A Bayesian Framework Based on Gaussian Mixture Model and Radial Basis Function Fisher Discriminant Analysis (BayGmmKda V1.1) for Spatial Prediction of Flood

Dieu, Tien Bui[1] and Nhat-Duc, Hoang[2]

[1]Geographic Information System Group, Department of Business Administration and IT, University College of Southeast Norway (USN), Gullbringvegen 36, N-3800, Bø i Telemark, Norway

[2]Faculty of Civil Engineering, Institute of Research and Development, Duy Tan University, P809 - K7/25 Quang Trung, Danang, Vietnam

Correspondence to: Nhat-Duc Hoang (hoangnhatduc@dtu.edu.vn)

Abstract. In this study, a probabilistic model, named as BayGmmKda, is proposed for flood susceptibility assessment with a study area in Central Vietnam. The new model is a Bayesian framework constructed by a combination of Gaussian Mixture Model (GMM), Radial Basis Function Fisher Discriminant Analysis (RBFDA), and a Geographic Information System (GIS) database. In the Bayesian framework, GMM is used for modeling the data distribution of flood influencing factors in the GIS database, whereas RBFDA is utilized to construct a latent variable aiming to enhance the model performance. As a result, the posterior probability of flood, which is the output of the BayGmmKda model, is used as flood susceptibility index. Experiment results showed that the proposed hybrid framework is superior to other benchmark models including the Adaptive Neuro Fuzzy Inference System and the Support Vector Machine. To facilitate the model implementation, a software program of BayGmmKda has been developed in Matlab. The BayGmmKda program can accurately establish a flood susceptibility map for the study region. Accordingly, local authorities can overlay this susceptibility map onto various land-use maps for the purpose of land-use planning or management.

Key words: Flood Evaluation; Bayesian Classifier; Gaussian Mixture Model; Discriminant Analysis; Latent Variable.

1 Introduction

Flood stands out as one of the most destructive natural hazards that destroys crops, infrastructures, and cause heavy loss of human lives in immense spatial extent (Dottori et al., 2016). Recent statistics on flood damages for the period 1995–2015 shows that flood affects to 109 million people around the global per year (Alfieri et al., 2017), in which, more than 220,000 people lost their lives (Winsemius et al., 2015). Although frequency of flood decreases in several regions i.e. Central Asia and America, flood occurrence is globally increased by 42% (Hirabayashi et al., 2013).

Among regions in the world, Southeast Asia is one of the most heavily flood damaged regions in the world due to monsoonal rainfalls and tropical hurricane patterns (Loo et al., 2015). Particularly, Vietnam is a storm center at the Western Pacific and this nation has faced the destructive consequence of flooding in many of its provinces. In Vietnam, floods are often triggered by tropical cyclones. More than 71% of the Vietnam's population and 59% of the total land area of Vietnam are susceptible to the impacts of these natural hazards (Tien Bui et al., 2016c). Based on a report done by Kreft et al. (2014), from 1994 to 2013, Vietnam endured an annual economic loss that is equivalent to $2.9 billion.

Additionally, the occurrences of flood in Vietnam are expected to rise rapidly in the near future due to the increases of poorly planned infrastructure developments and urbanization near watercourses, as well as an increased activity of deforestation and climate change. Hence, an accurate model of flood forecasting becomes a crucial need for land-use planning as well as establishment of disaster mitigation strategies. Based on flood prediction models, flood-prone area can be identified and mapped (Tien Bui et al., 2016c).

Needless to say, the identification of susceptible areas can significantly reduce damages of flood to the national economy and human lives by avoiding infrastructure developments and densely populated settlements in highly flood susceptible areas (Zhou et al., 2016). The prediction outcomes also help Government agency to issue appropriate flood management policies and to focus its limited financial resource to construct large-scale flood defense infrastructure in areas that feature great economic values but are highly susceptible to flood (Bubeck et al., 2012). Therefore, a tool of flood spatial modeling is of great usefulness.

To predict flood occurrence, conventional approaches require time series of meteorological and streamflow data at gauging stations (Machado et al., 2015). However, this is difficult for many areas in developing countries where no gauging station is available. Therefore, new modeling approaches should be explored and investigated. Given these motivations, this study proposes a novel methodology designed for enhancing the prediction accuracy as well as deriving probabilistic evaluation of flood susceptibility in a regional scale. Accordingly, spatial prediction of flood is carried out based on a statistical assumption that flood in the future will occur under the same conditions that triggered them in the past (Tien Bui et al., 2016b). In this way, the flood prediction problem boils down to an on-off supervised classification task, where the flood inventories are used as a flood class, whereas a non-flood class is derived from areas that have not yet damaged by flood. Consequently, spatial prediction of flood is derived based on probability of study area pixels belonging to the flood class. To yield probabilistic outputs of flood, this study proposes, for the first time, a Bayesian framework established based on a

Gaussian mixture model (GMM) and the Kernel Fisher Discriminant Analysis (KFDA). GMM is employed for density approximation to calculate the posterior probability of flood (flood susceptibility index), whereas KFDA is employed to construct a latent variable for from the geo-environmental conditions to enhance the performance of the Bayesian model.

In essence, the proposed integrated framework contains two phases of analysis. RBFDA is first employed for latent variable construction. The Bayesian approach assisted by GMM is then used to perform probabilistic pattern recognition. The first level performs pattern discriminant analysis task and the second level carries out the prediction to derive the model output of flood evaluation. Based on previous studies which indicate that hierarchical model structures can produce improving prediction accuracy, the proposed framework can potentially bring about desirable flood assessment results. The subsequent parts of this study are organized in the following order: Related works on flood prediction are summarized in the second section. The next section introduces the research method of the current paper, followed by the fourth part that describes the proposed Bayesian model for flood susceptibility forecasting. The next part reports the model prediction accuracy and comparison. The last section discusses some conclusions on this work.

2. A Review of Related Works on Flood Susceptibility Prediction

Because of the criticality of flood prediction, this problem has gained an increasing attention from the academic community. Following this trend, various flood analyzing tools have been developed, ranging from relatively simple methods to more sophisticated methodologies (Winsemius et al., 2013;Papaioannou et al., 2015;Gao et al., 2017). Basically, these tools could be classified into statistical analysis, rainfall-runoff models, and classification models. Statistical analysis uses long-term recorded time series data at gauged stations to establish regression models, and then, the models are used to transform flood information to ungauged basins (Yue et al., 1999;Cunnane, 1988;McCuen, 2016). Thus, these models are capable to provide flood predictions both in space and time. However, long-term data are not always available, and in many cases, they are general too short for reliable estimating of extreme quantiles (Seckin et al., 2013b;Nguyen et al., 2014).

Rainfall–runoff models, which deal with estimating of runoff from rainfall, are considered to be the most extensively used for flood prediction and management (Nayak et al., 2013;Ciabatta et al., 2016;Bennett et al., 2016). Various types of rainfall–runoff models can be found in literature, varying from empirical models to highly sophisticated physical processes. Empirical models could be established based on statistical techniques (Brocca et al., 2011) or advanced machine learning algorithms (Lohani et al., 2011) to model rainfall and runoff using historical time series data. In addition, physical processes models focus on simulating hydrological processes in a basin based on a set of mathematical equations governing physical processes of water flow and surfaces (Chiew et al., 1993;Birkel et al., 2010;Arnold et al., 1998;Beven et al., 1984;Grimaldi et al., 2013). In general, rainfall–runoff models require relative long term time series data at gauging stations. However, the density of gauging stations in developing countries is very low and this fact imposes a great obstacle for establishing accurate hydrological models (Fenicia et al., 2008). In addition, large-scale field works and deployments of measuring equipment are necessary for collecting data. Nevertheless, the complex and nonlinear nature of the flood modeling problem also bring about difficulties for hydrological methods and techniques (Sahoo et al., 2006).

In recent year, a new flood modeling approach called “on-off” classification of flood occurrence has been successfully proposed for spatial prediction of flood, also called flood susceptibility (Tien Bui et al., 2016d;Tehrany et al., 2014;Tehrany et al., 2015b). Accordingly, no time series data is required for the model calibration and the establishment of flood models is based on flood inventories (flood class) and non-flood areas (non-flood class). Accordingly, the probability of a pixel in the study area belongs to the flood class is used as flood susceptibility index. Although flood susceptibility map provides no temporal prediction or return period of flood, the flood map is capable delineating highly susceptible areas. Thus, it is a powerful flood analysis tool for decision-makers that could be used in landuse planning and flood management. Literature review shows that data-driven methods integrated with GIS databases have demonstrated their effectiveness and accuracy in large scaled flood susceptible predictions. An fuzzy logic based algorithm has been used to develop a map of flooded areas from synthetic aperture radar imagery, used for the operational flood management system in Italia, was established by Pulvirenti et al. (2011). A model based on the frequency ratio approach and GIS for spatial prediction of flooded regions was first introduced by Lee et al. (2012); the spatial database were constructed by field surveys and maps of the topography, geology, land cover, and infrastructure.

Prediction models with artificial neural network (ANN) have been employed for flood susceptibility evaluation by various scholars (Kia et al., 2012;Seckin et al., 2013a;Rezaeianzadeh et al., 2014;Radmehr and Araghinejad, 2014); previous works have shown ANN as a capable nonlinear modeling tool. Nevertheless, ANN learning is prone to overfitting and its performance has been shown to be inferior to that of support vector machine (Hoang and Pham, 2016).

Kazakis et al. (2015) introduced a multi-criteria index to assess flood hazard areas that relies on GIS and Analytical Hierarchy Process (AHP); in this methodology, the relative importance of each flood influencing factors for the occurrence and severity of flood were determined via AHP. More recently, Support Vector Machine-based flood susceptibility analysis approaches have been proposed by Tehrany et al. (2015a) and Tehrany et al. (2015b); the research finding is that SVM is more accurate than other benchmark models including the decision tree classifier and the conventional frequency ratio model.

Mukerji et al. (2009) constructed flood forecasting models based on an adaptive neuro-fuzzy interference system (ANFIS), Genetic Algorithm optimized ANFIS; experiments demonstrated that ANFIS attained the most desirable accuracy. Recently, a metaheuristic optimized neural fuzzy inference system, named as MONF, has been introduced by Tien Bui et al. (2016c); the research finding is that MONF is more capable than decision tree, ANN, SVM, and conventional ANFIS.

As can be seen from the literature review, various data-driven and advanced soft computing approaches have been proposed to construct different flood forecasting models. In most previous studies, the flood prediction was formulated as a binary pattern recognition problem in which the model output is either flood or no flood. Probabilistic models have rarely been examined to cope with the complexity as well as uncertainty of the problem under concern. Therefore, our research aims at enriching the body of knowledge by proposing a novel Bayesian probabilistic model to estimate the flood vulnerability with the use of a GIS database.

3 Research Method

3.1 Flood inventory map and flood influencing factors of the study area

3.1.1 The study area

In this research, Tuong Duong district (central Vietnam) is selected as the study area (see **Figure 1**). This is by far a heavily affected flood region in the country (Reynaud and Nguyen, 2016). The area of the district is approximately 2803 km^2 and locates between the longitudes of 18°58'42"N and 19°39'16"N, and between the latitudes of 104°15'58"E and 104°55'57"E. The topographical feature of the Tuong Duong district is inherently complex with mountainous areas, watersheds, and rivers. Drastic floods often divided the district into several isolated areas which are very difficult to approach for rescuing or evacuation purposes.

