Natural Hazards (2022) 111:2407–2429
https://doi.org/10.1007/s11069-021-05142-5

ORIGINAL PAPER

Assessment of flood-risk areas using random forest techniques: Busan Metropolitan City

Jihye Ha[1] · Jung Eun Kang[1]

Received: 31 December 2020 / Accepted: 24 November 2021 / Published online: 7 January 2022

Abstract
Climate change increases both the risks and effects of flooding in urban areas, which, without mitigation, may lead to social catastrophes. In Korea, devastating typhoons and overflows account for approximately 90% of the country's natural disasters, and the many man-made features of urban environments exacerbate the detrimental effects whenever flooding occurs. Many regression analysis methods exist for assessing geographical flood risk; furthermore, a handful of machine learning methods have been created for mitigation and estimation purposes—there are none for prevention. Therefore, in this study, we developed a machine learning flood assessment model that leverages several machine learning models for the Busan Metropolitan City. Each was applied to a test dataset, and their performances were evaluated based on accuracy, sensitivity, specificity, and area under the curve; thereafter, the model determined to be the most reliable was used to create a flood risk assessment map. The model was then used to assess the areas of highest probability of flooding. Upon its completion, we discovered that flooding may now occur with less rainfall than that of the 10-year return period. The derived map is expected to be used as a basic source for the development of preventive countermeasures against urban flooding, thus contributing to the enhancement of flood control and response capacities in applicable regions.

Keywords Machine learning · Prevention · Risk assessment · South Korea · Urban flooding

1 Introduction

With the increasing severity of climate change worldwide, our planet suffers from frequent and intense natural disasters. In Korea, devastating typhoons and flooding events account for approximately 90% of the country's natural disasters (Han and Park 2014). In urban areas with concentrated populations and advanced residential and industrial facilities, such disasters are exacerbated by many man-made factors, leading to chaos and great distress in and around urban infrastructures and facilities (Sim2008; Song 2012). Urban flooding damages human life, human health, and property from river inundation or inland flooding caused by the inadequacy or failure of sewer pipes and other drainage systems (Sim 2008).

✉ Jung Eun Kang
jekang@pusan.ac.kr

[1] Department of Urban Planning and Engineering, Pusan National University, Busan, South Korea

Recent cases of damage caused by urban flooding have been compounded due to the combined effects of the increase in extreme rainfall events from climate change, local heavy rainfall events with increased intensity in spatial and temporal terms, and urbanization (Park et al. 2007; Son et al. 2010). High-density development of urban built-up areas, increases in the area of impermeable layers caused by urban development, and vulnerable man-made spaces (e.g., subways and underground facilities) aggravate flood damage; this is further worsened by climate change. However, the damage can be mitigated or prevented through effective and comprehensive urban planning and disaster prevention strategies (Alexander 1993; Birkmann 2006).

Measures for preventing urban flooding can be classified as structural and nonstructural. In Korea, a country with a heavily concentrated population in relatively small areas, structural solutions mainly based on disaster prevention have been applied. However, with approaches centered on facilities, there is a limitation in the capacity for disaster prevention caused by heavy rainfall that exceeds design criteria. To overcome such limitations, there is now an increasing emphasis on integrative approaches that incorporate non-structural solutions. Although preventive strategies have been recommended and applied (e.g., restricted development in flood-risk areas) (Shim 2011; Lee and Kim 2015; Lee and Kang 2018), their actual application has been difficult in urban areas where high-density development has already taken place, especially when facing concerns over property rights. In extant flood-risk areas, structural solutions (e.g., disaster prevention facilities and infrastructure reinforcements) and non-structural solutions (e.g., disaster prevention districts and disaster insurance) have been emphasized; however, these measures are still insufficient. To minimize the casualties and property damage from recurring flood events, preemptive measures, risk identification, and areal management are required. Furthermore, public awareness must be increased to influence policy. In this regard, more reliable and detailed flood risk information is needed, and the development of a flood risk map that facilitates intuitive flood risk assessment is required.

