Geosci. Model Dev., 10, 3391–3409, 2017
https://doi.org/10.5194/gmd-10-3391-2017

Geoscientific
Model Development
Open Access

A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1.1) for spatial prediction of floods

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

[1]Geographic Information System Group, Department of Business and IT, University College of Southeast Norway (USN), Gullbringvegen 36, 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)

Received: 19 December 2016 – Discussion started: 17 January 2017
Revised: 17 July 2017 – Accepted: 10 August 2017 – Published: 14 September 2017

Abstract. In this study, a probabilistic model, named as BayGmmKda, is proposed for flood susceptibility assessment in a study area in central Vietnam. The new model is a Bayesian framework constructed by a combination of a 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 that aims at enhancing the model performance. As a result, the posterior probabilistic 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.

1 Introduction

Flooding is one of the most destructive natural hazards that cause heavy loss of human lives and property in immense spatial extent (Dottori et al., 2016; Komi et al., 2017). Recent statistics on flood damages for the period of 1995–2015 shows that flooding affected 109 million people around the globe per year (Alfieri et al., 2017) and killed more than 220 000 people (Winsemius et al., 2015). Although the frequency of flooding has decreased in several regions (i.e., in central Asia and America), flood occurrences have increased globally by 42 % (Hirabayashi et al., 2013).

Notably, 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). Located in this region, Vietnam is a storm center on 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 by Kreft et al. (2014), from 1994 to 2013, Vietnam endured an annual economic loss that is equivalent to USD 2.9 billion.

Additionally, the occurrences of flood in Vietnam are expected to rise rapidly in the near future due to the increases in poorly planned infrastructure developments and urbanization near watercourses, as well as an increased deforestation and climate change. Hence, an accurate model for evaluat-

Published by Copernicus Publications on behalf of the European Geosciences Union.

ing flood hazards for land-use planning becomes a crucial need for land-use planning as well as establishment of disaster mitigation strategies. Based on flood prediction models, flood-prone areas can be identified and mapped (Tien Bui et al., 2016c).

Needless to say, the identification of susceptible areas can significantly reduce flood damage to the national economy and human lives by avoiding infrastructure developments and densely populated settlements in highly flood-susceptible areas (Zhou et al., 2016). This identification also helps government agencies to issue appropriate flood management policies and to focus its limited financial resources on constructing large-scale flood defense infrastructure in areas that have great economic value but are highly susceptible to flood (Bubeck et al., 2012; Mason et al., 2010). Therefore, a tool for spatial flood 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 stations are available. Therefore, new modeling approaches should be explored and investigated. Given these motivations, this study proposes a novel methodology designed for achieving a high prediction accuracy as well as deriving probabilistic evaluations of flood susceptibility on a regional scale. Accordingly, spatial prediction of flooding is carried out based on a statistical assumption that flooding 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 flood inventories are used to define the class of flood occurrence. Moreover, the class nonflood occurrence is derived from areas that have not yet been damaged by flooding. Consequently, spatial prediction of flooding within the study area is achieved based on the probability of pixels belonging to the class of flood occurrences. To yield probabilistic outputs of flood susceptibility, this study proposes a Bayesian framework established on the basis of an integration of 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); in addition, KFDA constructs a latent variable based on the geoenvironmental 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 tasks and the second level carries out the prediction process to derive the model output of flood evaluation. Based on previous studies which indicate that hierarchical model structures can produce improved prediction accuracy, the proposed framework could 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 Sect. 2. The next section introduces the research method of the current paper, followed by Sect. 4 which describes the proposed Bayesian model for flood susceptibility forecasting. Section 5 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 (Winsemius et al., 2013; Papaioannou et al., 2015; Gao et al., 2017; Alfieri et al., 2014). 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; accordingly, the constructed regression models are used to transform flood information to ungauged basins (Yue et al., 1999; Cunnane, 1988; McCuen, 2016). Thus, these models are capable of providing discharge predictions both in space and time. However, long-term data are not always available; in many cases, they are generally too short for reliable estimations of extreme quantiles (Seckin et al., 2013b; Nguyen et al., 2014).

Rainfall–runoff models, which deal with estimation of runoff from rainfall, are considered to be the most extensively used approach 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 the literature, varying from empirical models to highly sophisticated physical processes. Empirical models could be established based on statistical techniques (Brocca et al., 2011; Neal et al., 2013) or advanced machine learning algorithms (Lohani et al., 2011); such models can be effectively employed to analyze rainfall and runoff on the basis of historical time series data. In addition, physical-process models focus on simulating hydrological processes in a basin based on a set of mathematical equations governing physical processes of water flow and surfaces (Aronica et al., 2012; Chiew et al., 1993; Beven et al., 1984; Birkel et al., 2010; Grimaldi et al., 2013). In general, rainfall–runoff models require relatively long-term time series data at gauging stations. However, the density of gauging stations in developing countries is very low and this fact creates a great obstacle to the establishment of accurate hydrological models (Fenicia et al., 2008). In addition, large-scale field works and deployments of measuring equipment are necessary for collecting data.

In recent years, a new flood modeling approach called “on–off” classification of flood occurrence has been successfully proposed for spatial prediction of flood (or alternatively called a flood susceptibility index; Tien Bui et al., 2016d;

Tehrany et al., 2014, 2015b). Accordingly, no time series data are required for the model calibration, and the establishment of flood models is based on flood inventories (flood class) and nonflood areas (nonflood class). Accordingly, the probability of a pixel in the study area belonging to the flood class is used as flood susceptibility index. Moreover, it is noted that the results of the model depend on the collection of sufficient training data. Although the 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 land-use planning and flood management.

The literature review shows that data-driven methods integrated with GIS databases have demonstrated their effectiveness and accuracy in large-scale flood susceptible predictions. An fuzzy-logic-based algorithm, established by Pulvirenti et al. (2011), has been used to develop a map of flooded areas from synthetic aperture radar imagery; this algorithm is used for the operational flood management system in Italy. 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 was constructed by field surveys and maps of the topography, geology, land cover, and infrastructure.

Prediction models with artificial neural networks (ANNs) 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 that an ANN is 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 machines (SVMs; Hoang and Pham, 2016). Kazakis et al. (2015) introduced a multicriteria index to assess flood hazard areas that relies on GIS and analytical hierarchy processes (AHPs); in this methodology, the relative importance of each flood-influencing factor for the occurrence and severity of flood was determined via AHP. More recently, support-vector-machine-based flood susceptibility analysis approaches have been proposed by Tehrany et al. (2015a, b); 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 neuro-fuzzy inference system, named as MONF, has been introduced by Tien Bui et al. (2016c); this research pointed out that MONF is more capable than decision tree, ANN, SVM, and conventional ANFIS methods.