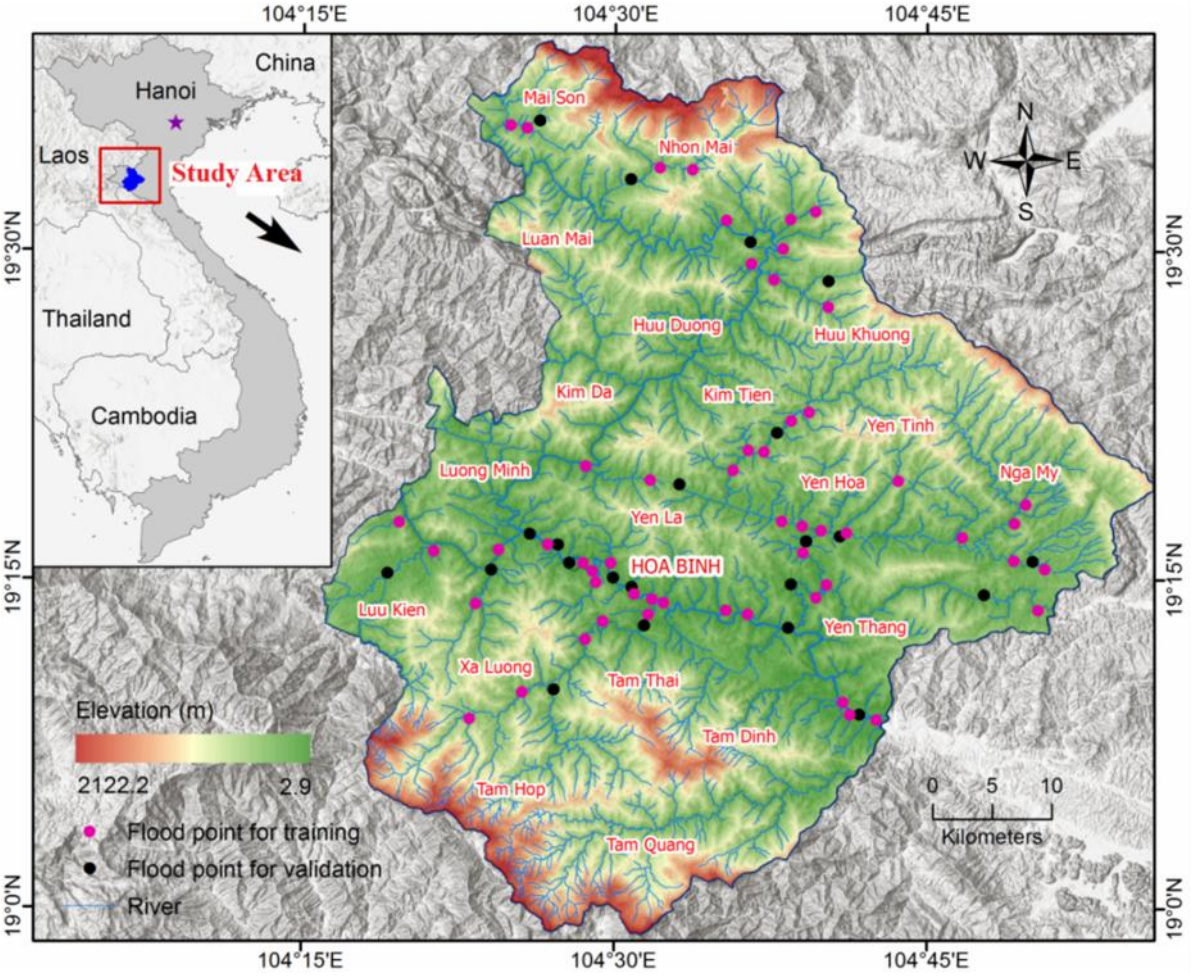


Figure 1 Location of the Tuong Duong district (Central Vietnam)

The district has two separated seasons, namely a cold season (from November to March) and a hot season (from April to October). The yearly rainfall of the district is within the range of 1679 mm and 3259 mm. The rainfall amount is primarily intensified during the rainy period which contributes to roughly 90% of the total annual rainfall. Due to the district's location as well as its topographic and climatic features, the study area is highly susceptible to flood events with immense infliction to human casualty and economic value. An examination carried out by Reynaud and Nguyen (2016) reported that approximately 40% of families have been damaged by floods and roughly 20% of families must be relocated from the flooded areas; the average loss of flood goes up to 24% of the family income each year.

3.1.2 Flood inventory map

Prediction of flood zones can be based on an assumption that future flood events are governed by the very similar conditions of flooded zones in the past. Therefore, flood inventories and their geo-environmental conditions (e.g. topological and hydrological features) produced them must be extensively determined and collected (Tien Bui et al., 2016c;Tehrany et al., 2015b). The first step of this analysis is to establish a flood inventory map for the region under investigation. In this study, the flood inventory map established by Tien Bui et al. (2016c) was used to analyze the relationships between flood occurrences and influencing factors.

The flood inventory map stores documentations of past flood events (see **Figure 1**). It is noted that the type of floods in this study area is flash flood. This is the main flood type in this region due to characteristics of the terrain. The map was constructed by gathering information of the study area, field works at flood areas, and analyses from results of the Landsat 8 Operational Land Imagery (from 2010 to 2014) with the resolution of 30m (retrieved from http://earthexplorer.usgs.gov). Furthermore, the location of flood events was also verified by field works carried out in 2014 with handhold GPS devices. In summary, the total number of flood locations during the last five years was recorded to be 76. It is noted that flood locations were determined by overlaying the flood polygons in the inventory map and the Digital Elevation Model (DEM). Moreover, only pixels in the map that associate with flood points are used to extract the influencing factors used for flood prediction.

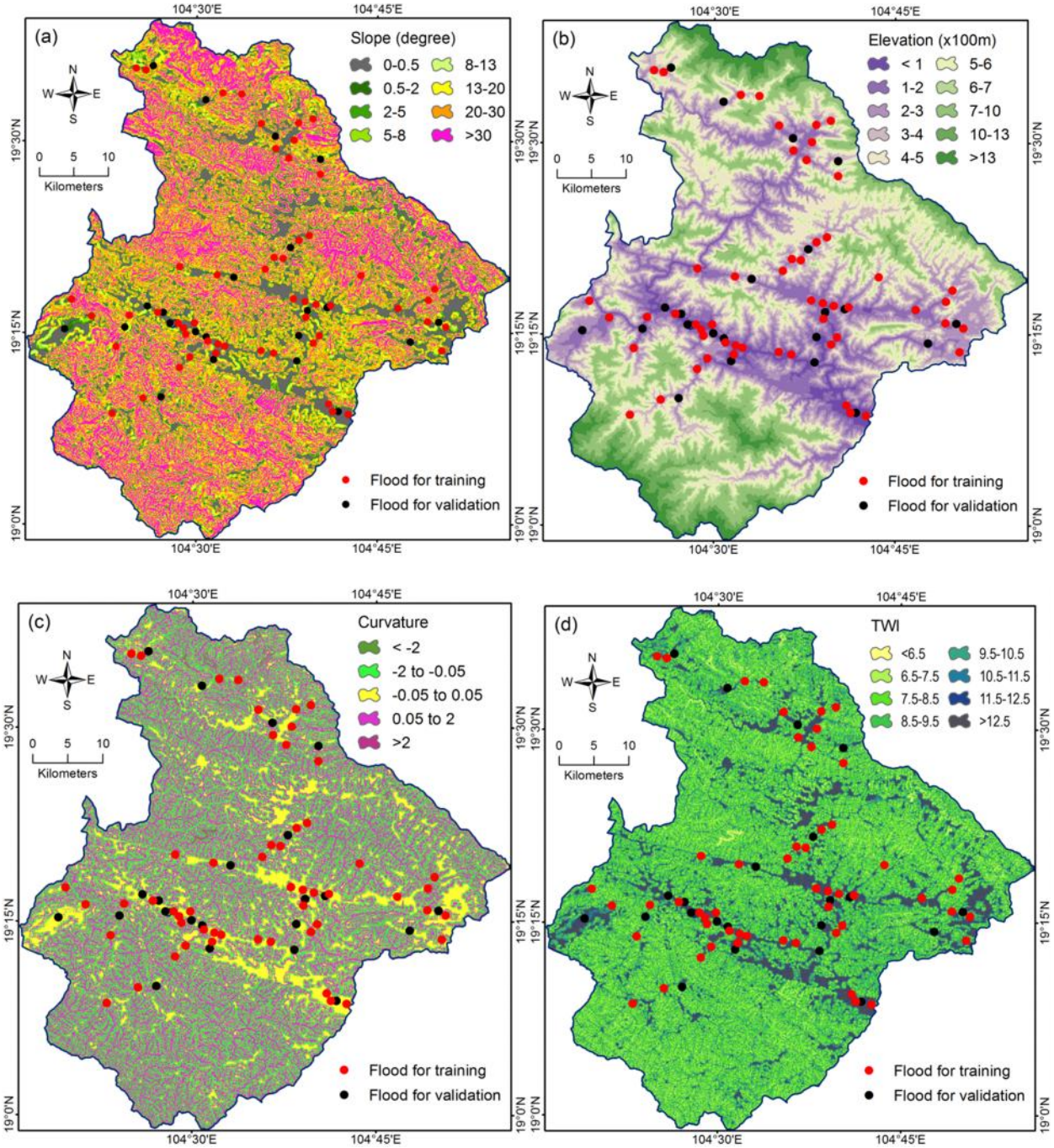
3.1.3 Flood influencing factors

To construct a flood prediction model, besides the flood inventory map, it is crucial to determine the flood influencing factors (Tehrany et al., 2015a). It is worth to notice that the selection of the flood governing factors varies due to different

160 characteristics of study areas and the availability of data (Papaioannou et al., 2015). Based on the previous work of Tien Bui et al. (2016c), the physical relationships between influencing factors and flood processes have been analyzed. Accordingly, a total of ten influencing factors were selected in this study; they include slope (IF_1), elevation (IF_2), curvature (IF_3), topographic wetness index (TWI) (IF_4), stream power index (SPI) (IF_5), distance to river (IF_6), stream density (IF_7), normalized difference vegetation index (NDVI) (IF_8), lithology (IF_9), and rainfall (IF_{10}). These factors are used to analyze
165 the flood vulnerability for the studied area and a GIS database consisting of the flood inventory map and the chosen factors has been established. The information of ten influencing factors of flood occurrence employed in this study is summarized in **Table 1**. The distributions of the ten factors within the studied region are illustrated in **Figure 2**.

