SRC are ranked the least important. Three steps were performed to provide further clarity on the rationality of RFW. The steps were as follows: (1) removing the two most important and the least important indexes, so that only six indexes remained; (2) determining the index weight again according to the proportion of MGD value shown in Figure 7; and (3) generating landslide susceptibility maps. Figure 9(a) and 9(b) show the landslide susceptibility maps that removed the least important indexes (SEF and SRC) and the most important indexes (EL and SL), respectively. Comparing Figure 8(a) and Figure 9(a) shows that some areas with medium susceptibility level are found in downstream areas. Nevertheless, there is a basic similarity in the distribution of dangerous zones in both maps. Moreover, the low and very low susceptibility level zones also display similar distribution patterns. Table 5 indicates that the accuracy reached 76.81% despite removing SEF and SRC, with a reduction of only 2.90%, suggesting that these two indexes are dispensable for the assessment system. However, the dangerous zones in Figure 9(b) differ greatly from those of Figure 8(a) since the most important indexes (EL and SL) have been removed. The result with the very high and high susceptibility level in the midstream areas, such as Heyuan, Boluo, Longmen, central Zijin, and central Huidong, is obviously inaccurate since most of these zones belong to plain areas. The accuracy in Figure 9(b) is

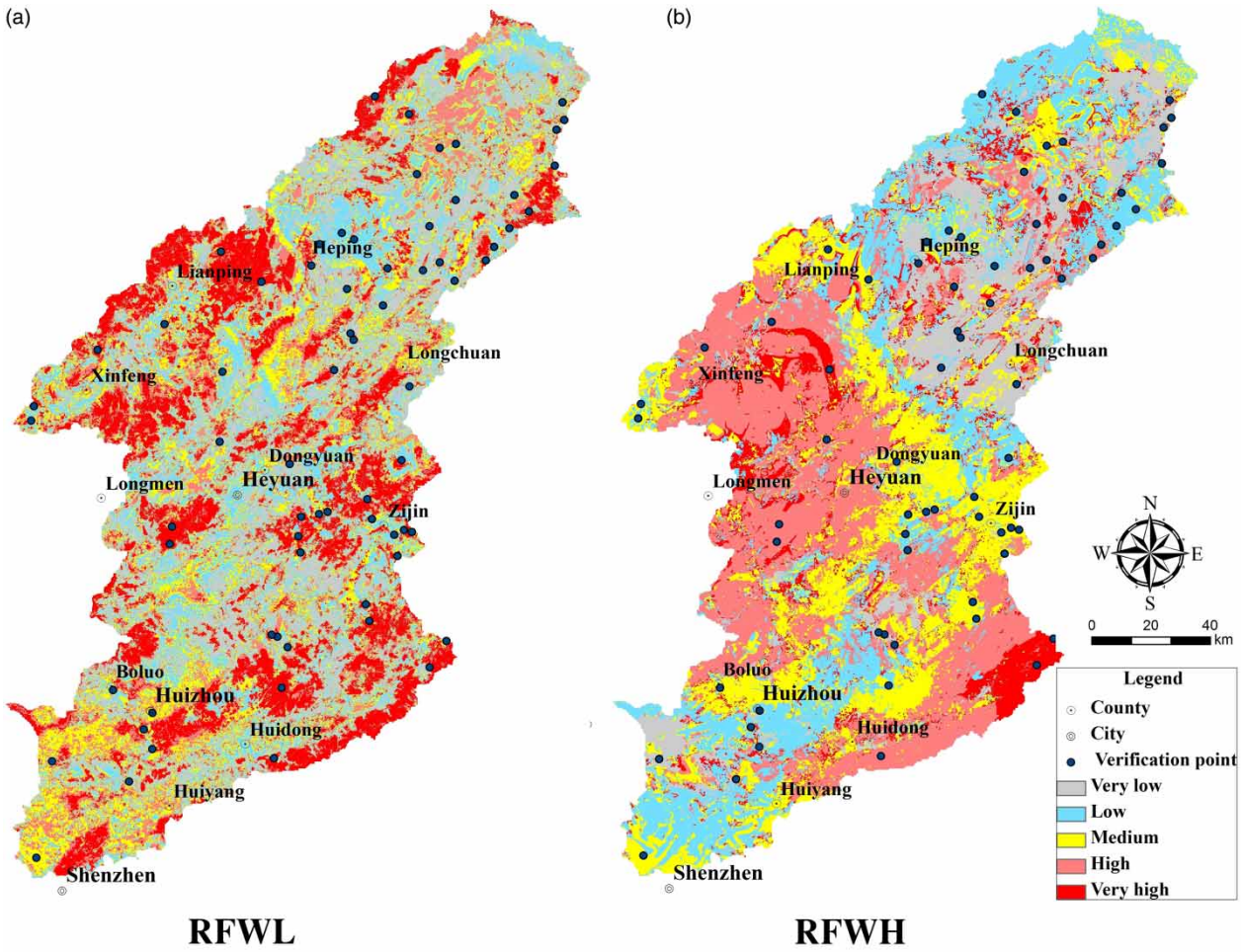

Figure 9 | Landslide susceptibility maps (six indexes) based on the RFW: (a) RFWL removing the least important indexes (SEF and SRC) and (b) RFWH removing the most important indexes (EL and SL).

merely 30.43%, which is far lower than that of Figure 8(a), suggesting that the EL and SL are indispensable and play a significant role in the assessment system. The RFW proposed by this study can be regarded as reasonable and feasible, and evidence has been provided and substantiated via precision validation and contrastive analysis. Additionally, the RFW also provides a clue for using high-resolution data, for example, RF ranking EL as the most important indexes means that the errors in EL might strongly affect the assessment results. In this case, the EL layer requires high-resolution data to obtain more accurate results (e.g., this research used the modified SRTM 30-m data instead of the SRTM 90-m data).

RF is an efficient and easy to operate learning machine and was used in this study to determine index weights. This is a novel approach for landslide susceptibility assessments. However, certain issues still remain. For example, the landslide susceptibility and index weights were assessed by the statistic method; however, they were not strongly involved in the formation mechanism, which may lead to difficulties to reflect the physical process of landslide. As far as the evaluation system is concerned, rainfall-induced landslide susceptibility