Various studies have evaluated flood risks. However, owing to the nature of flood damage in which multiple factors have complex effects, the development of reliable predictive modeling has been difficult. In Korea, the number of days with more than 50 mm of precipitation per hour increased from an average of 5.1 in the 1970s to 12.3 in the 2000s. In the 2000s, the risk of urban flooding increased because of increasing numbers of guerrilla torrential rains, i.e., a significant amount of rain in a specific area in a short period of time. Additionally, interest in urban flooding has increased, leading to its emergence as a key issue since the 2000s and resulting in several studies on the development of elaborate flood analysis technologies utilizing meteorological, spatial information, and hydrological techniques. Although a considerable amount of data has been accumulated, and necessary technologies have been further developed, they have limitations in that they focus on topographical analysis of floods in lowlands, difficulties in data collection, and handling excessive number of parameters (Lee SE et al., 2016). Previous domestic and international studies on the evaluation of the vulnerability and risk of urban flooding mainly employed hydrological approaches, stochastic methods, and geospatial information service (GIS) techniques (Cançado et al. 2008; Ahmad et al. 2013; Son et al. 2013; Ouma et al. 2014; Lee et al. 2016; Lee and Kang 2018). Dutta and Herath (2001) developed and proposed a GIS-based flood loss estimation method and applied it to the Ichinomiya River basin located in Japan. This method is composed of a physically based flood inundation model and loss estimation model as an integrated model that can estimate real-time losses and losses due to floods for the past and future. Herath et al. (2003) developed a method for estimating flood damage in urban watersheds using RS (remote sensing) and GIS and applied it to the Chiba Prefecture, Japan. In their study, an intensive statistical method was applied according to the administrative unit area and land cover data obtained from satellite images using the water level–damage function estimated from the GIS, disaster flood area, inundation depth, and past flood data. Buchele et al. (2006) presented a multipronged approach that can improve the risk assessment method for the existing extreme events in Germany using the water level–damage curve method and GIS. Carlos and Tucci (2007) analyzed the impact of flooding on urban development to conduct a study on urban flood management in Brazil. Tsakiris (2014) applied the concept of expected annual damage, with the view that vulnerability changes according to the size of the risk, to evaluate the risk of flooding in Rapentoza, Greece.

However, recent overseas research trends have shown the application of machine learning techniques, which have been reported to have superior predictive performance compared to those of linear regression (Choi et al. 2018). For identification and evaluation of flood- and inundation-risk areas, multiple criteria decision-making techniques, including analytical hierarchy processes and expert evaluations, have been applied (Chen et al. 2011; Tang et al. 2018; Vojtek et al. 2019). Recent studies have reported the application of various machine learning models, such as artificial neural networks (Kia et al. 2012; Zhao et al. 2018), support vector machines (SVMs) (Tehrany et al. 2014, 2015), decision trees (DTs) (Tehrany et al. 2013), naïve Bayes (NB) (Khosravi et al. 2019; Liu et al. 2016; Chen et al. 2020), and random forests (RFs) (Chapi et al. 2017; Rahmati et al. 2017; Hong et al. 2018; Chen et al. 2020) to the identification of flood-risk area identification and assessment using large amount of data.

There have been few machine learning studies on urban flooding in Korea. Jang et al. (2009) employed DT modeling and performed vulnerability analyses per watershed for South Korea using GIS data. As a result, they established and evaluated preventive measures against extreme rainfall events. Lee (2017) applied RF and boosted tree models to analyze flood and landslide vulnerabilities in Seoul Metropolitan City. Choi et al. (2018) developed a heavy rainfall damage assessment function for the Seoul metropolitan area using an RF model with an SVM.

This study aims to develop a highly reliable flood assessment model using machine learning techniques and big data for urban flood assessment and to construct a flood risk assessment map using the developed model. The research procedure is outlined as follows. First, the extant research is reviewed, and variables used in this study are derived. Next, data on variables are collected for application to the study area (i.e., Busan Metropolitan City), and a spatial information system is developed via GIS analysis. DT, RF, SVM, and NB techniques are applied to develop a machine-learning-based flood assessment function based on data over 3 years (i.e., 2014, 2015, and 2016) for the derived variables. For each model, accuracy, receiver operating characteristic (ROC) curve creation, and area-under-the-curve (AUC) calculation are performed to determine the reliability of the results, and the model determined to be the most reliable is used to create a flood risk assessment map. The map thus derived is expected to be used as a basic source for the development of countermeasures against urban flooding, thus contributing to the enhancement of flood control and response capacities in applicable regions.