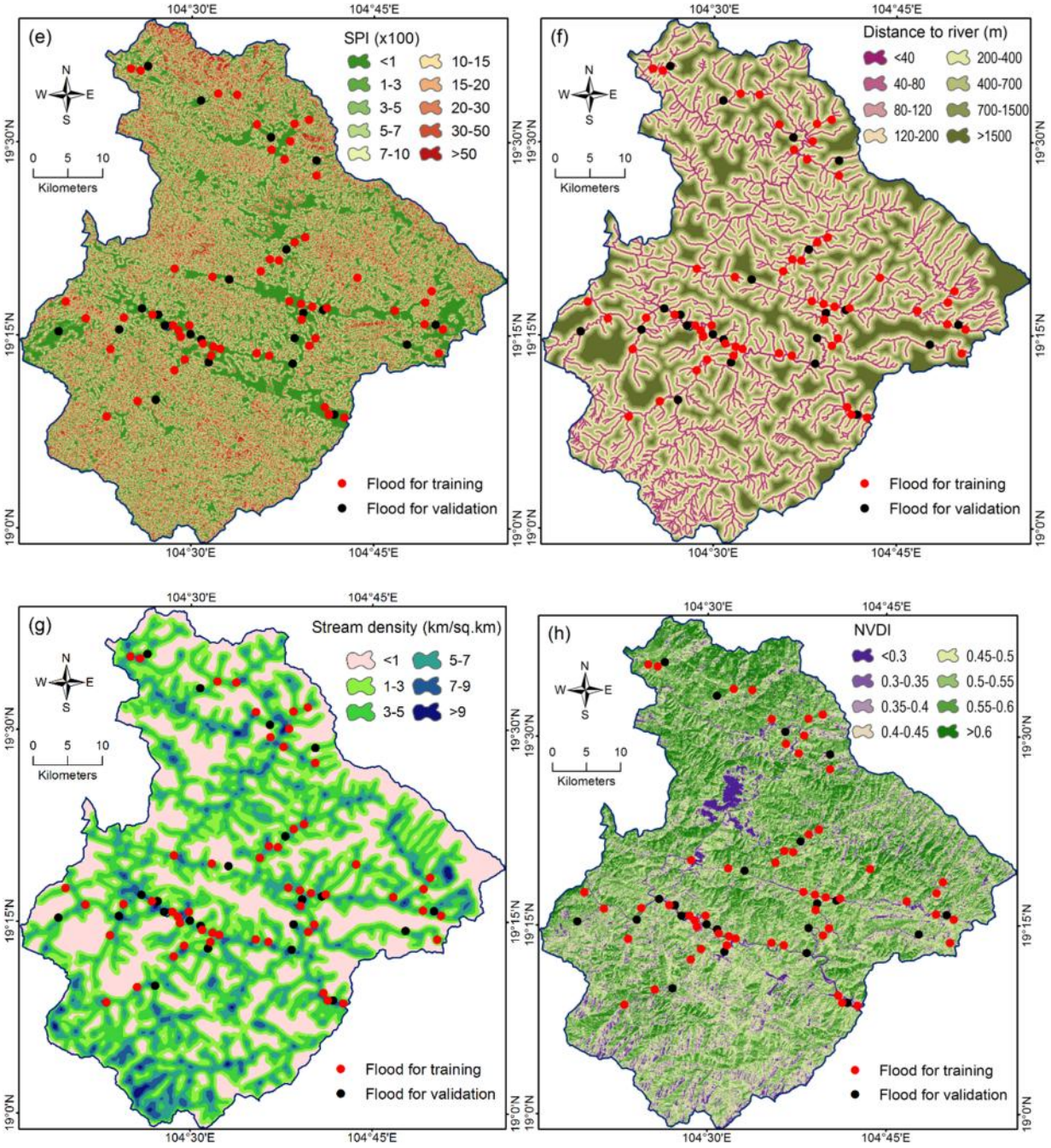
Table 1 Flood influencing factors and their categories

Factors	**Coding**	**Description of factor categories**
Slope (^{o})	IF_1	1 (0 to 0.5); 2 (0.5 to 2); 3 (2 to 5); 4 (5to 8); 5 (8 to 13); 6 (13 to 20); 7 (20 to 30); 8 (>30)
Elevation (100m)	IF_2	1 (<1); 2 (1 to 2); 3 (2 to 3); 4 (3 to 4); 5 (4 to 5); 6 (5 to 6); 7 (6 to 7); 8 (7 to 10); 9 (10 to 13); 10 (>13)
Curvature	IF_3	1 (<-2); 2 (-2 to -0.05) ; 3 (-0.05 to 0.05); 4 (0.05 to 2); 5 (>2)
Topographic Wetness Index (TWI)	IF_4	1 (<6.5); 2 (6.5 to 7.5); 3 (7.5 to 8.5); 4 (8.5 to 9.5); 5 (9.5 to 10.5); 6 (10.5 to 11.5); 7 (11.5 to 12.5); 8 (>12.5)
Stream Power Index (SPI)	IF_5	1 (<1); 2 (1 to 3); 3 (3 to 5); 4 (5 to 7); 5 (7 to10); 6 (10 to 15); 7 (15 to 20); 8 (20 to 30); 9 (30 to 50); 10 (>50)
Distance to river (m)	IF_6	1 (<40); 2 (40 to 80); 3 (80 to 120); 4 (120 to 200); 5 (200 to400); 6 (400 to 700); 7 (700 to 1500); 8 (>1500)
Stream density (km/km2)	IF_7	1 (<1); 2 (1 to 3); 3 (3 to 5); 4 (5 to 7); 5 (7 to9); 6 (>9)
Normalized Difference Vegetation Index (NDVI)	IF_8	1 (<0.3); 2 (0.3to 0.35); 3 (0.35 to 0.4); 4 (0.4 to 0.45); 5 (0.45 to0.5); 6 (0.5 to 0.55); 7 (0.55 to 0.6); 8 (>0.6)
Lithology (rock type)	IF_9	1 (Q); 2 (Nkb); 3 (Jmh); 4 (T3npb); 5 (T2); 6 (C-bslk); 7 (D-ntdl); 8 (S2-D1hn); 9 (O3-S1sc3); 10 (O3-S1sc2); 11 (O3-S1sc1); 12 (PR2bk)
Rainfall (1000mm)	IF_{10}	1 (<1.82); 2 (1.82 to 1.92); 3 (1.92 to 2.02); 4 (2.02 to 2.12); 5 (2.12 to 2.22); 6 (2.22 to 2.32); 7 (2.32 to 2.42); 8 (>2.42)



170

Figure 2 Flood influencing factors: (a) Slope, (b) Elevation, (c) Curvature, (d) Topographic wetness index



175

Figure 2 (Cont.) Flood influencing factors: (e) Stream power index, (f) Distance to river, (g) Stream density, (h) Normalized Difference Vegetation Index,

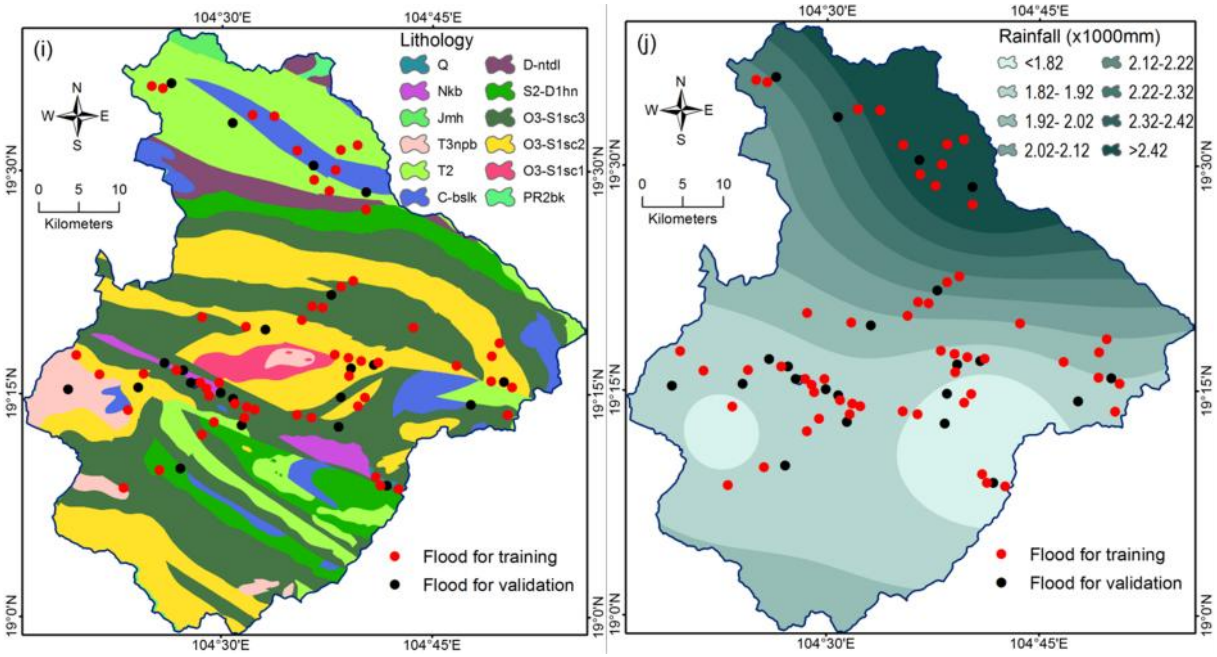


Figure 2 (Cont.) (i) Lithology, (j) Rainfall

3.2 Bayesian Framework for flood classification

The flood prediction in this study is considered as a pattern classification problem within which 'flood' and 'non-flood' are the two class labels of interest. As a result, the probability (posterior probability) of pixels belonging to the flood class, which are derived from the model, will be used as susceptibility indices. These susceptibility indices of the pixels are then used to generate the flood susceptibility map. To cope with the complexity as well as the uncertainty of the problem of interest, Bayesian framework is employed in this study to evaluate the flood susceptibility of each data sample. **Figure 3** demonstrates the general concept of the Bayesian framework used for classification.

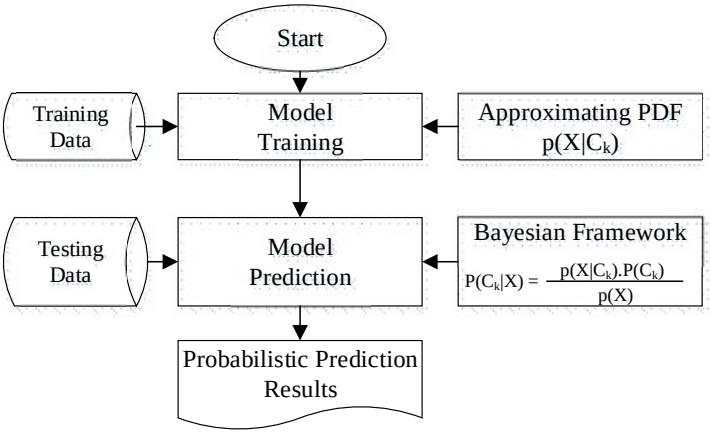


Figure 3 General concept of the Bayesian Framework for flood classification

The Bayesian framework provides a flexible way for probabilistic modeling. This method features a strong ability for dealing with uncertainty and noisy data (Theodoridis, 2015;Cheng and Hoang, 2016). Nevertheless, previous studies have rarely examined the capability of this approach for inferring flood susceptibility. Basically, pattern classification aims at assigning a pattern to one of $M = 2$ distinctive class labels C_k which k is either 1 or 2. $C_1 = 1$ and $C_2 = 0$ denote the flood class and the non-flood class, respectively. To recognize an input pattern based on the information supplied by its feature vector X, we need to attain the posterior probability $P(C_k|X)$, which indicates the likelihood that the feature vector X falls into a certain group C_k. Based on such information, the pattern will be categorized to the group with the highest posterior probability. The posterior probability $P(C_k|X)$ is calculated as follows (Webb and Copsey, 2011):

$$P(C_k \mid X) = \frac{p(X \mid C_k) \times P(C_k)}{p(X)} \quad (1)$$

where $P(C_k \mid X)$ denotes the posterior probability. $p(X \mid C_k)$ represents the likelihood which is also called the class-conditional probability density function (PDF). $P(C_k)$ denotes the prior probability, which implies the probability of the class before any feature is measured. The denominator $p(X)$ is the evidence factor; this quantity is merely a scale factor for guaranteeing that the posterior probabilities are valid; it can be calculated as follows:

$$P(X) = \sum_{k=1}^{M} p(X \mid C_k) \times P(C_k) \quad (2)$$

Generally, the prior probabilities $P(C_k)$ can be calculated by computing the ratio of training instances in each class. Thus, the bulk in establishing a Bayesian classification model is to calculate the likelihood $p(X|C_k)$. This likelihood expresses the density of input patterns in the learning space within a certain group of data. In most of situations, $p(X|C_k)$ is unknown and must be estimated from the available data. In this research, the Gaussian Mixture Model is utilized for computing the class-conditional probability density function $p(X|C_k)$.