was considered at basin scale and only eight indexes were selected to construct the index system. This neglects the influence of seismic factors, river degradation, and groundwater movement due to data restrictions. However, these factors also play an important role in the formation and development of landslides. A more comprehensive index system needs to be developed in future studies to account for these as well as for other missing factors. Also, the RFW of this study requires a large number of historical landslide sites. Increasing the number of sites would significantly improve the accuracy of the results. Merely 208 sites were used in this study due to data limitations. This study proved that, despite a few drawbacks, the application of RF to weight determination shows significant potential. Thus, RFW is highly recommended for landslide susceptibility assessments.

CONCLUSIONS

Landslide susceptibility assessment is an appropriate tool for analyzing and predicting the spatial distribution of rainfall-induced landslides. However, the determination of suitable index weights is a critical step that significantly influences the assessment results. To solve this problem, this study proposed a new weight determining method based on RF. Using the Dongjiang River basin for a case study, eight indexes were selected to construct a susceptibility index system. Using OW computed via RF, the landslide susceptibility was evaluated by the FCE method. The study concludes that:

1. RF ranks EL and SL as the two most important of the eight indexes, occupying 46.31% of the total MGD. NDVI was regarded as the third most important index, with a percentage of 14.07%. M1DP, DF, and RC ranked fourth, fifth, and sixth with percentages of 11.56%, 10.46%, and 6.40%, respectively. The indexes SEF and SRC are less consequential, with an MGD of only 11.2% of the total.
2. The accuracy based on RFW reached up to 76.81%, while it only reached 66.67% based on EW. Both the precision validation and contrastive analysis demonstrated that the assessment results of RFW were more reasonable and satisfactory than those of EW. Generally, the landslide susceptibility situation in the Dongjiang River basin was quite significant. Therefore, preventative actions, including both engineering and non-engineering measures, should be conducted in the dangerous zones to predict and prevent landslides, and to reduce losses of capital and human life as much as possible.
3. Two experiments removed the two most important and the least important indexes, respectively, providing further evidence that the index weight calculated via RF was reasonable and feasible for landslide susceptibility assessment. The initial application of the RF to weight determination in this study showed significant potential despite a number of drawbacks. Accordingly, the index weight calculated via RF was more consistent with scientific principles and can therefore be recommended for use in this field.