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 to enrich 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 Fig. 1). This is by far one of the most heavily affected flood regions in the country (Reynaud and Nguyen, 2016). The area of the district is approximately 2803 km^2. The district is located 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 features of the Tuong Duong district are 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.

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 1679–3259 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 effects on the rate of human casualties and economic loss. An examination carried out by Reynaud and Nguyen (2016) reported that approximately 40 % of families have been affected by floods and roughly 20 % of families must be relocated away from the flooded areas; the average loss from flooding is 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 the geoenvironmental conditions (e.g., topological and hydrological features) that 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.

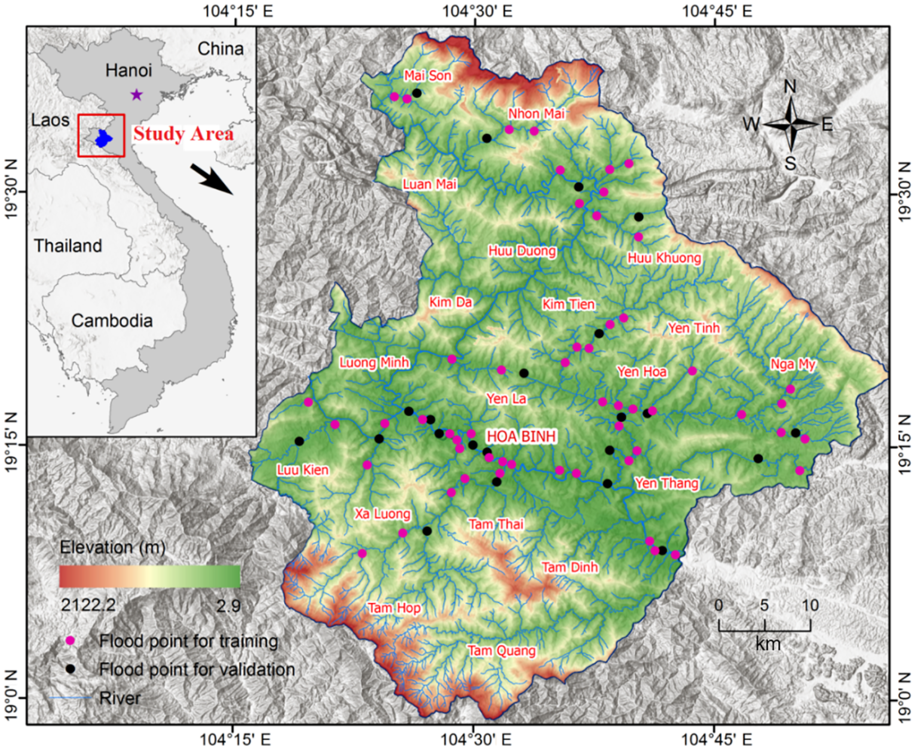

Figure 1. Location of the Tuong Duong district (central Vietnam).

The flood inventory map stores documentations of past flood events (see Fig. 1). It is noted that the type of floods in this study area are flash floods. 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 a resolution of 30 m (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 5 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 are associated with flood points are used to extract the influencing factors used for flood prediction.

Although the data for this study were collected from 2010 to 2014, there were recurrent flash floods which occurred during tropical typhoons in this period. Thus, it is reasonable to conclude that all significant flash flood locations in the study area have been revealed and determined. It should be noted that due to the statistical assumption used in this study, the inclusion of flood locations in the distant past (i.e., before the year of 2009) for flood susceptibility analysis may cause bias. It is because the construction of new hydropower dams such as Ban Ve (from 2010) and Nam Non (from 2011) and deforestation or forestation have changed the geoenvironmental conditions in the study area (Dao, 2017; Manley et al., 2013). In other words, the geoenvironmental conditions of the distant past are very different to those of the present time; therefore, flood locations in the distant past should not be included in the current analysis.

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 proper to note that the selection of the flood-governing factors varies due to different 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 10 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 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 description of the 10 influencing factors of flood occurrence employed in this study

Table 1. Flood-influencing factors and their categories.

Factors	Coding	Description of factor categories
Slope (°)	IF_1	1 (0–0.5); 2 (0.5–2); 3 (2–5); 4 (5–8); 5 (8–13); 6 (13–20); 7 (20–30); 8 (> 30)
Elevation (100 m)	IF_2	1 (< 1); 2 (1–2); 3 (2–3); 4 (3–4); 5 (4–5); 6 (5–6); 7 (6–7); 8 (7–10); 9 (10–13); 10 (> 13)
Curvature	IF_3	1 (< −2); 2 (−2 to −0.05) ; 3 (−0.05–0.05); 4 (0.05–2); 5 (> 2)
Topographic wetness index (TWI)	IF_4	1 (< 6.5); 2 (6.5–7.5); 3 (7.5–8.5); 4 (8.5–9.5); 5 (9.5–10.5); 6 (10.5–11.5); 7 (11.5–12.5); 8 (> 12.5)
Stream power index (SPI)	IF_5	1 (< 1); 2 (1–3); 3 (3–5); 4 (5–7); 5 (7 to10); 6 (10–15); 7 (15–20); 8 (20–30); 9 (30–50); 10 (> 50)
Distance to river (m)	IF_6	1 (< 40); 2 (40–80); 3 (80–120); 4 (120–200); 5 (200–400); 6 (400–700); 7 (700–1500); 8 (> 1500)
Stream density (km km^{-2})	IF_7	1 (< 1); 2 (1–3); 3 (3–5); 4 (5–7); 5 (7–9); 6 (> 9)
Normalized difference vegetation index (NDVI)	IF_8	1 (< 0.3); 2 (0.3–0.35); 3 (0.35–0.4); 4 (0.4–0.45); 5 (0.45–0.5); 6 (0.5–0.55); 7 (0.55–0.6); 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 (1000 mm)	IF_{10}	1 (< 1.82); 2 (1.82–1.92); 3 (1.92–2.02); 4 (2.02–2.12); 5 (2.12–2.22); 6 (2.22–2.32); 7 (2.32–2.42); 8 (> 2.42)

is summarized in Table 1. The distributions of the 10 factors within the studied region are illustrated in Fig. 2.

3.2 Bayesian framework for flood classification

The flood prediction in this study is considered as a pattern classification problem within which "flood" and "nonflood" 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, a 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.