3.3 Gaussian Mixture Model for Density Estimation

3.3.1 Gaussian Mixture Model

It is noted that the posterior probability value (Eq.1) for each pixel of the study area is used as flood susceptibility index. To obtain the posterior probability, the class-conditional probability density function (PDF) must be estimated. This section presents how PDF is estimated by a Gaussian Mixture Model (GMM). GMM is selected in this research because it has been shown to be an effective parametric method for modeling of data distribution especially in high dimensional space (McLachlan and Peel, 2000;Theodoridis and Koutroumbas, 2009). Previous studies (Paalanen, 2004;Figueiredo and Jain, 2002;Gómez-Losada et al., 2014;Arellano and Dahyot, 2016) point out that any continuous distribution can be approximated arbitrarily well by a finite mixture of Gaussian distributions. Due to their usefulness as a flexible modeling tool, GMMs have

received an increasing attention from the academic community (Zhang et al., 2016;Khanmohammadi and Chou, 2016;Ju and Liu, 2012).

In a *d*-dimensional space the Gaussian PDF is defined mathematically in the following form:

$$N(x \mid \theta) = \frac{1}{(2\pi)^{d/2} \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\} \quad (3)$$

where « denotes the vector of variable mean and Σrepresents the matrix of covariance; and $\bar{} = \{«, \Sigma\}$ denotes a set of distribution parameter.

A GMM is, in essence, an aggregation of several multivariate Normal distributions; hence, its PDF for each data sample is computed as a weighted summation of Gaussian distributions (see **Figure 4**):

$$p(x \mid \Theta) = \sum_{i=1}^{k} \alpha_i p_i(x \mid \theta_i) = \sum_{i=1}^{k} \alpha_i N(x \mid \theta_i) \quad (4)$$

where $\Theta = \{\sim_1, \sim_2, ..., \sim_k, \bar{}_1, \bar{}_2, ..., \bar{}_k\}.\{\sim_1, \sim_2, ..., \sim_k,\}$ is called the mixing coefficients of *k* Gaussian components and $\sum_{i=1}^{k} \sim_i = 1$.

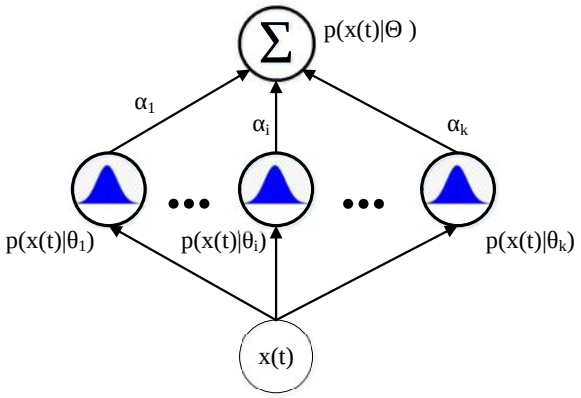


Figure 4 Structure of Gaussian Mixture Model

Accordingly, the PDF for all data samples can be expressed as follows (Ju and Liu, 2012):

$$p(X \mid \Theta) = \prod_{t=1}^{n} p(x_t \mid \Theta) = L(\Theta \mid X) \quad (5)$$

Identifying a GMM's parameters Θ can be considered as unsupervised learning task within which a dataset of independently distributed data points $X = \{x_1, ..., x_N\}$ generated from an integrated distribution dictated via the PDF $p(X \mid \Theta)$. The goal is to find the most appropriate value of Θ, denoted as Θ_e, that maximizes the log-likelihood function:

$$\Theta_e = \arg\max_{\Theta} \log(L(X, \Theta)) = \log(\prod_{t=1}^{n} p(x_t \mid \Theta)) = \sum_{t=1}^{n} \log(\sum_{i=1}^{k} \alpha_i p_i(x_t \mid \theta_i)) \quad (6)$$

Practically, instead of dealing with the log-likelihood function, an equivalent objective function *Q* is optimized (Ju and Liu, 2012):

$$\text{Max. } Q = \sum_{t=1}^{n}\sum_{i=1}^{k} w_{it} \log[\alpha_i p_i(x_t \mid \theta_i)] \tag{7}$$

where w_{it} is a posteriori probability for the *i*th class, $i=1,...k$ and w_{it} satisfies the following conditions:

$$w_{it} = \frac{\sim_i p_i(x_t \mid \bar{}_i)}{\sum_{s=1}^{k} \sim_s p_s(x_t \mid \bar{}_s)} \; ; \; \sum_{i=1}^{k} w_{it} = 1 \tag{8}$$

In order to compute Θ_e in Eq.6, the Expectation Maximization (EM) algorithm is employed. In addition, an unsupervised learning approach proposed by Figueiredo and Jain (2002) is used for determining Θ. These two algorithms are briefly reviewed in the next section of the paper.

3.3.2 Learning of finite Mixture Model with the Expectation Maximization Algorithm

The Expectation Maximization **(**EM) method is a statistical approach to fit a GMM based on historical data; this method converges to a maximum likelihood estimate of model parameters (McLachlan and Krishnan 2008). It can be recapitulated as follows (McLachlan and Peel, 2000). Commencing from an initial parameter Θ_o, an iteration of the EM algorithm consists of the *E-step* in which the current conditional probabilities $p_i(x_t \mid \bar{}_i) = N(x_i \mid «_i, \Sigma_i)$ that x_t generated from the *i*th mixture component are calculated, and the *M-step* within which the maximum likelihood estimates of $\bar{}_i$ are updated. The iteration of EM algorithm terminates when the change value of the objective function is lower than a threshold value.

These two steps of the EM procedure are stated as follows: (i) *E-step*: estimating the expected classes of all data samples for each class w_{it} based on Eq. (8); and (ii) *M-step*: calculating maximum likelihood given the data's class membership distribution using the following equations:

$$\sim_i^{new} = \frac{1}{n}\sum_{t=1}^{n} w_{it} \tag{9}$$

$$\mu_i^{new} = \sum_{t=1}^{n} w_{it} x_t \Big/ \sum_{t=1}^{n} w_{it} \tag{10}$$

$$\Sigma_i^{new} = \Big(\sum_{t=1}^{n} w_{it}(x_t - \mu_i^{new})(x_t - \mu_i^{new})^T\Big) \Big/ \sum_{t=1}^{n} w_{it} \tag{11}$$

3.3.3 Unsupervised learning of finite mixture model

The EM algorithm increases the log-likelihood iteratively until convergence is detected; and this approach generally can derive a good set of estimated parameters. Nonetheless, EM suffers from low convergence speed in some data sets, high

sensitivity to initialization condition, and sub-optimal estimated solutions (Biernacki et al., 2003). Moreover, additional efforts are required to determine an appropriate number of Gaussian distributions within the mixture.

As an attempt to alleviate such drawbacks of EM, Figueiredo and Jain (2002) put forward an unsupervised algorithm for learning a GMM from multivariate data. The algorithm features the capability of identifying a suitable number of Gaussian components autonomously; and by experiments, the authors show that the algorithm is not sensitive to initialization. In other words, this unsupervised approach incorporates the tasks of model estimation and model selection in a unified algorithm. Generally, this method can initiate with a large number of components. The initial values for component means can be assigned to all data points in the training set; in an extreme case, it is possible to distribute the component number equal to the data point number. This algorithm gradually fine-tunes the number of mixture components by casting out element of Normal distributions that are irrelevant for the data modeling process (Paalanen, 2004).

Furthermore, Figueiredo and Jain (2002) employed the Minimum Message Length (MML) criterion (Wallace and Dowe, 1999) as an index for model selection; the application of this criterion for the case of GMM learning leads to the following objective function (Figueiredo and Jain, 2002):

$$\Omega(\Theta \mid \mathrm{X}) = \frac{\mathrm{N}}{2} \sum_{\mathrm{i}:\alpha_{\mathrm{i}}>0} \ln(\frac{\mathrm{n}\alpha_{\mathrm{i}}}{12}) + \frac{\mathrm{C}_{\mathrm{nz}}}{2} \ln(\frac{\mathrm{n}}{12}) + \frac{\mathrm{C}_{\mathrm{nz}}(\mathrm{N}+1)}{2} - \ln \mathrm{L}(\mathrm{X}, \Theta) \tag{12}$$

where *n* denotes the size of the training set, *N* represents the number of hyper-parameters needed to construct a Gaussian distribution, and C_{nz} is the number of Gaussian distribution component featuring nonzero weight ($\sim_i > 0$). Accordingly, the EM method is then utilized to minimized Eq. 12 with a fixed number of C_{nz}.

In detail, the EM algorithm is employed to estimate $\sim_i$ as follows:

$$\alpha_{\mathrm{i}}^{\mathrm{new}} = \frac{\max\{0, (\sum_{\mathrm{t}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{it}}) - \frac{\mathrm{N}}{2}\}}{\sum_{\mathrm{j}=1}^{\mathrm{k}} \max\{0, (\sum_{\mathrm{t}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{jt}}) - \frac{\mathrm{N}}{2}\}} \tag{13}$$

Accordingly, the parameters $«_i^{new}$ and Σ_i^{new} are updated based on Eqs. 10 and 11, respectively. The algorithm stops when the relative decrease in the objective function $\Omega(\Theta \mid \mathrm{X})$ becomes smaller than a preset threshold (e.g. 10^{-5}).

3.4 Radial Basis Function Fisher Discriminant Analysis for Generation of Latent Variable

In machine learning, the performance of a model may be enhanced if latent variables were used (Yu, 2011).Therefore, latent variable approach is employed in this research. Accordingly, Radial Basis Function Fisher Discriminant Analysis (RBFDA) proposed Mika et al. (1999), an extension of the Fisher Discriminant Analysis for dealing with data nonlinearity, is used to generate a latent factor for flood analysis. Thus, RBFDA is utilized to project the feature from the original learning space to a projected space that expresses a high degree of class reparability (Theodoridis and Koutroumbas, 2009). Using

this kernel technique, the data from an input space *I* is first mapped into a high dimensional feature space *F*. Hence, discriminant analysis tasks can be performed nonlinearly in *I*.

Herein, $\varphi(.)$ is defined as a transformation from an input space *I* to a high dimensional feature space *F*, to compute *w* (the projecting vector), it is necessary to maximize the Fisher discriminant ratio as follows:

$$J(w) = \frac{w^T S_B^{\varphi} w}{w^T S_W^{\varphi} w} \tag{14}$$

where $S_B^{\square} = (m_1^{\square} - m_2^{\square})(m_1^{\square} - m_2^{\square})^T$ (15)

$$S_W^{\varphi} = \sum_{k=1}^{C} \sum_{i=1}^{Nk} (\varphi(x_i) - m_k^{\varphi})(\varphi(x_i) - m_k^{\varphi})^T \tag{16}$$

$$m_k^{\varphi} = \frac{1}{Nk} \sum_{i=1}^{Nk} \varphi(x_i^k) \tag{17}$$

To obtain *w*, the kernel trick is applied. Thus, one only needs to establish a formulation of the algorithm which only requires dot-product $\varphi(x).\varphi(y)$ of the training data and employ kernel functions which calculate $\varphi(x).\varphi(y)$. The widely-employed Radial Basis Kernel Function (RBKF) is expressed in the following formula (with σ denotes the kernel function bandwidth):

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{18}$$

Since a solution of the vector w lies in the span of all data samples in the projected space, the transformation vector w is shown in the following formula:

$$w = \sum_{i=1}^{N} \sim_i \square(x_i) \tag{19}$$

From Eq. (17) and Eq. (19), we have: $w^T m_k^{\varphi} = \frac{1}{Nk} \sum_{j=1}^{N} \sum_{i=1}^{Nk} \alpha_j k(x_j, x_i^k) = \alpha^T M_k$; $M_k = \frac{1}{Nk} \sum_{i=1}^{Nk} k(x_j, x_i^k)$ (20)

Taking into account the formulas of $J(w)$, $S_B^{\square}$, as well as Eq. (20), we can restate the numerator of Eq. (14) in the following manner:

$w^T S_B^{\square} w = \sim^T M \sim$; where $M = (M_1 - M_2)(M_1 - M_2)^T$ (21)

Based on the Eq. (17) that defines $m_k^{\square}$, the denominator of Eq. (14) can be demonstrated in the following way:

$$w^T S_W^{\square} w = \sim^T N \sim \tag{22}$$

where $N = \sum_{k=1}^{2} K_k (I - 1_{l_k}) K_k^T$; K_k denotes a *N*-by-N_k kernel matrix with a typical element is $k(x_n, x_m^k)$; *I* represents the identity matrix and 1_{l_k} is a matrix within which all positions are $1/l_k$.