However, in Korea, most of the existing research is being conducted in the metropolitan area (Seoul Metropolitan City). In fact, there are few studies on Busan, where casualties and property damage due to flooding occur periodically. Therefore, it is differentiated from existing studies in terms of the construction of a flood risk assessment map of Busan, which exhibits a characteristic topographical vulnerability to natural disasters. In addition, in constructing the flood assessment model, by selecting the final model after comparative analysis using four machine learning techniques, the rationality and objectivity of the analysis method were secured, and the reliability of the analysis results was secured by verifying various results. Furthermore, the precision in the analysis was increased by using grid-type data of 30 × 30 m^2, which is smaller than the administrative district.

2 Methods

2.1 Study area

The study area, Busan Metropolitan City, has topographical characteristics vulnerable to natural disasters, including many hilly mountains and lowlands around rivers and seas. According to the analysis of Lee et al. (2018), Busan Metropolitan City not only has a relatively high intensity of heavy rainfall compared with other regions, but it also has a significant proportion of developed land. Thus, when flooding occurs, the damage may quickly spread over extensive areas. Furthermore, the city has been classified as an area having a high risk of devastating flood damage owing to its large number of facilities inside flood plains, its high proportion of lowlands and old buildings, and the concentrated distribution of population. According to the inundation trace information data provided by the Land and Geospatial Informatix Corporation, a total of 145 areas have had flooding incidents of depths 0.3 m or during the past 11 years (2009–2019), most of which were subject to recurrent damage for every heavy rainfall or typhoon. Because the flood damage tends to recur yearly, Busan Metropolitan City has designated flooding-risk districts (i.e., one in Yeongdo-gu, three in Buk-gu, and two in Gangseo-gu), resulting in a total of six areas as of 2020. They are constantly monitored. Nevertheless, in July 2020, intensive heavy rainfall of up to 87 mm/h persisted, resulting in 313 damage cases, comprising 182 cases of building flooding and fracture, six cases of sewer pipe failure and sewer backflow, 61 cases of inundation of roads and bridges, and 64 cases of damage to other facilities, as well as human life casualties.

In addition, according to the Representative Concentration Pathway (RCP) 6.0 scenario, in which the greenhouse gas policy is to some extent realized in the Busan Metropolitan City, the average annual precipitation for the past 10 years (2001–2010) is projected to increase by 1.8% to 1560.5 mm from 1,532 mm by 2050. It is predicted to be 1522.7 mm based on the RCP 2.6 scenario that the Earth itself can recover the effects of human activities (Korea Meteorological Administration 2017). Precipitation is expected to increase in all areas of Busan Metropolitan City; hence, preemptive measures are necessary. The examination of flood damage in Busan Metropolitan City from past to present shows that damage is caused by the combined effects of changes in rainfall characteristics related to climate change, the increase in impermeable areas caused by urban development, aging of flood protection facilities, and insufficient water movement capacity, indicating that there is a pressing need for measures customized for the urban characteristics of Busan Metropolitan City.

2.2 Selection of variables and data collection for development of flood assessment model

Prior to developing a flood assessment model in this study, variables affecting flooding were selected by reviewing existing research in Korea and overseas. Factors affecting the occurrence of urban flooding were categorized into hydrologic, socioeconomic, facility, geographical, meteorological, and response measure factors (Smith 1994; Penning-Rowsell et al. 1999; USACE 1996; Nicholas et al. 2001; Kelman and Spence 2004). Based on the factors shown in Table 1, the dependent variable of the analysis model (i.e., *flooded area*), which directly shows the flood damage caused by flooding, was selected. The independent variables were classified into *climate exposure*, *geographical*, *development*, *facility*, and *urban flooding risk* factors, as derived.

Urban flooding shows highly complex patterns in contrast to those in non-urban areas. During a rainfall event, rainwater infiltration does not occur in the impermeable layers of most urban ground surfaces; hence, immediate surface runoff is generated. This runoff flows from higher to lower altitudes, and some of the runoff water is discharged through the sewer pipe network. Finally, the runoff flows to the outlet points of drainage pipes, and if the outlet point is lower than the water level of the river, it is drained through the pump of a drainage pumping station (Lee et al. 2019). During this process, in the case of heavy rainfall in which the precipitation temporarily exceeds the peak capacity of the drainage pipe, sewer backflow occurs in the connected pipes in the downstream region, leading to area flooding (Falconer et al. 2009; Golding 2009). The following variables were selected in consideration of the pattern of inundation in urban areas and in reference to previous studies.