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, in which k is either 1 or 2. $C_1 = 1$ and $C_2 = 0$ denote the flood class and the nonflood class, respectively. To recognize an input pattern based on the information supplied by its feature vector $\boldsymbol{X}$, we need to attain the posterior probability $P(C_k|\boldsymbol{X})$, which indicates the likelihood that the feature vector $\boldsymbol{X}$ 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 $P(C_k|\boldsymbol{X})$ is calculated as follows (Webb and Copsey, 2011):

$$P(C_k|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|C_k) \times P(C_k)}{p(\boldsymbol{X})}, \tag{1}$$

where $P(C_k|\boldsymbol{X})$ denotes the posterior probability. The term $p(\boldsymbol{X}|C_k)$ 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 $p(\boldsymbol{X})$ 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:

$$P(\boldsymbol{X}) = \sum_{k=1}^{M} p(\boldsymbol{X}|C_k) \times P(C_k). \tag{2}$$

Generally, the prior probabilities $P(C_k)$ can be calculated by computing the ratio of training instances in each class. Thus, the bulk of establishing a Bayesian classification model is the calculation of the likelihood $p(\boldsymbol{X}|C_k)$. This likelihood expresses the density of input patterns in the learning space within a certain group of data. In most of situations,

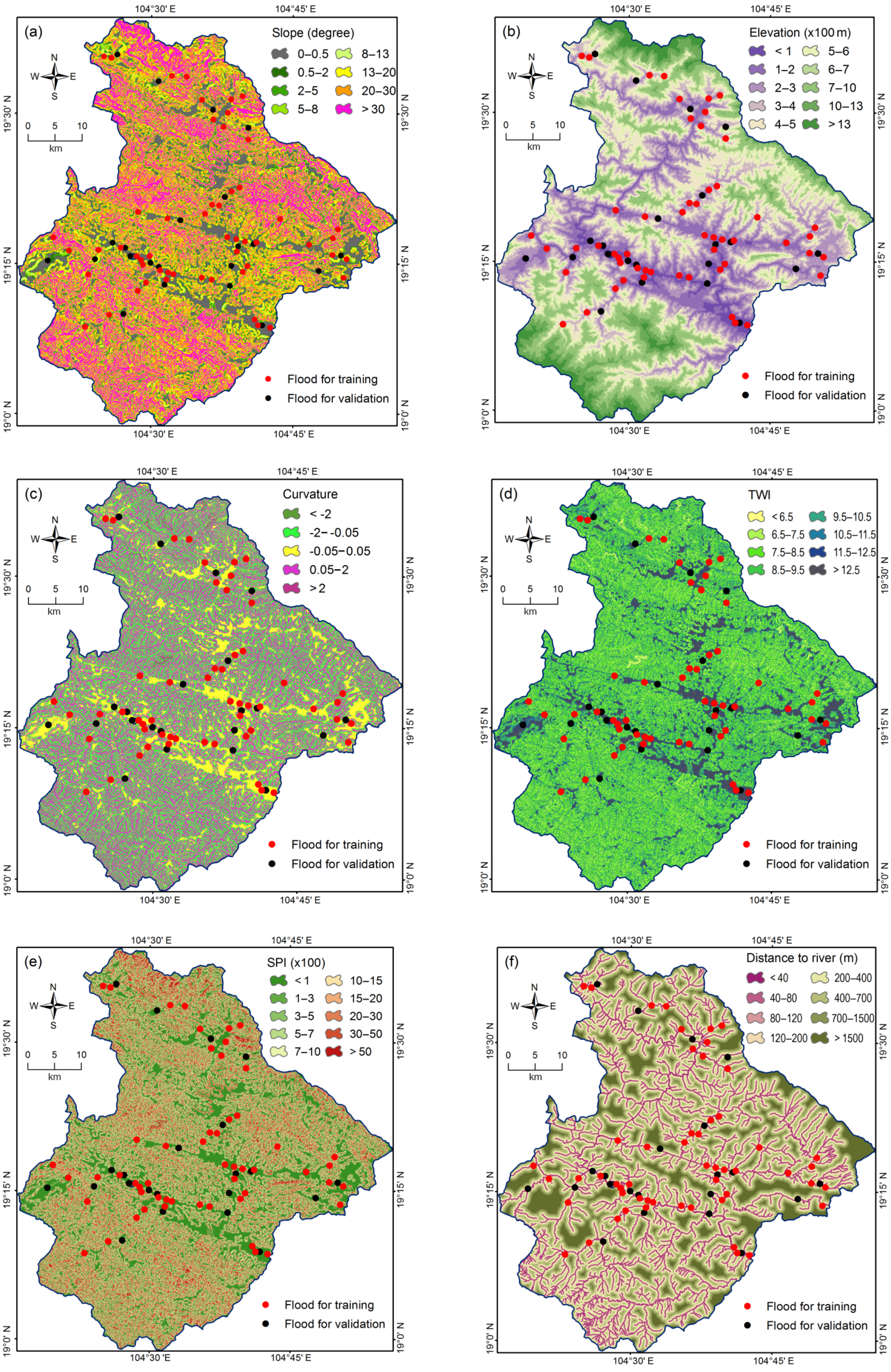
(a)
Slope (degree)
0–0.5
0.5–2
2–5
5–8
8–13
13–20
20–30
> 30
(b)
Elevation (x100 m)
< 1
1–2
2–3
3–4
4–5
5–6
6–7
7–10
10–13
> 13
(c)
Curvature
< -2
-2–-0.05
-0.05–0.05
0.05–2
> 2
(d)
TWI
<6.5
6.5–7.5
7.5–8.5
8.5–9.5
9.5–10.5
10.5–11.5
11.5–12.5
> 12.5
(e)
SPI (x100)
< 1
1–3
3–5
5–7
7–10
10–15
15–20
20–30
30–50
> 50
(f)
Distance to river (m)
< 40
40–80
80–120
120–200
200–400
400–700
700–1500
> 1500
Flood for training
Flood for validation
N
W
E
S
0 5 10
km
104°30′ E
104°45′ E
19°30′ N
19°15′ N
19°0′ N

Figure 2.