Considering all Eq. (14), Eq. (21), and Eq. (22), the solution of RBFDA can be found by maximizing:

$$J(\alpha) = (\alpha^T M \alpha) / (\alpha^T N \alpha) \tag{23}$$

The optimization problem with the objective function expressed in Eq. (23) is found by identifying the primal eigenvector of $N^{-1}M$. Based on the optimization results, an input patter in I is projected on to a line defined by the vector w in the following manner:

$$w.\varphi(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x) \tag{24}$$

4 The proposed Bayesian Framework for Flood Susceptibility Prediction

4.1 The established GIS database

To formulate a flood assessment model, the first stage is to construct a GIS database (see **Figure 5**) within which locations of past flood events, maps of topographic feature, Landsat 8 imagery, maps of geological feature, and precipitation statistical records are acquired and integrated. In this study, the data acquisition, processing, and integration were performed with ArcGIS (version 10.2) and IDRISI Selva (version 17.01) software packages.

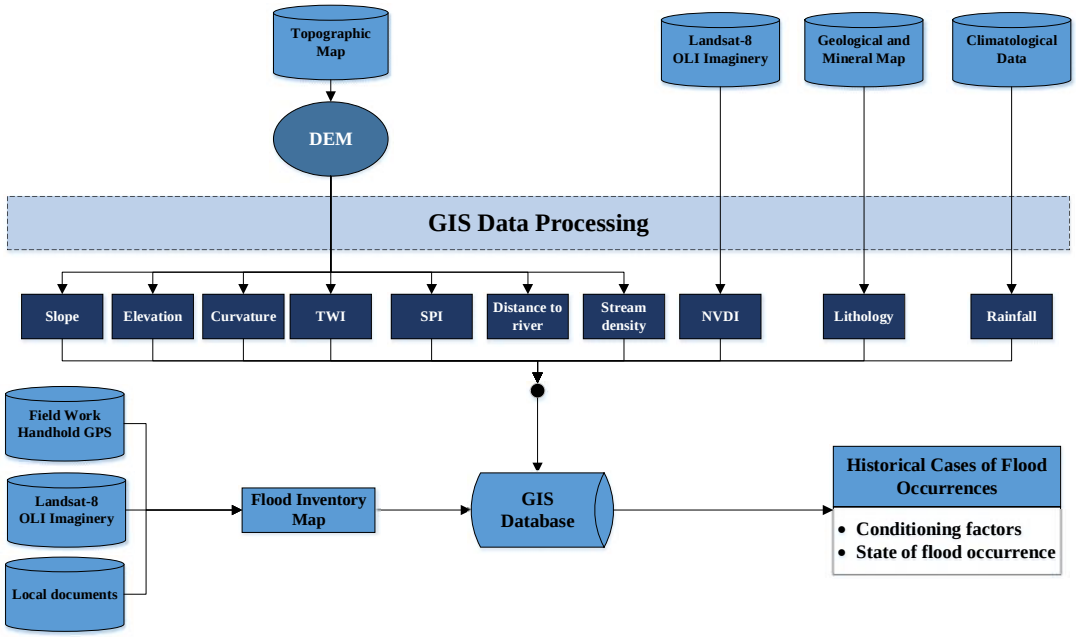


Figure 5 The established GIS database

Furthermore, a C++ application has been developed by the authors to transform the flood susceptibility indices into a GIS format for ArcGIS implementation. Accordingly, the compiled outcomes are employed to form a database that includes

the aforementioned flood influencing features with two class outputs: “flood” and “non-flood”. As mentioned earlier, a total of 76 flood locations have been recorded. To balance the dataset and reliably construct the flood prediction model, 76 locations of non-flood areas are randomly sampled and included for analysis. Hence, the total database consists of 152 data samples.

4.2 The Proposed Model Structure

The proposed model for flood susceptibility assessment that incorporates RBFDA, the Bayesian classification framework, and GMM is presented in this section of the study. The overall flowchart of the proposed Bayesian framework based on GMM and RBFDA for flood susceptibility prediction, named as BayGmmKda, is demonstrated in **Figure 6**.

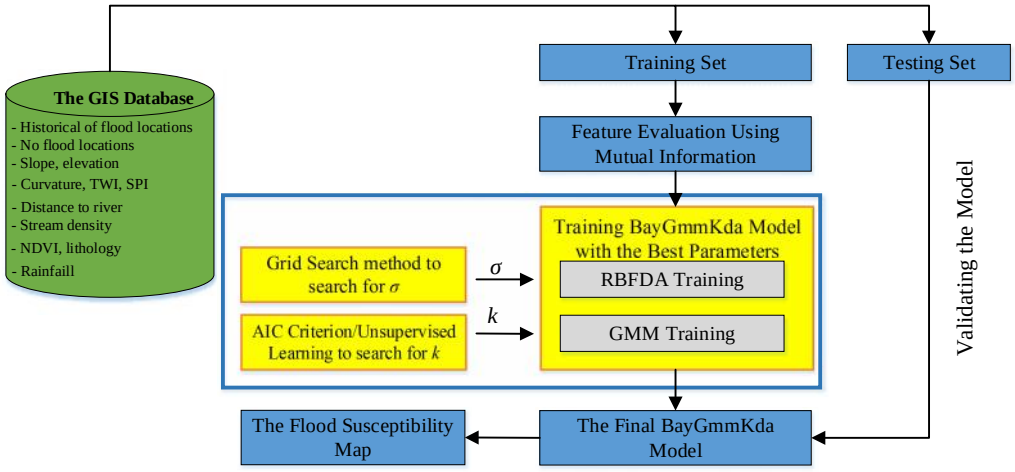


Figure 6 The proposed BayGmmKda

Firstly, the whole dataset, including 152 data samples, was separated into two sets: Training Set (90% or 137 samples) employed for model establishing and Testing Set (10% or 15 samples) used for model testing. It is noted that the input variables of the dataset have been normalized using the Min-Max normalization; the purpose of data normalization was to hedge against the situation of unbalanced variable magnitudes.

Secondly, a latent input factor was generated using the RBFDA (explained in section 3.4) and added to the training dataset aiming to enhance the classification performance. Subsequently, the feature evaluation was performed to quantify the degree of relevance of each input factors with the flood inventories in the Training Set. Any non-relevant factor should be eliminated from the modeling process to reduce noise and enhance the model performance (Tien Bui et al., 2016a;Tien Bui et al., 2017). For this purpose, in this research, the Mutual Information Criterion (Kwak and Choi, 2002;Hoang et al., 2016), a widely employed techniques for feature selection in machine learning, was selected to express the pertinence of each influencing factor to the flood. It is noticed that the larger the mutual information, the stronger the relevancy between the influencing factor and flood.

In the next step, the BayGmmKda model was trained and established using the Training Set. The purpose of the training process was to find the best parameters for the mixture component (*k*) used in GMM and the kernel function bandwidth (±) used in RBFDA of the BayGmmKda model. To determine the best *k*, the EM algorithm that employs Akaike Information Criterion (AIC) (Akaike, 1974) was used. Thus, the value of *k* was varied from 1 to 20, and then, AIC was estimated and used to select the model that exhibits the best fit to the data at hand. It is noted that a model with a few number of mixture components (*k*) indicates a less degree of complexity (Olivier et al., 1999). In addition, the unsupervised GMM learning (Figueiredo and Jain, 2002) is also used for autonomously determining the best *k*. Accordingly, the model starts with a maximum component number (*k*) of 20; the algorithm carries out the model selection process by removing irrelevant mixture components if applicable. To determine the best ± , the Grid Search procedure is performed and the parameter ± corresponding to the highest classification accuracy rate was selected.

Using the best *k* and ± in the previous step, the final BayGmmKda model was finally constructed and the Bayesian classification framework was derived. The Bayesian framework was then used to estimate the posterior probability (flood susceptibility index) for all the pixels in the study areas. The flood susceptibility index was then transferred to a raster format to open in ArcGIS.

4.3 The Developed Matlab Interface of BayGmmKda

It is noted that GMM with the EM training algorithm is implemented with the Matlab statistical toolbox (MathWorks, 2012a); meanwhile, the BayGmmKda performs the unsupervised algorithm with the program code provided by Figueiredo (2002). The RBFDA algorithm and the unified BayGmmKda model have been coded in Matlab by the authors. In addition, a software program with a graphical user interface (GUI) (see **Figure 7**) for the implementation of theBayGmmKda model has been coded in Matlab environment by the authors. The GUI development aims at providing a user-friendly system for performing flood susceptibility predictions.

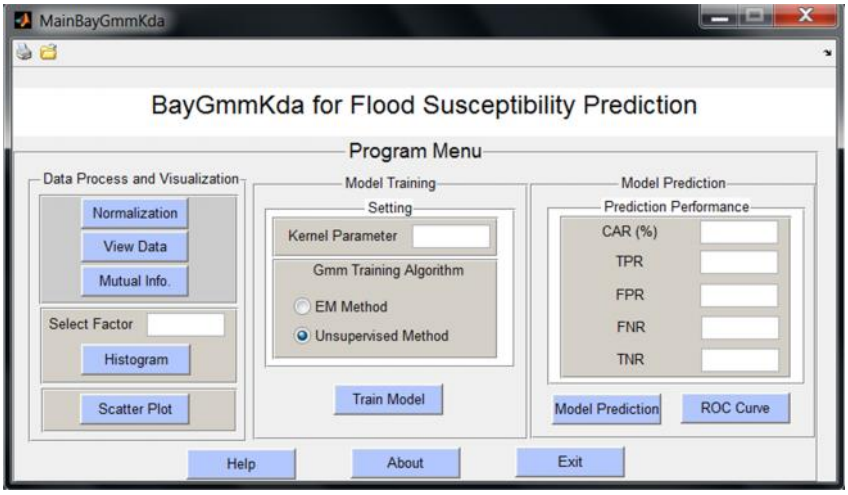


Figure 7 Main Menu of BayGmmKda

As shown in **Figure 7**, **t**he program consists of three modules: Data Process and Visualization, Model Training, and Model Prediction. The first module provides basic functions for data inspection and visualization including data normalization, data viewing, and preliminary feature selection with mutual information. In the second module, the users simply provide model parameters including the kernel function parameter and the GMM training method. The trained model is employed to carry out prediction tasks in the third module, within which the model prediction performance is reported.

5. Experimental Results

5.1 Feature selection and training of the BayGmmKda model

The outcome of the preliminary examination on the pertinence of flood influencing factors is reported in **Figure 8a**. As mentioned earlier, the relevancies of influencing factors are exhibited by the mutual information criterion. Based on the outcome, IF_5 (SPI) features the highest mutual dependence, followed by IF_7 (stream density) and IF_8 (NVDI). Influencing factors of IF_4 (TWI) and IF_{10} (rainfall) exhibit comparatively low values of mutual information. Because all the mutual information values are not null, all influencing factors deem to be relevant and should be retained for the subsequent processes of model training and prediction.