ACKNOWLEDGEMENT

The research was financially supported by the National Natural Science Foundation of China (Grant No. 91547202, 51479216, 51579105, 51210013), the China Postdoctoral Science Foundation (2017M612662), the

Chinese Academy of Engineering Consulting Project (2015-ZD-07-04-03), the Public Welfare Project of Ministry of Water Resources (Grant No. 200901043-03), the Project for Creative Research from Guangdong Water Resources Department (Grant No. 2016-07, 2016-01), and Research program of Guangzhou Water Authority (2017).

REFERENCES

Ai, F. F., Bin, J. & Zhang, Z. M. 2014 Application of random forests to select premium quality vegetable oils by their fatty acid composition. *Food Chem.* **143**, 472–478.

Avtar, R., Singh, C. K., Singh, G., Verma, R. L., Mukherjee, R. S. & Sawada, H. 2011 Landslide susceptibility zonation study using remote sensing and GIS technology in the Ken-Betwa River Link area, India. *Bull. Eng. Geol. Environ.* **70**, 595–606.

Balteanu, D., Chendes, V., Sima, M. & Enciu, P. 2010 A country-wide spatial assessment of landslide susceptibility in Romania. *Geomorphology* **124** (3–4), 102–112.

Behzadian, M., Otaghsara, S. K., Yazdani, M. & Ignatius, J. 2012 A state-of the-art survey of TOPSIS applications. *Expert Syst. Appl.* **39**, 13051–13069.

Breiman, L. 2001 Random forests. *Mach. Learn.* **45** (1), 5–32.

Chang, C. L., Tsai, C. H. & Chen, L. 2003 Applying grey relational analysis to the decathlon evaluation model. *Int. J. Comput. Internet Manage.* **11**, 54–62.

Chen, X. & Ishwaran, H. 2012 Random forests for genomic data analysis. *Genomics* **99** (6), 323–329.

Christian, C., Marilena, C., Nathalie, A., Caraballo, A., Alvaro, G. G., Edoardo, R. & Valerio, A. 2015 Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: a case of the Belice River basin (Western Sicily, Italy). *Geomorphology* **242**, 49–64.

Diakoulaki, D., Mavrotas, G. & Papayannakis, L. 1995 Determining objective weights in multiple criteria problems: the critic method. *Comput. Oper. Res.* **22** (7), 763–770.

EM-DAT. Disaster Profiles 2016 The OFDA/CRED international disaster database. http://www.emdat.be/database (accessed 19 December 2016).

Feng, L. H. & Luo, G. Y. 2009 Practical study on the fuzzy risk of flood disasters. *Acta. Appl. Math.* **106**, 421–432.

Giachetti, R. E. & Young, R. E. 1997a Analysis of the error in the standard approximation used for multiplication of triangular and trapezoidal fuzzy numbers and the development of a new approximation. *Fuzzy Sets Syst.* **91** (1), 1–13.

Giachetti, R. E. & Young, R. E. 1997b A parametric representation of fuzzy numbers and their arithmetic operators. *Fuzzy Sets Syst.* **91** (2), 185–202.

Heung, B., Bulmer, C. E. & Schmidt, M. G. 2014 Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma* **214–215**, 141–154.

Huang, J. H., Xie, H. L., Yan, J., Lu, H. M., Xu, Q. S. & Liang, Y. Z. 2013 Using random forest to classify T-cell epitopes based on amino acid properties and molecular features. *Anal. Chim. Acta* **804**, 70–75.

Jesmin, F. K. & Sharif, M. B. 2014 Weighted entropy for segmentation evaluation. *Opt. Laser. Technol.* **57**, 236–242.