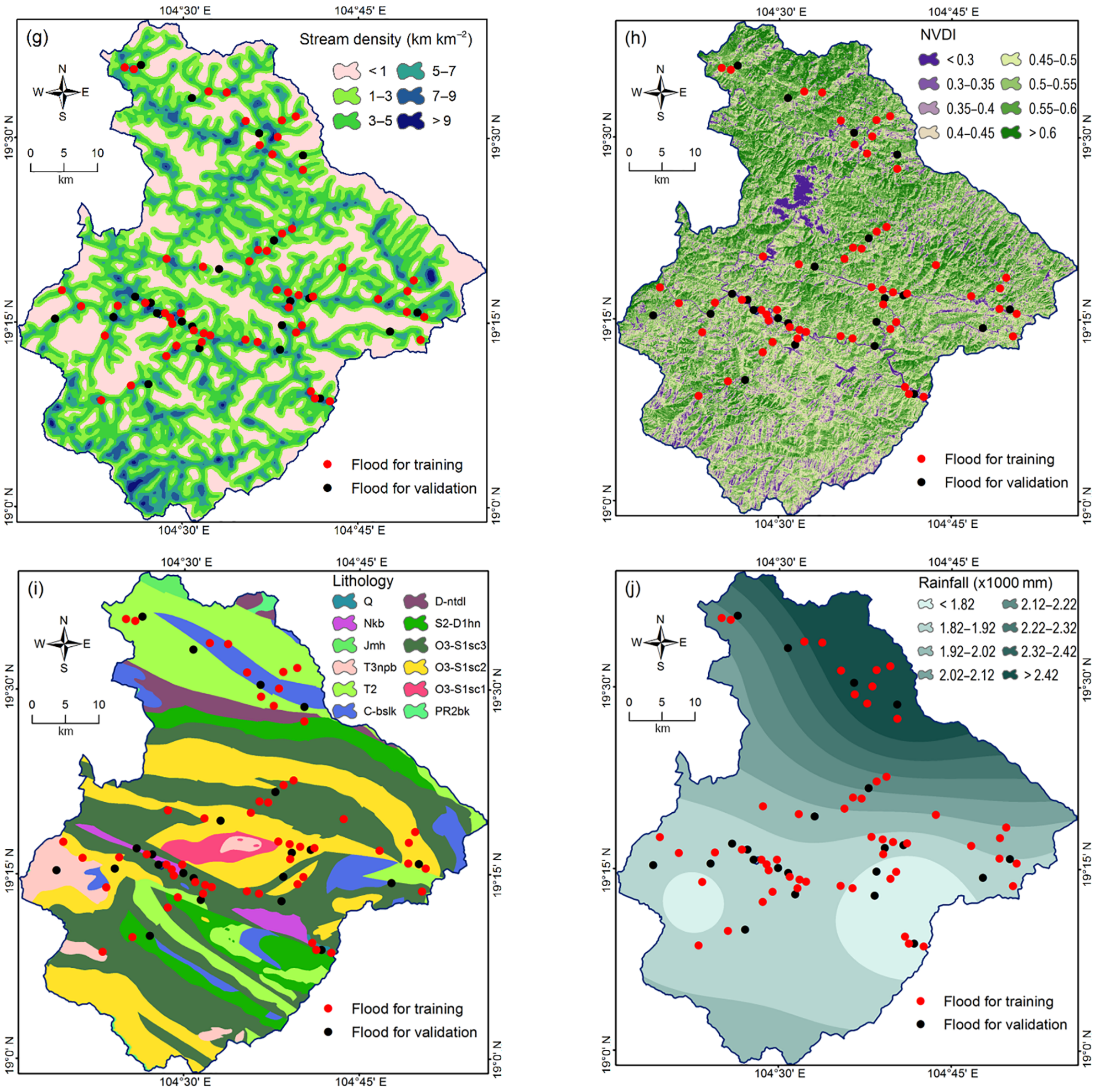

Figure 2. Flood-influencing factors: **(a)** slope, **(b)** elevation, **(c)** curvature, **(d)** topographic wetness index, **(e)** stream power index, **(f)** distance to river, **(g)** stream density, **(h)** normalized difference vegetation index, **(i)** lithology, and **(j)** rainfall.

$p(\boldsymbol{X}|C_k)$ 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 $p(\boldsymbol{X}|C_k)$.

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 PDF must be estimated. This section presents how PDF is estimated by a Gaussian mixture model. A 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 amount of attention from the academic community (Zhang et al., 2016; Khanmohammadi and Chou, 2016; Ju and Liu, 2012).

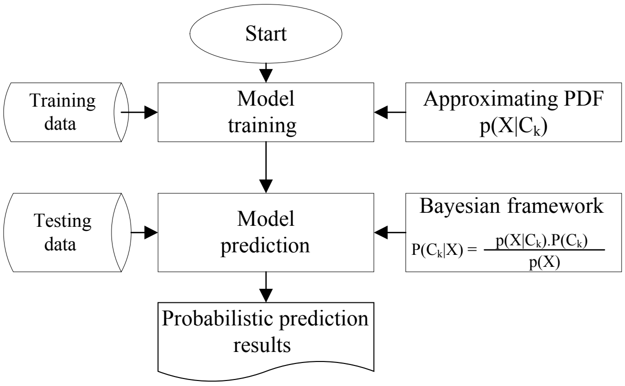
Start
Training data
Model training
Approximating PDF p(X|C_k)
Testing data
Model prediction
Bayesian framework
P(C_k|X) = p(X|C_k).P(C_k) / p(X)
Probabilistic prediction results

Figure 3. General concept of the Bayesian Framework for flood classification.

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

$$N(x|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{3}$$

where μ denotes the vector of variable mean, Σ represents the matrix of covariance, and $\theta = \{\mu, \Sigma\}$ 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 Fig. 4):

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

where $\Theta = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \theta_1, \theta_2, \ldots, \theta_k\}$. $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ is called the mixing coefficients of k Gaussian components and $\sum_{i=1}^{k}\alpha_i = 1$.

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

$$p(\boldsymbol{X}|\Theta) = \prod_{t=1}^{n} p(x_t|\Theta) = L(\Theta|\boldsymbol{X}). \tag{5}$$

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

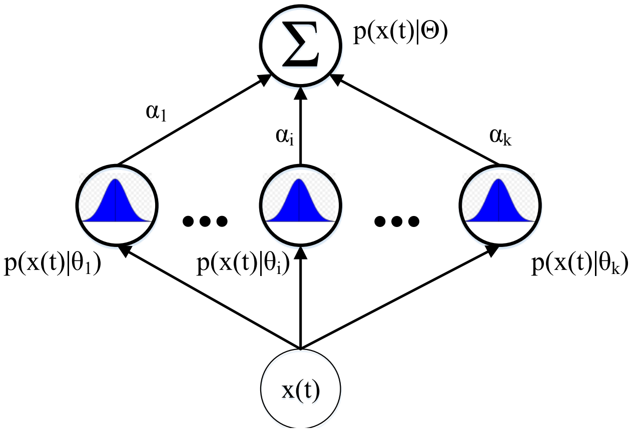

Figure 4. Structure of a Gaussian mixture model.

$$\Theta_e = \arg\max_{\Theta} \log(L(\boldsymbol{X}, \Theta)) = \log\left(\prod_{t=1}^{n} p(x_t|\Theta)\right) = \sum_{t=1}^{n}\log\left(\sum_{i=1}^{k}\alpha_i p_i(x_t|\theta_i)\right). \tag{6}$$