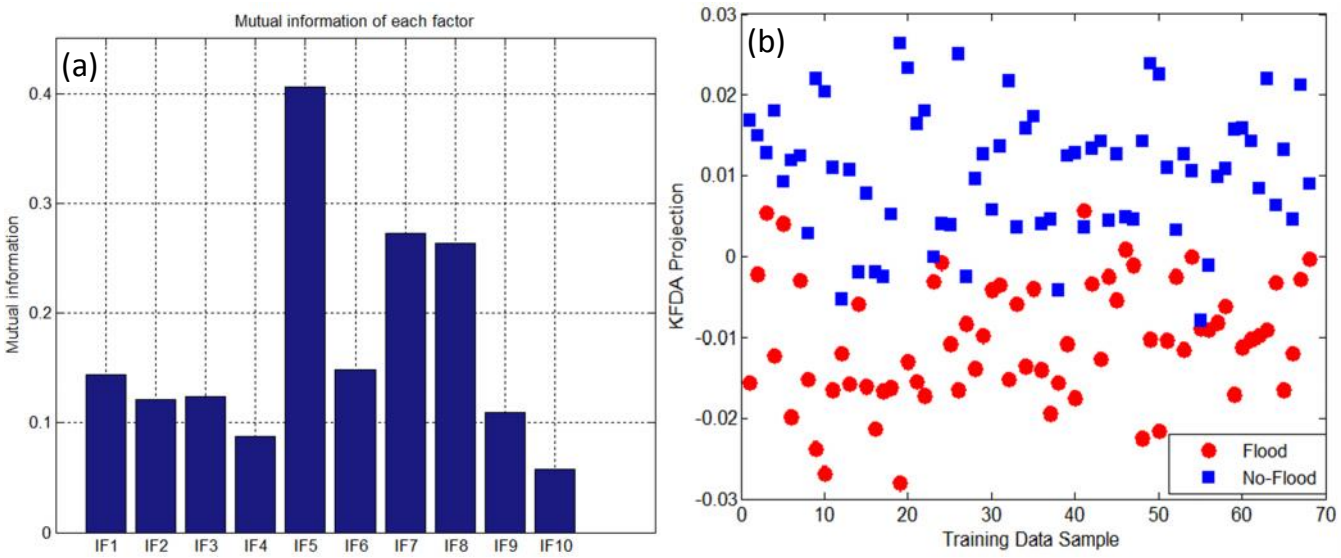


Figure 8 (a) Mutual information of flood influencing factors; (b) RBFDA-based latent factor derived in this study

It is worth reminding that the BayGmmKda's training phase is executed in two consecutive steps, training RBFDA and training GMM. RBFDA analyzes the data in the Training Set to establish a latent factor which is a one-dimensional representation of the original input pattern. **Figure 8b** shows the resulted latent factor constructed by RBFDA. In the next step of the training phase, GMM is constructed by the original input patterns with their corresponding labels which consist of ten input factors and with the RBFDA -based latent factor.

Classification Accuracy Rate (CAR) is employed to exhibit the rate of correctly classified instances. In addition, a more detailed analysis on the model capability can be presented by calculating True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR). These four rates are also widely utilized to exhibit the predictive capability of a prediction model (Hoang and Tien-Bui, 2016).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,;\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}\,;\ \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}\,;\ \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{25}$$

where TP, TN, FP, and FN represent the values of true positive, true negative, false positive, and false negative, respectively.

In addition to the four rate, the Receiver Operating Characteristic (ROC) curve (van Erkel and Pattynama, 1998) is used to summary the global performance of the model. The ROC curve basically demonstrates the trade-off between the two aforementioned TPR and FPR when the threshold for accepting the positive class of 'flood' varies. In addition, the Area Under the ROC Curve (AUC) is employed to quantify the global performance. In generally, a better model is characterized by a larger value of AUC.

As aforementioned, the data set is randomly separated into the Training Set and the Testing Set which occupy 90% and 10% of the data samples, respectively. The Training Set is employed to train the mode; meanwhile, the Testing Set is used for validating the model capability after being trained. Since one selection of data for the Training Set and the Testing Set may not truly demonstrate the model's predictive capability, this study carries out a repetitive sub-sampling procedure within which 30 experimental runs is carried out. In each experimental run, 10% of the data set is retrieved in a random manner from the database to constitute Testing Set; the rest of the database is included in the Training Set.

The testing performance of the proposed Bayesian framework for flood susceptibility is reported in **Table 2** and **Figure 9**, which provides the average ROC curves of the proposed model framework, obtained from the random subsampling process, with two methods of GMM training. Herein, the two Bayesian models that employ the EM algorithm and the Unsupervised Learning algorithm (UL) for training GMM are denoted as BayGmmKda-EM and BayGmmKda-UL, respectively. As can be seen that the BayGmmKda-UL model demonstrates clearly better predictive performance (CAR = 89.58%, AUC = 0.94, TPR = 0.96, TNR = 0.91) than that of the BayGmmKda- EM model (CAR = 86.67%, AUC = 0.93, TPR = 0.95, TNR = 0.85). Although the performances of the BayGmmKda-EM model and the BayGmmKda-UL model are comparable in TPR, however, the BayGmmKda-UL model deems more accurate than the BayGmmKda-EM model when the two models predict samples with the non-flood class.

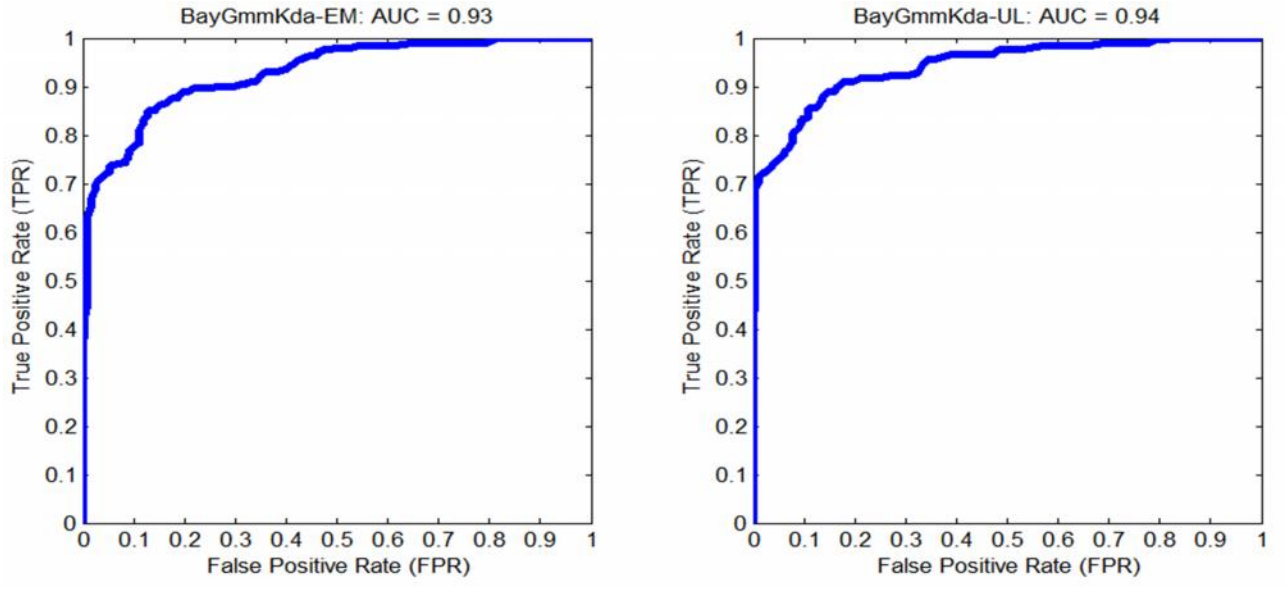


Figure 9 ROC plots of the proposed BayGmmKda

Table 2 Prediction Results of BayGmmKda

Data Set	CAR (%)	AUC	TPR	FPR	FNR	TNR
Average						
BayGmmKda-EM	86.67	0.93	0.95	0.12	0.15	0.85
BayGmmKda-UL	89.58	0.94	0.96	0.12	0.09	0.91
Standard deviation						
BayGmmKda-EM	6.51	0.07	0.05	0.10	0.12	0.12
BayGmmKda-UL	7.22	0.05	0.04	0.11	0.10	0.10

5.2 Model comparison

Because this is the first time the BayGmmKda model proposed for the flood susceptibility, therefore the valid of the proposed model should be assessed. For this task, the benchmarks were used for the comparison, including support vector machine (SVM), Adaptive Neuro Fuzzy Inference System (ANFIS), and the GMM-based Bayesian classifier. The above machine learning techniques were selected because SVM and ANFIS have been recently verified to be effective tools for predicting flood susceptibility (Tien Bui et al., 2016c;Tehrany et al., 2015b). Whereas the GMM-based Bayesian classifier (BayGmm) is the Bayesian framework for classification which employs GMM for density estimation, but is not integrated with the RBFDA algorithm as in the BayGmmKda model. It is emphasis that BayGmm is included for the comparison to confirm the advantage of the newly constructed BayGmmKda and to verify the usefulness of RBFDA in enhancing the discriminative capability of the hybrid framework.

To construct the SVM model, the model's hyperparameters of the regularization constant (C) and the parameter of the radial basis kernel function ($\pm$) need to be specified. Herein, a grid search process, which is identical to the one used to identify the kernel function bandwidth used in RBFDA, is employed to fine-tuned such hyperparameters of the SVM model. It is noted that SVM method is implemented in Matlab package (MathWorks, 2012b). Meanwhile, the ANFIS model used in this section is trained with the metaheuristic approach described in the previous work of Tien Bui et al. (2016c).

Table 3 Performance comparison of the BayGmmKda model with the three benchmarks, the SVM model, the ANFIS model, and the BayGmm model.

Models	CAR (%)	AUC	TPR	FPR	FNR	TNR
Average						
BayGmmKda	89.58	0.94	0.96	0.12	0.09	0.91
ANFIS	85.63	0.83	0.84	0.13	0.16	0.87
BayGmm	85.02	0.92	0.82	0.13	0.17	0.88
SVM	83.75	0.82	0.78	0.10	0.22	0.90
Standard deviation						
BayGmmKda	7.22	0.05	0.04	0.11	0.10	0.10
ANFIS	6.17	0.05	0.14	0.10	0.14	0.10
BayGmm	7.24	0.08	0.11	0.10	0.11	0.10
SVM	10.33	0.06	0.16	0.11	0.16	0.11

It is noted that a random subsampling with 30 runs is employed for all models in this experiment. The result comparison between the proposed BayGmmKda model and three benchmark models is shown in **Table 3**. The result shows that the proposed model yields the best results (CAR = 89.58% and AUC = 0.94). It is followed by the ANFIS model (CAR = 85.63%, AUC = 0.83); the BayGmm model (85.02%, AUC = 0.92), and the SVM model (83.75%, AUC = 0.82).

Table 4 Model Comparison Based on the Wilcoxon signed-rank test

	BayGmmKda	ANFIS	BayGmm	SVM
BayGmmKda		++	++	++
ANFIS	--		+	+
BayGmm	--	-		+
SVM	--	-	-	

To confirm the performance of the proposed BayGmmKda model is significantly higher than that of the three benchmark model, the Wilcoxon signed-rank test is employed. The Wilcoxon signed-rank test is widely used to evaluate whether classification outcomes of prediction models are significantly dissimilar (Tien Bui et al., 2016e). Using this test, the

p-values that obtained from experimental results of the four models can be computed using a threshold value of 0.05. The result of the Wilcoxon signed-rank test is shown in **Table 4**. It is noted that the signs "++", "+", "--", and "-" represent a significant win, a win, a significant loss, and a loss, respectively. The result confirms that the proposed BayGmmKda model achieves significant wins over the other models.