Climate exposure factors are those reflecting weather conditions, which include the *maximum hourly precipitation* (Seo et al. 2016; Kim and Kim 2018), *maximum daily precipitation* (Son et al. 2011; Kang and Lee 2012; Lee et al. 2016; Kim and Kim 2018; Lee and Kang 2018), and *days over 80 mm precipitation* (Kang and Lee 2012). Geographical factors including *altitude* (Tehrany et al. 2014; Sowmya et al. 2015; Bui et al. 2016; Marconi et al. 2016; Rahmati et al. 2017; Youssef et al. 2016), *slope* (Pradhan 2010; Tehrany et al. 2014; Bui et al. 2016; Khosravi et al. 2016; Marconi et al. 2016; Seekao and Pharino 2016; Youssef et al. 2016), and *distance from the river* (Bui et al. 2016; Khosravi et al. 2016; Marconi et al. 2016; Youssef et al. 2016) were considered. For development factors, in consideration of indiscriminate urban development and concentration, *soil drainage* (Tehrany et al. 2014; Seekao and Pharino 2016; Youssef et al. 2016) and *impermeable area* (Lee and Kang 2018) were selected. For facility factors, considering the process of urban flooding occurrence, *length of sewer pipe* (Lim et al. 2010), *distance from detention reservoir*, and *number of drainage systems* were selected.

For data collection, flood damage data in the smallest possible spatial unit within the spatiotemporal range were collected. Inundation trace data of the Busan area have been collected since 2009, and according to the database constructed thus far, the flooded areas included 42 regions of 1.385 km^2 in 2009, 18 regions of 0.811 km^2 in 2011, 11 regions of 0.247 km^2 in 2012, 201 regions of 8.89 km^2 in 2014, 6 regions of 0.322 km^2 in 2015, 32 regions of 0.642 km^2 in 2016, 12 regions of 0.156 km^2 in 2017, and 9 regions of 0.653 km^2 in 2019. Considering these flooded areas and the size of the areas with reference to 2014 (i.e., the year having the most significant damage), data for years 2014, 2015, and 2016 corresponding to three consecutive time series were constructed as the datasets in this study. For the analysis unit, considering the spatial analysis and mapping, 30×30 m grid cells were used. The study area was resampled with approximately 900,000 grid cells, and the average values were used for the raster data included in the grid cells. The 30×30 m dimension is most frequently used in GIS grid systems; hence, they can be utilized in the establishment of various geographic information-based disaster prevention measures (Kang and Lee 2015; Kim et al. 2015).

Table 1 Analysis variables used in assessment of flood-risk areas

Classification	Indicators	Variable name	Source
Dependent variable			
Experience factor	Flooded area	flooding_14, 15, 16	National Spatial Data Information Portal
Independent variable			
Exposure factor	Days over 80 mm precipitation	x80_14, 15, 16	Korea Meteorological Administration
	Maximum time precipitation	maxT_14, 15, 16	
	Maximum daily precipitation	maxD_14, 15, 16	
Geographical factor	Altitude	Altitude	National Geographic Information Institute
	Slope	Slope	
	Distance from the river	river_dis	National Spatial Data Information Portal, City of Busan
Development factor	Soil drainage	soil_drain	National Spatial Data Information Portal, Rural Development Administration
	Impermeable area	impermeable_area	Environmental Geographic Information Service
Facility factor	Distance from the detention reservoir	detension_dis	City of Busan
	Length of sewer pipe	sewer_pipe	
	Drainage system	drainage_system	