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

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

where w_{it} is a posteriori probability for the ith class, $i = 1, \ldots, k$, and w_{it} satisfies the following conditions:

$$w_{it} = \frac{\alpha_i p_i(x_t|\theta_i)}{\sum_{s=1}^{k}\alpha_s p_s(x_t|\theta_s)} \; ; \; \sum_{i=1}^{k} w_{it} = 1. \tag{8}$$

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 the 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 $p_i(x_t|\theta_i) = N(x_i|\mu_i, \Sigma_i)$ that x_t generated from the ith mixture component are calculated, and the M step within which the maximum likelihood estimates of θ_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:

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

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

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

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 datasets, high sensitivity to initialization condition, and suboptimal 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 through 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 elements 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):

$$\Omega\left(\Theta|\boldsymbol{X}\right) = \tag{12}$$
$$\frac{N}{2}\sum_{i:\alpha_i>0}^{\ln\left(\frac{n\alpha_i}{12}\right)+\frac{C_{\text{nz}}}{2}} \ln\left(\frac{n}{12}\right) + \frac{C_{\text{nz}}(N+1)}{2} - \ln L\left(\boldsymbol{X}, \Theta\right),$$

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 components featuring nonzero weight ($\alpha_i > 0$). Accordingly, the EM method is then utilized to minimize Eq. (12) with a fixed number of C_{nz}.

In detail, the EM algorithm is employed to estimate α_i as follows:

$$\alpha_i^{\text{new}} = \frac{\max\{0, \left(\sum\limits_{t=1}^{n} w_{it}\right) - \frac{N}{2}\}}{\sum\limits_{j=1}^{k} \max\left\{0, \left(\sum\limits_{t=1}^{n} w_{jt}\right) - \frac{N}{2}\right\}}. \tag{13}$$

Accordingly, the parameters μ_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 $\Omega(\Theta|\boldsymbol{X})$ becomes smaller than a preset threshold (e.g., 10^{-5}).

3.4 Radial-basis-function Fisher discriminant analysis for generation of latent variables

In machine learning, the performance of a model may be enhanced if latent variables are 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 $\boldsymbol{w}$ (the projecting vector), it is necessary to maximize the Fisher

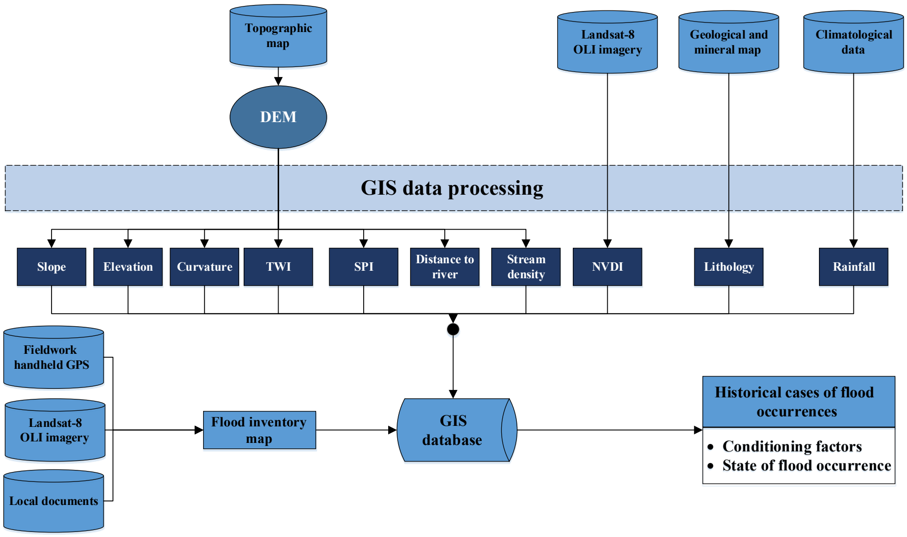

Figure 5. The established GIS database.

discriminant ratio as follows:

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

$$\text{where} S_B^{\varphi} = \left(m_1^{\varphi} - m_2^{\varphi}\right)\left(m_1^{\varphi} - m_2^{\varphi}\right)^T, \tag{15}$$

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

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

To obtain $\boldsymbol{w}$, the kernel trick is applied. Thus, one only needs to establish a formulation of the algorithm which only requires dot-product $\varphi(x) \cdot \varphi(y)$ of the training data and employ kernel functions which calculate $\varphi(x) \cdot \varphi(y)$. The widely employed radial-basis kernel function (RBKF) is expressed in the following formula (with σ denoting 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 $\boldsymbol{w}$ lies in the span of all data samples in the projected space, the transformation vector $\boldsymbol{w}$ is shown in the following formula:

$$\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i \varphi(x_i). \tag{19}$$

From Eqs. (17) and (19), we have the following:

$$\boldsymbol{w}^T m_k^{\varphi} = \frac{1}{Nk}\sum_{j=1}^{N}\sum_{i=1}^{Nk}\alpha_j k\left(x_j, x_i^k\right) = \alpha^T M_k, \tag{20}$$

$$M_k = \frac{1}{Nk}\sum_{i=1}^{Nk} k\left(x_j, x_i^k\right).$$

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

$$\boldsymbol{w}^T S_B^{\varphi} \boldsymbol{w} = \alpha^T M \alpha, \tag{21}$$

where $M = (M_1 - M_2)(M_1 - M_2)^T$.

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

$$\boldsymbol{w}^T S_W^{\varphi} \boldsymbol{w} = \alpha^T N \alpha, \tag{22}$$

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

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

$$J(\alpha) = \frac{\left(\alpha^T M \alpha\right)}{\left(\alpha^T N \alpha\right)}. \tag{23}$$

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 $\mathbf{I}$ is projected on to a line defined by the vector $\boldsymbol{w}$ in the following manner:

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

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 Fig. 5) within which locations of past flood events, maps of topographic feature, Landsat-8 imagery, maps of geological features, 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.

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 nonflood. 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 nonflood 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 Fig. 6.

Firstly, the whole dataset, including 152 data samples, was separated into two sets: a training set (90 % or 137 samples), employed for model establishing, and a 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 minimum–maximum 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 Sect. 3.4) and added to the training dataset, with the aim of enhancing 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 nonrelevant factor should be eliminated from the modeling process to reduce noise and enhance the model performance (Tien Bui et al., 2016a, 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 factors 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 (σ) 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 number of mixture components (k) indicates a lesser 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 σ, the grid search procedure is performed and the parameter σ corresponding to the highest classification accuracy rate was selected.