5.3 Construction of the flood susceptibility map

Experimental outcomes have indicated that the BayGmmKda model is the best for this study area, therefore the model was used to compute the posterior probability for all the pixels of the study area. The posterior probability values that were used as flood susceptibility indices were further transformed to a raster format and open in ArcGIS 10.4 software package. Using these indices, the flood susceptibility map (see **Figure 10**) was derived and visualized by mean of five classes: very high (10%), high (10%), moderate (10%), low (20%), and very low (50%). The threshold values for separating these classes were determined by overlaying the historical flood locations and the flood susceptibility indices map (Tien Bui et al., 2016c), and then, a graphical curve (see **Figure 10**) was constructed and the threshold values were derived.

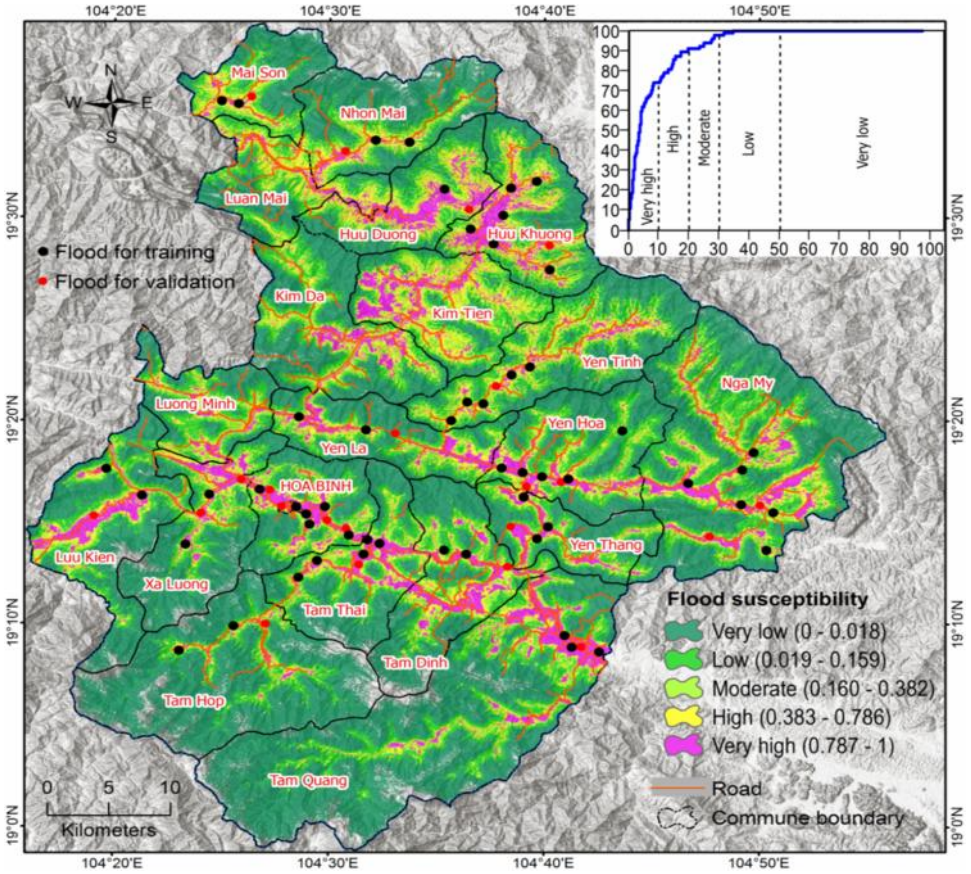


Figure 10 The flood susceptibility map using the proposed BayGmmKda model for the study area

Interpretation of the map shows that 10% of the Tuong Duong district was classified into the very high class and this class covers 73.68% of the total historical flood locations. Meanwhile, both the high class and the moderate classes cover 10% of the region but account for only 15.79% and 7.9% of the total historical flood locations, respectively. Whereas, the low class covers 20% of the district but it contains only 2.63% of the total historical flood locations. Particularly, 50% of the district, which is categorized to the very low class, contains no flood location. These indicate that the proposed BayGmmKda model has successfully delineated susceptible flood prone-areas. In other words, the interpretation results confirm the reliability of the proposed Bayesian framework in this work.

6. Conclusion

This research has developed a new tool, named as BayGmmKda, for flood susceptibility evaluation with a case study in a high frequency flood area at Central Vietnam. The newly constructed model is a Bayesian framework that combines GMM and RBFDA for spatial prediction of flood. A GIS database has been established to train and test the BayGmmKda method. The training phase of BayGmmKda consists of two steps: (i) discriminant analysis with RBFDA in which a latent factor is generated and (ii) density estimation using GMM. After the training phase, the Bayesian framework is employed to compute the posterior probability. The posterior probability was then used as flood susceptibility index. Furthermore, a Matlab program with GUI has been developed to ease the implementation of the BayGmmKda model in flood vulnerability assessment.

It is noted that in this study, the GMM training is performed with two methods: the EM algorithm and the unsupervised learning approach. Furthermore, a repeated subsampling process with 30 experimental runs is carried to evaluate the model prediction outcome. The subsampling process verified by statistical test confirms that the GMM method trained by the unsupervised learning approach has attained a better prediction accuracy compared with the EM algorithm. Therefore, this method of GMM learning is strongly recommended for other studies in the same field.

Furthermore, the experiments demonstrate that the latent factor created by RBFDA is really helpful in boosting the classification accuracy of the BayGmmKda model. This melioration in accuracy of the BayGmmKda stems from its integrated learning structure. As described earlier, the classification task is performed by a hybridization of discrimination analysis and Bayesian framework. The Bayesian model carried out the classification task by consideration of the patterns in the original dataset and an additional factor produced from the discrimination analysis. As result, the performance of the BayGmmKda model is better than those obtained from the three benchmarks (SVM ANFIS, and BayGmm).

The main limitation in this work is that the BayGmmKda is a data-driven tool; therefore, field works and GIS-based geo-environmental data are necessary for the model construction phase. These data collecting and analyzing can be time-consuming. In addition, the grid search procedure is used for hyper-parameter setting in the BayGmmKda model requires a high computational cost especially for large-scale data sets. Furthermore, the outcome of this grid search procedure may not

be optimal; therefore, more advanced model selection approaches i.e. metaheuristics optimization algorithms could be utilized to further improve the model accuracy.

Despite such limitations, the proposed BayGmmKda model, featured by its high predictive accuracy and the capability of delivering probabilistic outputs, is a promising alternative for flood susceptibility prediction. Future extensions of this research may include the model application in flood prediction for other study areas, investigations of other flood influencing factors i.e. streamflow and antecedent soil moisture which may be relevant for flood analysis, and improving the current model with other novel soft computing methods i.e. feature selection, pattern classification, and dimension reduction to alleviate the aforementioned drawbacks as well as to enhance the model performance.

7. Code availability

The Matlab code of the BayGmmKda model is given in the Supplement.

8. Data availability

The dataset used in this research is given in the Supplement.

Acknowledgements

This research was partially supported by Department of Business and IT, School of Business, University College of Southeast Norway. Data for this research are from the Project No. B2014-02-21 and were provided by Dr. Quoc-Phi Nguyen (Hanoi University of Mining and Geology, Vietnam).

References

Akaike, H.: A new look at the statistical identification model, IEEE Trans. Automat. Control., 19, 716–723, 10.1109/TAC.1974.1100705, 1974.

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., Roo, A., Salamon, P., Wyser, K., and Feyen, L.: Global projections of river flood risk in a warmer world, Earth's Future, 5, 171-182, 2017.

Arellano, C., and Dahyot, R.: Robust ellipse detection with Gaussian mixture models, Pattern Recognit., 58, 12-26, http://dx.doi.org/10.1016/j.patcog.2016.01.017, 2016.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: Model development1. Wiley Online Library, 1998.

Bennett, J. C., Robertson, D. E., Ward, P. G., Hapuarachchi, H. P., and Wang, Q.: Calibrating hourly rainfall-runoff models with daily forcings for streamflow forecasting applications in meso-scale catchments, Environmental Modelling & Software, 76, 20-36, 2016.

Beven, K., Kirkby, M., Schofield, N., and Tagg, A.: Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments, Journal of Hydrology, 69, 119-143, 1984.

Biernacki, C., Celeux, G., and Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Comput. Stat. Data. Anal., 41, 561-575, http://dx.doi.org/10.1016/S0167-9473(02)00163-9, 2003.

Birkel, C., Tetzlaff, D., Dunn, S., and Soulsby, C.: Towards a simple dynamic process conceptualization in rainfall–runoff models using multi-criteria calibration and tracers in temperate, upland catchments, Hydrological Processes, 24, 260-275, 2010.

Brocca, L., Melone, F., and Moramarco, T.: Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting, Hydrological processes, 25, 2801-2813, 2011.