Jia, X. L., Li, C. H., Cai, Y. P., Wang, X. & Lian, S. 2015 An improved method for integrated water security assessment in the Yellow River basin, China. *Stoch. Environ. Res. Risk Assess.* **29**, 2213–2227.

Jiang, H. & Eastman, J. R. 2000 Application of fuzzy measures in multi-criteria evaluation in GIS. *Int. J. Geogr. Inform. Sci.* **14** (2), 173–184.

Jiang, W. G., Deng, L., Chen, L. Y., Wu, J. & Li, J. 2009 Risk assessment and validation of flood disaster based on fuzzy mathematics. *Prog. Natural Sci.* **19** (10), 1419–1425.

Jimenez-Peralvarez, J. D., Irigaray, C., El Hamdouni, R. & Chacon, J. 2011 Landslide-susceptibility mapping in a semi-arid mountain environment: an example from the southern slopes of Sierra Nevada (Granada, Spain). *Bull. Eng. Geol. Environ.* **70**, 265–277.

Lai, C. G., Chen, X. H., Chen, X. Y., Wang, Z. L., Wu, X. S. & Zhao, S. W. 2015 A fuzzy comprehensive evaluation model for flood risk based on the combination weight of game theory. *Nat. Hazards* **77**, 1243–1259.

Lai, C. G., Shao, Q. X., Chen, X. H., Wang, Z. L., Zhou, X. W., Yang, B. & Zhang, L. L. 2016a Flood risk zoning using a rule mining based on ant colony algorithm. *J. Hydrol.* **542**, 268–280.

Lai, C. G., Wang, Z. L., Chen, X. H., Xu, C.-Y., Yang, B., Meng, Q. Q. & Huang, B. 2016b A procedure for assessing the impacts of land-cover change on soil erosion at basin scale. *Hydrol. Res.* **47** (5), 903–918.

Lee, S. & Evangelista, D. G. 2006 Earthquake-induced landslide-susceptibility mapping using an artificial neural network. *Nat. Hazards Earth Syst. Sci.* **6**, 687–695.

Lee, S. & Pradhan, B. 2007 Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* **4**, 33–41.

Lei, T. C., Huang, Y. M., Lee, B. J., Hsieh, M. H. & Lin, K. T. 2014 Development of an empirical model for rainfall-induced hillside vulnerability assessment: a case study on Chen-Yu-Lan watershed, Nantou, Taiwan. *Nat. Hazards* **74**, 341–373.

Li, Q. 2013 Fuzzy approach to analysis of flood risk based on variable fuzzy sets and improved information diffusion methods. *Nat. Hazards Earth Syst. Sci.* **13**, 239–249.

Li, L. H. & Mo, R. 2015 Production task queue optimization based on multi-attribute evaluation for complex product assembly workshop. *Plos One* **10** (9), e0134343.

Li, X. G., Wei, X. & Huang, Q. 2012 Comprehensive entropy weight observability–controllability risk analysis and its application to water resource decision-making. *Water Res.* **38** (4), 573–579.

Li, B., Yang, G. S., Wan, R. R., Dai, X. & Zhang, Y. H. 2016 Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. *Hydrol. Res.* **47**, 69–83.

Liaw, A. & Wiener, M. 2002 Classification and regression by random Forest. *R News* **2**, 18–22.

Liu, X. P. & Li, X. 2008 Simulating complex urban development using kernel-based non-linear cellular automata. *Ecol. Modell.* **211** (1–2), 169–181.

Liu, X. P., Li, X., Liu, L. & He, J. Q. 2008 A bottom-up approach to discover transition rules of cellular automata using ant intelligence. *Int. J. Geogr. Inform. Sci.* **22** (11–12), 1247–1269.

Liu, D. E., Chen, C. H., Lian, Y. Q. & Lou, Z. H. 2010 Impacts of climate change and human activities on surface runoff in the Dongjiang River Basin of China. *Hydrol. Process.* **24** (11), 1487–1495.

Mckean, J. & Roering, J. 2001 Objective landslide detection and surface morphology mapping using high-resolution airborne laser altimetry. *Geomorphology* **57** (3–4), 331–351.