Using the best k and σ 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 the coupling of the 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 Mário A. T. Figueiredo (http://www.lx.it.pt/~mtf/, last access: 1 April 2016). 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 Fig. 7) for the implementation of the BayGmmKda model has been coded in a MATLAB environment by the authors. The GUI development aims at providing a user-friendly system for performing flood susceptibility predictions.

As shown in Fig. 7, the 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,

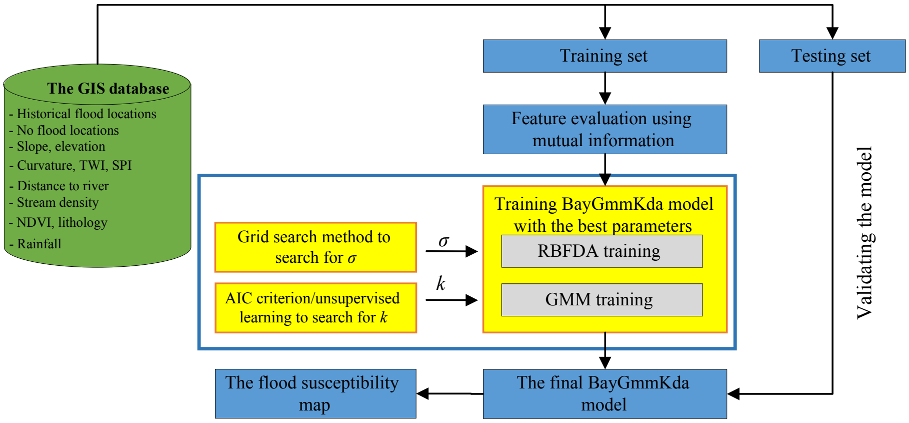
The GIS database
- Historical flood locations
- No flood locations
- Slope, elevation
- Curvature, TWI, SPI
- Distance to river
- Stream density
- NDVI, lithology
- Rainfall
Training set
Testing set
Feature evaluation using mutual information
Training BayGmmKda model with the best parameters
RBFDA training
GMM training
Grid search method to search for σ
AIC criterion/unsupervised learning to search for k
σ
k
Validating the model
The flood susceptibility map
The final BayGmmKda model

Figure 6. The proposed BayGmmKda.

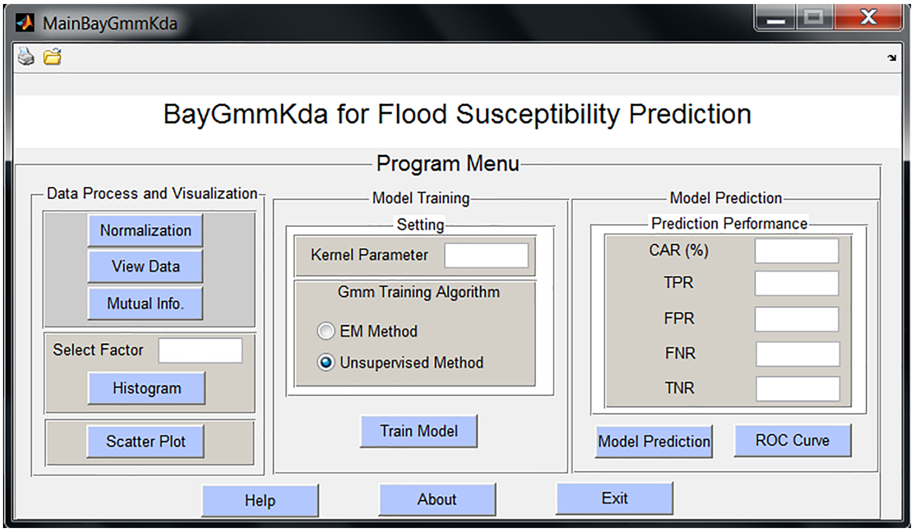

Figure 7. Main menu of BayGmmKda.

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 Fig. 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 are deemed to be relevant and should be retained for the subsequent processes of model training and prediction.

It is worth keeping in mind 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 10 input factors and with the RBFDA-based latent factor.

The 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 pre-

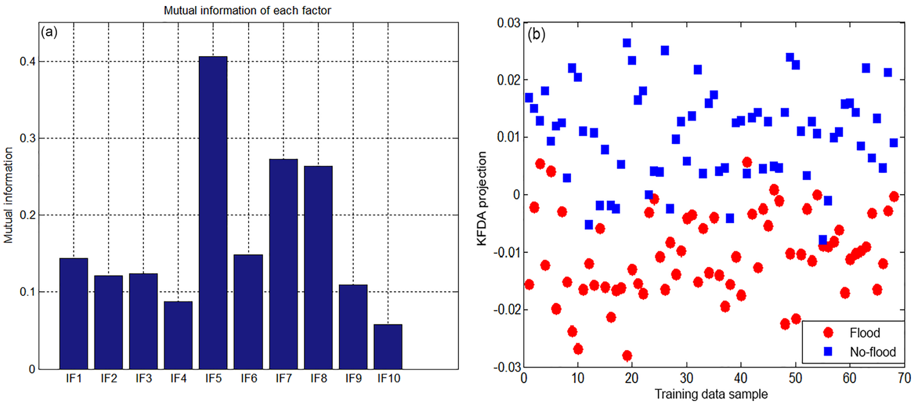

Figure 8. (a) Mutual information of flood-influencing factors; **(b)** RBFDA-based latent factor derived in this study.

sented 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).

$$\begin{aligned}\mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}},\\ \mathrm{FNR} &= \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}},\\ \mathrm{TNR} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\end{aligned} \tag{25}$$

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 rates, the receiver operating characteristic (ROC) curve (van Erkel and Pattynama, 1998) is used to summarize 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 the AUC.

As aforementioned, the dataset 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 subsampling procedure within which 30 experimental runs are carried out. In each experimental run, 10 % of the dataset is retrieved in a random manner from the database to constitute the testing set; the rest of the database is included in the training set.

Table 2. Prediction results of BayGmmKda.

Dataset	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

The testing performance of the proposed Bayesian framework for flood susceptibility is reported in Table 2 and Fig. 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 (UL) algorithm for training GMM are denoted as BayGmmKda-EM and BayGmmKda-UL, respectively. It 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 is deemed more accurate than the BayGmmKda-EM model when the two models predict samples with the nonflood class.