Bubeck, P., Botzen, W., and Aerts, J.: A review of risk perceptions and other factors that influence flood mitigation behavior, Risk. Anal. , 32, 1481–1495, 10.1111/j.1539-6924.2011.01783.x, 2012.
Cheng, M.-Y., and Hoang, N.-D.: Slope Collapse Prediction Using Bayesian Framework with K-Nearest Neighbor Density Estimation: Case Study in Taiwan, J. Comput. Civ. Eng., 30, 04014116, doi:10.1061/(ASCE)CP.1943-5487.0000456, 2016.
Chiew, F. H. S., Stewardson, M. J., and McMahon, T. A.: Comparison of six rainfall-runoff modelling approaches, Journal of Hydrology, 147, 1-36, http://dx.doi.org/10.1016/0022-1694(93)90073-I, 1993.
Ciabatta, L., Brocca, L., Massari, C., Moramarco, T., Gabellani, S., Puca, S., and Wagner, W.: Rainfall-runoff modelling by using SM2RAIN-derived and state-of-the-art satellite rainfall products over Italy, International Journal of Applied Earth Observation and Geoinformation, 48, 163-173, 2016.
Cunnane, C.: Methods and merits of regional flood frequency analysis, Journal of Hydrology, 100, 269-290, 1988.
Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L.: Development and evaluation of a framework for global flood hazard mapping, Adv. Water Resour., 94, 87-102, http://dx.doi.org/10.1016/j.advwatres.2016.05.002, 2016.
Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, n/a-n/a, 10.1029/2006WR005563, 2008.
Figueiredo, M. A. T.: http://www.lx.it.pt/~mtf/, Access Date: 01/04/2016, 2002.
Figueiredo, M. A. T., and Jain, A. K.: Unsupervised learning of finite mixture models, IEEE Trans. Pattern Anal. Mach. Intell., 24, 381-396, 10.1109/34.990138, 2002.
Gao, Z., Long, D., Tang, G., Zeng, C., Huang, J., and Hong, Y.: Assessing the potential of satellite-based precipitation estimates for flood frequency analysis in ungauged or poorly gauged tributaries of China's Yangtze River basin, Journal of Hydrology, 550, 478-496, https://doi.org/10.1016/j.jhydrol.2017.05.025, 2017.
Gómez-Losada, Á., Lozano-García, A., Pino-Mejías, R., and Contreras-González, J.: Finite mixture models to characterize and refine air quality monitoring networks, Sci. Total Environ., 485–486, 292-299, http://dx.doi.org/10.1016/j.scitotenv.2014.03.091, 2014.
Grimaldi, S., Petroselli, A., Arcangeletti, E., and Nardi, F.: Flood mapping in ungauged basins using fully continuous hydrologic–hydraulic modeling, Journal of Hydrology, 487, 39-47, 2013.
Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S.: Global flood risk under climate change, Nature Climate Change, 3, 816-821, 2013.
Hoang, N.-D., and Pham, A.-D.: Hybrid artificial intelligence approach based on metaheuristic and machine learning for slope stability assessment: A multinational data analysis, Expert. Syst. Appl., 46, 60–68, http://dx.doi.org/10.1016/j.eswa.2015.10.020, 2016.
Hoang, N.-D., and Tien-Bui, D.: A Novel Relevance Vector Machine Classifier with Cuckoo Search Optimization for Spatial Prediction of Landslides, J. Comput. Civ. Eng., 30, 04016001, 10.1061/(ASCE)CP.1943-5487.0000557, 2016.
Hoang, N.-D., Tien Bui, D., and Liao, K.-W.: Groutability estimation of grouting processes with cement grouts using Differential Flower Pollination Optimized Support Vector Machine, Appl. Soft Comput., 45, 173-186, http://dx.doi.org/10.1016/j.asoc.2016.04.031, 2016.
Ju, Z., and Liu, H.: Fuzzy Gaussian Mixture Models, Pattern Recognit., 45, 1146-1158, http://dx.doi.org/10.1016/j.patcog.2011.08.028, 2012.
Kazakis, N., Kougias, I., and Patsialis, T.: Assessment of flood hazard areas at a regional scale using an index-based approach and Analytical Hierarchy Process: Application in Rhodope–Evros region, Greece, Sci Total Environ., 538, 555-563, http://dx.doi.org/10.1016/j.scitotenv.2015.08.055, 2015.
Khanmohammadi, S., and Chou, C.-A.: A Gaussian mixture model based discretization algorithm for associative classification of medical data, Expert. Syst. Appl., 58, 119-129, http://dx.doi.org/10.1016/j.eswa.2016.03.046, 2016.
Kia, M. B., Pirasteh, S., Pradhan, B., Mahmud, A. R., Sulaiman, W. N. A., and Moradi, A.: An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia, Environ. Earth Sci., 67, 251-264, 10.1007/s12665-011-1504-z, 2012.
Kreft, S., Eckstein, D., Junghans, L., Kerestan, C., and Hagen, U.: Global climate risk index 2015: Who suffers most from extreme weather events, Report from Germanwatch, 1-31, 2014.
Kwak, N., and Choi, C.-H.: Input feature selection by mutual information based on Parzen window, IEEE Trans. Pattern Anal. Mach. Intell., 24, 1667-1671, 10.1109/TPAMI.2002.1114861, 2002.
Lee, M. J., Kang, J. e., and Jeon, S.: Application of frequency ratio model and validation for predictive flooded area susceptibility mapping using GIS, In Proc. of the 2012 IEEE International Geoscience and Remote Sensing Symposium, 2012, 895-898,
Lohani, A. K., Goel, N., and Bhatia, K.: Comparative study of neural network, fuzzy logic and linear transfer function techniques in daily rainfall-runoff modelling under different input domains, Hydrological Processes, 25, 175-193, 2011.
Loo, Y. Y., Billa, L., and Singh, A.: Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia, Geosci. Front., 6, 817-823, http://dx.doi.org/10.1016/j.gsf.2014.02.009, 2015.
Machado, M. J., Botero, B., López, J., Francés, F., Díez-Herrero, A., and Benito, G.: Flood frequency analysis of historical flood data under stationary and non-stationary modelling, Hydrology and Earth System Sciences, 19, 2561, 2015.
MathWorks: Statistics Toolbox, The MathWorks, Inc., 2012a.
MathWorks: Bioinformatics Toolbox, The MathWorks, Inc., 2012b.
McCuen, R. H.: Modeling hydrologic change: statistical methods, CRC press, 2016.

McLachlan, G., and Peel, D.: Finite Mixture Models, Wiley-Interscience; 1 edition, Printed United States 2000.
McLachlan, G., and Krishnan , T.: The EM Algorithm and Extensions, 2nd Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, New Jersey, USA, 2008.
Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.: Fisher discriminant analysis with kernels, In Proc. of the 1999 IEEE Neural Networks for Signal Processing, Madison, WI , 23 Aug 1999-25 Aug 1999, 41–48, 10.1109/NNSP.1999.788121, 1999.
Mukerji, A., Chatterjee, C., and Raghuwanshi, N. S.: Flood Forecasting Using ANN, Neuro-Fuzzy, and Neuro-GA Models, J. Hydrol. Eng., 14, 647-652, doi:10.1061/(ASCE)HE.1943-5584.0000040, 2009.
Nayak, P. C., Venkatesh, B., Krishna, B., and Jain, S. K.: Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach, Journal of Hydrology, 493, 57-67, http://dx.doi.org/10.1016/j.jhydrol.2013.04.016, 2013.
Nguyen, C. C., Gaume, E., and Payrastre, O.: Regional flood frequency analyses involving extraordinary flood events at ungauged sites: further developments and validations, Journal of Hydrology, 508, 385-396, 2014.
Olivier, C., Jouzel, F., and Matouat, A. E.: Choice of the Number of Component Clusters in Mixture Models by Information Criteria, In Proc. of the Vision Interface '99, May 18-21 1999, Trois-Rivieres, Quebec, Canada, 74 – 81, 1999.
Paalanen, P.: Bayesian classification using Gaussian mixture model and EM estimation: implementations and comparisons, Technical Report, Department of Information Technology, Lappeenranta University of Technology, 2004.
Papaioannou, G., Vasiliades, L., and Loukas, A.: Multi-criteria analysis framework for potential flood prone areas mapping, Water. Resour. Manage., 29, 399–418., 2015.
Pulvirenti, L., Pierdicca, N., Chini, M., and Guerriero, L.: An algorithm for operational flood mapping from Synthetic Aperture Radar (SAR) data using fuzzy logic, Nat. Hazards Earth Syst. Sci., 11, 529-540, 10.5194/nhess-11-529-2011, 2011.
Radmehr, A., and Araghinejad, S.: Developing Strategies for Urban Flood Management of Tehran City Using SMCDM and ANN, J. Comput. Civ. Eng., 28, 05014006, doi:10.1061/(ASCE)CP.1943-5487.0000360, 2014.
Reynaud, A., and Nguyen, M.-H.: Valuing Flood Risk Reductions, Environ. Model. Assess., 21, 603-617, 10.1007/s10666-016-9500-z, 2016.
Rezaeianzadeh, M., Tabari, H., Arabi Yazdi, A., Isik, S., and Kalin, L.: Flood flow forecasting using ANN, ANFIS and regression models, Neural. Comput. & Applic., 25, 25-37, 10.1007/s00521-013-1443-6, 2014.
Sahoo, B., Chatterjee, C., Raghuwanshi, N. S., Singh, R., and Kumar, R.: Flood Estimation by GIUH-Based Clark and Nash Models, J. Hydrol. Eng., 11, 515-525, doi:10.1061/(ASCE)1084-0699(2006)11:6(515), 2006.
Seckin, N., Cobaner, M., Yurtal, R., and Haktanir, T.: Comparison of Artificial Neural Network Methods with L-moments for Estimating Flood Flow at Ungauged Sites: the Case of East Mediterranean River Basin, Turkey, Water. Resour. Manage., 27, 2103-2124, 10.1007/s11269-013-0278-3, 2013a.
Seckin, N., Cobaner, M., Yurtal, R., and Haktanir, T.: Comparison of artificial neural network methods with L-moments for estimating flood flow at ungauged sites: the case of East Mediterranean River Basin, Turkey, Water resources management, 27, 2103-2124, 2013b.
Tehrany, M. S., Pradhan, B., and Jebur, M. N.: Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS, Journal of Hydrology, 512, 332-343, 2014.
Tehrany, M. S., Pradhan, B., and Jebur, M. N.: Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method, Stoch. Environ. Res. Risk. Assess., 29, 1149-1165, 10.1007/s00477-015-1021-9, 2015a.
Tehrany, M. S., Pradhan, B., Mansor, S., and Ahmad, N.: Flood susceptibility assessment using GIS-based support vector machine model with different kernel types, Catena, 125, 91-101, 2015b.
Theodoridis, S., and Koutroumbas, K.: Pattern Recognition, Academic Press, Elsevier Inc., Printed in the United States of America, 2009.
Theodoridis, S.: Machine Learning: A Bayesian and Optimization Perspective, Academic Press, Elsevier, Printed in The United States, 2015.
Tien Bui, D., Le, K.-T., Nguyen, V., Le, H., and Revhaug, I.: Tropical Forest Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, Using GIS-Based Kernel Logistic Regression, Remote Sensing, 8, 347, 2016a.
Tien Bui, D., Nguyen, Q. P., Hoang, N.-D., and Klempe, H.: A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS, Landslides, 1-17, 10.1007/s10346-016-0708-4, 2016b.
Tien Bui, D., Pradhan, B., Nampak, H., Bui, Q.-T., Tran, Q.-A., and Nguyen, Q.-P.: Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS, J. Hydrol., 540, 317-330, http://dx.doi.org/10.1016/j.jhydrol.2016.06.027, 2016c.
Tien Bui, D., Pradhan, B., Nampak, H., Quang Bui, T., Tran, Q.-A., and Nguyen, Q. P.: Hybrid Artificial Intelligence Approach Based on Neural Fuzzy Inference Model and Metaheuristic Optimization for Flood Susceptibility Modelling in A High-Frequency Tropical Cyclone Area using GIS, Journal of Hydrology, 540, 317-330, http://dx.doi.org/10.1016/j.jhydrol.2016.06.027, 2016d.
Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., and Revhaug, I.: Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree, Landslides, 13, 361-378, 10.1007/s10346-015-0557-6, 2016e.

Tien Bui, D., Bui, Q.-T., Nguyen, Q.-P., Pradhan, B., Nampak, H., and Trinh, P. T.: A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area, Agricultural and Forest Meteorology, 233, 32-44, http://dx.doi.org/10.1016/j.agrformet.2016.11.002, 2017.

van Erkel, A. R., and Pattynama, P. M. T.: Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology, Eur. J. Radiol., 27, 88-94, http://dx.doi.org/10.1016/S0720-048X(97)00157-5, 1998.

Wallace, C. S., and Dowe, D. L.: Minimum Message Length and Kolmogorov Complexity, Comput. J., 42, 270-283, 10.1093/comjnl/42.4.270, 1999.

Webb , A. R., and Copsey, K. D.: Statistical Pattern Recognition, John Wiley & Sons, United Kingdom, 2011.

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., and Bouwman, A.: A framework for global river flood risk assessment, Hydrol. Earth. .System. Sci, 17, 1871–1892, 10.5194/hess-17-1871-2013, 2013.

Winsemius, H. C., Aerts, J. C., van Beek, L. P., Bierkens, M. F., Bouwman, A., Jongman, B., Kwadijk, J. C., Ligtvoet, W., Lucas, P. L., and van Vuuren, D. P.: Global drivers of future river flood risk, Nature Climate Change, 2015.

Yu, J.: Localized Fisher discriminant analysis based complex chemical process monitoring, AIChE Journal, 57, 1817-1828, 2011.

Yue, S., Ouarda, T., Bobée, B., Legendre, P., and Bruneau, P.: The Gumbel mixed model for flood frequency analysis, Journal of hydrology, 226, 88-100, 1999.

Zhang, G., Mahfouf, M., Abdulkareem, M., Gaffour, S.-A., Yang, Y.-Y., Obajemu, O., Yates, J., Soberanis, S. A., and Pinna, C.: Hybrid-modelling of compact tension energy in high strength pipeline steel using a Gaussian Mixture Model based error compensation, Appl. Soft Comput., 48, 1-12, http://dx.doi.org/10.1016/j.asoc.2016.06.007, 2016.

Zhou, Z., Liu, S., Zhong, G., and Cai, Y.: Flood Disaster and Flood Control Measurements in Shanghai, Nat. Hazards Rev., Just Released, 10.1061/(ASCE)NH.1527-6996.0000213, 2016.