Miller, S., Brewer, T. & Harris, N. 2009 Rainfall thresholding and susceptibility assessment of rainfall-induced landslides: application to landslide management in St Thomas, Jamaica. *Bull. Eng. Geol. Environ.* **68**, 539–550.

Ok, A. O., Akar, O. & Gungor, O. 2012 Evaluation of random forest method for agricultural crop classification. *Eur. J. Remote Sens.* **45**, 421–432.

Opricovic, S. & Tzeng, G. H. 2004 Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **156**, 445–455.

Pistocchi, A., Luzi, L. & Napolitano, P. 2002 The use of predictive modeling techniques for optimal exploitation of spatial databases: a case study in landslide hazard mapping with expert system-like methods. *Environ. Geol.* **41** (7), 765–775.

Pradhan, B. 2011 Manifestation of an advanced fuzzy logic model coupled with Geo-information techniques to landslide susceptibility mapping and their comparison with logistic regression modelling. *Environ. Ecol. Stat.* **18**, 471–493.

Roodposhti, M. S., Rahimi, S. & Beglou, M. J. 2014 PROMETHEE II and fuzzy AHP: an enhanced GIS-based landslide susceptibility mapping. *Nat. Hazards* **73**, 77–95.

Shannon, C. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **5** (1), 3–53.

Smith, P. F., Ganesh, S. & Liu, P. 2013 A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *J. Neurosci. Methods* **220**, 85–91.

Su, C., Wang, L. L., Wang, X. Z., Huang, Z. C. & Zhang, X. C. 2015 Mapping of rainfall-induced landslide susceptibility in Wencheng, China, using support vector machine. *Nat. Hazards* **76**, 1759–1779.

Wang, Z. L., Lai, C. G., Chen, X. H., Yang, B., Zhao, S. W. & Bai, X. Y. 2015 Flood hazard risk assessment model based on random forest. *J. Hydrol.* **527**, 1130–1141.

Williams, J. R., Dyke, P. T. & Fuchs, W. W. 1990 EPIC: erosion productivity impact calculator. In: *Users Manual: EPIC: Erosion Productivity Impact Calculator, Model Documentation, USDA–ARS Tech. Bull. No. 1768, USDA–ARS Grassland* (A. N. Sharpley & J. R. Williams, eds), Soil and Water Research Laboratory, Temple, TX, p. 127.

Xue, X. H. & Yang, X. G. 2014 Seismic liquefaction potential assessed by fuzzy comprehensive evaluation method. *Nat. Hazards* **71**, 2101–2112.

Yan, J. H., Feng, C. H. & Li, L. 2014 Sustainability assessment of machining process based on extension theory and entropy weight approach. *Int. J. Adv. Manuf. Technol.* **71**, 1419–1431.

Yang, X. J., Hou, D. G., Hao, Z. L. & Wang, E. Y. 2016 Fuzzy comprehensive evaluation of landslide caused by underground mining subsidence and its monitoring. *Int. J. Environ. Pollut.* **59**, 284–302.

Yeh, C. C., Chi, D. J. & Lin, Y. R. 2014 Going-concern prediction using hybrid random forests and rough set approach. *Inform. Sci.* **254**, 98–110.

Zhang, J., Zulkernine, M. & Haque, A. 2008 Random-forests-based network intrusion detection systems. *IEEE T. Syst. Man. Cy. C.* **38**, 649–659.

Zhao, J., Jin, J. L., Zhang, X. M. & Chen, Y. Q. 2012 Dynamic risk assessment model for water quality on projection pursuit cluster. *Hydrol. Res.* **43** (6), 798–807.

Zou, Q., Zhou, J. Z., Zhou, C., Song, L. X. & Guo, J. 2013 Comprehensive flood risk assessment based on set pair analysis variable fuzzy sets model and fuzzy AHP. *Stoch. Environ. Res. Risk Assess.* **27**, 525–546.

First received 11 March 2017; accepted in revised form 6 June 2017. Available online 27 July 2017