5.2 Model comparison

Because this is the first time the BayGmmKda model has been proposed for the measurement flood susceptibility, the validity of the proposed model should be assessed. Hence, the benchmarks were used for the comparison, including the support vector machine, adaptive neuro-fuzzy inference system, and the GMM-based Bayesian classifier. The above

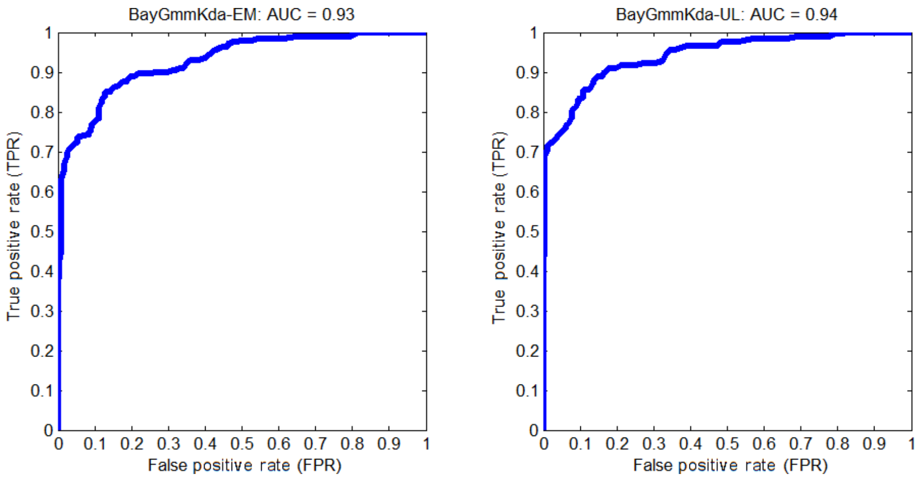

Figure 9. ROC plots of the proposed BayGmmKda.

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). It is noted that the GMM-based Bayesian classifier (BayGmm) is the Bayesian framework for classification which employs GMM for density estimation; however, BayGmm is not integrated with the RBFDA algorithm. BayGmm is used in the performance comparison section 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 (σ) 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-tune such hyperparameters of the SVM model. It is noted that the SVM method is implemented in a MATLAB package (MathWorks, 2012b). Meanwhile, the ANFIS model is trained with the metaheuristic approach described in the previous work of Tien Bui et al. (2016c).

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).

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 were 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.

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

Table 4. Model comparison based on the Wilcoxon signed-rank test.

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

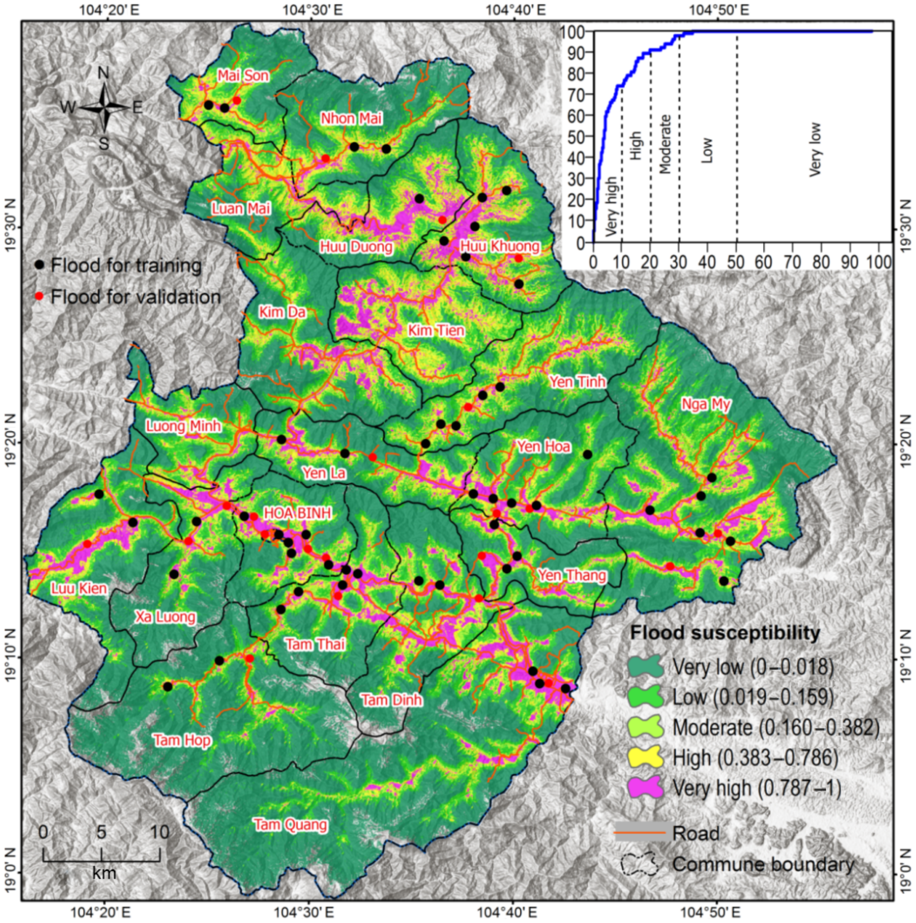

Figure 10. The flood susceptibility map using the proposed BayGmmKda model for the study area.

5.3 Construction of the flood susceptibility map

Experimental outcomes have indicated that the BayGmmKda model is the best for this study area, and 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 Fig. 10) was derived and visualized by the 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 Fig. 10) was constructed and the threshold values were derived.

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. In particular, 50 % of the district, which is categorized to the very low class, contains no flood location. These results 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 in central Vietnam. The newly constructed model is a Bayesian framework that combines GMM and RBFDA for spatial prediction of flooding. 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 out 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 a 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 geoenvironmental data are necessary for the model construction phase. This data collection and analysis 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 datasets. Furthermore, the outcome of this grid search procedure may not be optimal; therefore, more advanced model selection approaches, i.e., metaheuristic 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.

Code and data availability. The MATLAB code of the BayGmmKda model is given in the Supplement.

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

The Supplement related to this article is available online at https://doi.org/10.5194/gmd-10-3391-2017-supplement.

Competing interests. The authors declare that they have no conflict of interest.

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 Quoc-Phi Nguyen (Hanoi University of Mining and Geology, Vietnam).

Edited by: Jeffrey Neal
Reviewed by: two anonymous referees

References

Akaike, H.: A new look at the statistical identification model, IEEE T. Automat. Contr., 19, 716–723, https://doi.org/10.1109/TAC.1974.1100705, 1974.

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P., and Feyen, L.: Advances in pan-European flood hazard mapping, Hydrol. Process., 28, 4067–4077, 10.1002/hyp.9947, 2014.

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, https://doi.org/10.1016/j.patcog.2016.01.017, 2016.

Aronica, G. T., Franza, F., Bates, P. D., and Neal, J. C.: Probabilistic evaluation of flood hazard in urban areas using Monte Carlo simulation, Hydrol. Process., 26, 3962–3972, https://doi.org/10.1002/hyp.8370, 2012.

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, Environ. Model. Softw., 76, 20–36, 2016.

Beven, K. J., Kirkby, M. J., Schofield, N., and Tagg, A. F.: Testing a physically-based flood forecasting model (TOPMODEL) for three U.K. catchments, J. Hydrol., 69, 119–143, https://doi.org/10.1016/0022-1694(84)90159-8, 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 An., 41, 561–575, https://doi.org/10.1016/S0167-9473(02)00163-9, 2003.

Birkel, C., Tetzlaff, D., Dunn, S. M., and Soulsby, C.: Towards a simple dynamic process conceptualization in rainfall–runoff models using multi-criteria calibration and tracers in temperate, upland catchments, Hydrol. Process., 24, 260–275, https://doi.org/10.1002/hyp.7478, 2010.

Brocca, L., Melone, F., and Moramarco, T.: Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting, Hydrol. Process., 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, https://doi.org/10.1111/j.1539-6924.2011.01783.x, 2012.

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, Int. J. Appl. Earth Obs., 48, 163–173, 2016.

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, https://doi.org/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, J. Hydrol., 147, 1–36, https://doi.org/10.1016/0022-1694(93)90073-I, 1993.

Cunnane, C.: Methods and merits of regional flood frequency analysis, J. Hydrol., 100, 269–290, 1988.

Dao, N.: Reflecting on the role of academics–activists in shifting hydropower narratives in Vietnam, Crit. Asian Stud., 49, 444–447, https://doi.org/10.1080/14672715.2017.1339450, 2017.

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, https://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, W01402, https://doi.org/10.1029/2006WR005563, 2008.

Figueiredo, M. A. T. and Jain, A. K.: Unsupervised learning of finite mixture models, IEEE T. Pattern Anal., 24, 381–396, https://doi.org/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, J. Hydrol., 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, https://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, J. Hydrol., 487, 39–47, https://doi.org/10.1016/j.jhydrol.2013.02.023, 2013.

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S.: Global flood risk under climate change, Nat. Clim. 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, https://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, https://doi.org/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, https://doi.org/10.1016/j.asoc.2016.04.031, 2016.

Ju, Z. and Liu, H.: Fuzzy Gaussian Mixture Models, Pattern Recognit., 45, 1146–1158, https://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, https://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, https://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, https://doi.org/10.1007/s12665-011-1504-z, 2012.

Komi, K., Neal, J., Trigg, M. A., and Diekkrüger, B.: Modelling of flood hazard extent in data sparse areas: a case study of the Oti River basin, West Africa, J. Hydrol., 10, 122–132, https://doi.org/10.1016/j.ejrh.2017.03.001, 2017.

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 T. Pattern Anal., 24, 1667–1671, https://doi.org/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, Int. Geosci. Remote Se., 895–898, 2012.

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, Hydrol. Process., 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, https://doi.org/10.1016/j.gsf.2014.02.009, 2015.

Machado, M. J., Botero, B. A., López, J., Francés, F., Díez-Herrero, A., and Benito, G.: Flood frequency analysis of historical flood data under stationary and non-stationary modelling, Hydrol. Earth Syst. Sci., 19, 2561–2576, https://doi.org/10.5194/hess-19-2561-2015, 2015.

Manley, P. N., Mortenson, L., Halperin, J. J., and Quyen, N. H.: Options for monitoring forest degradation in Northern Viet Nam: An assessment in systems design and capacity building needs in Con Cuong District, Nghe An Province, USAID Asia Final Report, 2013.

Mason, D. C., Speck, R., Devereux, B., Schumann, G. J. P., Neal, J. C., and Bates, P. D.: Flood Detection in Urban Areas Using TerraSAR-X, IEEE T. Geosci. Remote, 48, 882–894, https://doi.org/10.1109/TGRS.2009.2029236, 2010.

MathWorks: Statistics Toolbox, The MathWorks, Inc., 2012a.

MathWorks: Bioinformatics Toolbox, The MathWorks, Inc., 2012b.

McCuen, R. H.: Modeling hydrologic change: statistical methods, CRC press, 448 pp., 2016.

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.

McLachlan, G. and Peel, D.: Finite Mixture Models, Wiley-Interscience, 1st Edn., Printed United States 2000.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.: Fisher discriminant analysis with kernels, Proceedings of the 1999 IEEE Neural Networks for Signal Processing, Madison, WI, 23–25 August 1999, 41–48, https://doi.org/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, https://doi.org/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, J. Hydrol., 493, 57–67, https://doi.org/10.1016/j.jhydrol.2013.04.016, 2013.

Neal, J., Keef, C., Bates, P., Beven, K., and Leedal, D.: Probabilistic flood risk mapping including spatial dependence, Hydrol. Process., 27, 1349–1363, https://doi.org/10.1002/hyp.9572, 2013.

Nguyen, C. C., Gaume, E., and Payrastre, O.: Regional flood frequency analyses involving extraordinary flood events at ungauged sites: further developments and validations, J. Hydrol., 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, Proceedings of the Vision Interface'99, Trois-Rivieres, Quebec, Canada, 18–21 May 1999, 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, https://doi.org/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, https://doi.org/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. Appl., 25, 25–37, https://doi.org/10.1007/s00521-013-1443-6, 2014.

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, https://doi.org/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 Resour. Manage., 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, J. Hydrol., 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. Env. Res. Risk A, 29, 1149–1165, https://doi.org/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, https://doi.org/10.1016/j.catena.2014.10.017, 2015b.

Theodoridis, S.: Machine Learning: A Bayesian and Optimization Perspective, Academic Press, Elsevier, USA, 2015.

Theodoridis, S. and Koutroumbas, K.: Pattern Recognition, Academic Press, Elsevier Inc., USA, 2009.

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 Sens., 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, 14, 1–17, https://doi.org/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, https://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, J. Hydrol., 540, 317–330, https://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, https://doi.org/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, Agr. Forest Meteorol., 233, 32–44, https://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, https://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, https://doi.org/10.1093/comjnl/42.4.270, 1999.

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

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., and Bouwman, A.: A framework for global river flood risk assessments, Hydrol. Earth Syst. Sci., 17, 1871–1892, https://doi.org/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, Nat. Clim. Change, 6, 381–385, 2015.

Yu, J.: Localized Fisher discriminant analysis based complex chemical process monitoring, AICHE J., 57, 1817–1828, 2011.

Yue, S., Ouarda, T., Bobée, B., Legendre, P., and Bruneau, P.: The Gumbel mixed model for flood frequency analysis, J. Hydrol., 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, https://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., 18, B5016001, https://doi.org/10.1061/(ASCE)NH.1527-6996.0000213, 2016.