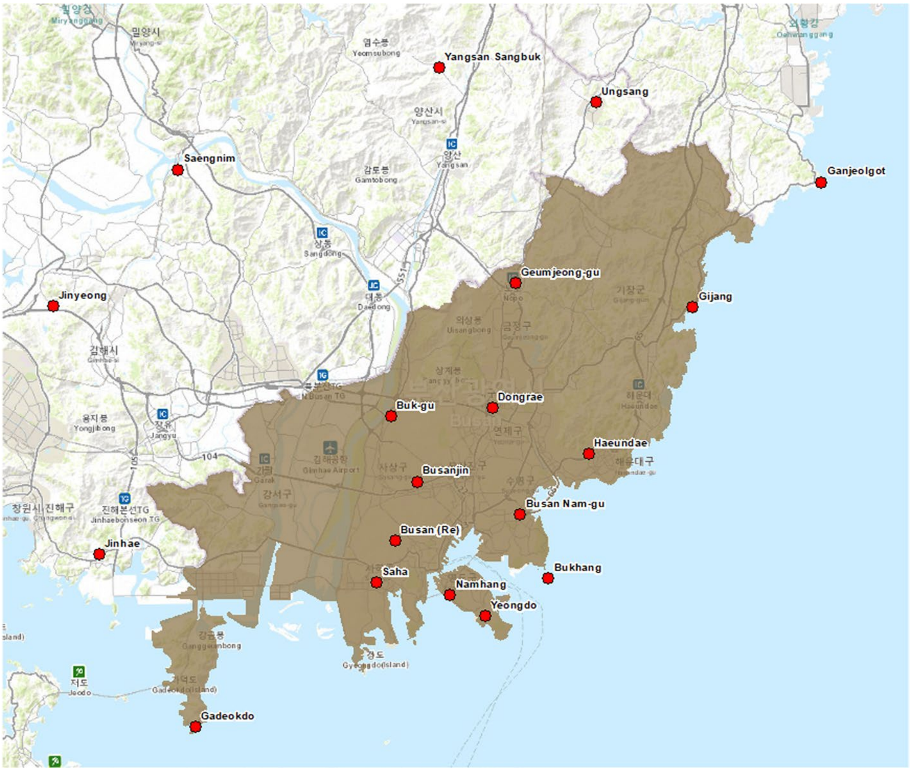


Fig. 1 Automatic weather station measurement points

For data of days having *more than 80 mm precipitation*, *maximum hourly precipitation*, and *maximum daily precipitation*, automatic weather station (AWS) data provided by the Korea Meteorological Administration were used. For the 13 measurement points in Busan Metropolitan City (i.e., Busan (Re), Yeongdo, Gadeokdo, Gijang, Haeundae, Busanjin, Geumjeong-gu, Dongrae, Buk-gu, Busan Nam-gu, Saha, Namhang, and Bukhang) and six measurement points in the Gyeongnam area adjacent to Busan (i.e., Ungsang, Jinyeong, Ganjeolgot, Yangsan Sangbuk, Jinhae, and Saengnim), daily precipitation and hourly precipitation for the years 2014, 2015, and 2016 were used (Fig. 1). The inverse-distance-weighted method was used among ArcGIS geostatistical spatial interpolation methods to calculate the weights of adjacent values according to distance, and the values for points of estimation were interpolated using the weighted average.

For altitude and slope data, a digital elevation model (DEM) was used. The DEM dataset is a numerical elevation model made by measuring the height of the topographical surface at regular intervals. In this study, data with a resolution of 90 m were used. Because the current law on national security restricts the public disclosure of precise 3D topography with a grid spacing of 90 m or more, the public data distribution was used. For the data on the distance from the river, data from national rivers, class 1 local rivers, class 2 local rivers, and small rivers were integrated for use in this study using the Near tool, which inputs the distance to the nearest point to the input feature class, the shortest distance from the river from the center of each grid cell was extracted. For data on the distances from the detention reservoir, the nearest distance from the detention reservoir was extracted the same way as above. Soil drainage data were classified into seven classes using a detailed soil map: extremely well-drained (EW), well-drained (W), moderately well-drained (MW), somewhat poorly drained (WP), poorly drained (P), very poorly drained (VP), and others (etc.). For impermeable data, we used the land cover map, the sum of the area for residential, industrial, commercial, and transportation areas corresponding to urban built-up areas in the major category. The data for the length of the sewer pipe and the number of drainage systems were provided by Busan Metropolitan City, and the sum of the length of the sewer pipe and the number of drainage systems included in each grid cell was calculated (Fig. 2).

2.3 Flood assessment model using machine learning methods

As shown in Fig. 3, the variables used for flood risk assessment were constructed using GIS data composed of spatial information and attribute data; data preprocessing was performed for the application of machine learning techniques. Here, data preprocessing refers to the process of replacing existing data with data suitable for machine learning algorithms and removing missing values and noise data.

In machine learning, the datasets must be divided for model training and validation; the training dataset included 617,122 regions and the test dataset included 264,482 regions, representing a ratio of 7:3 Furthermore, the data were randomly extracted, and it was ensured that the same region would not be duplicated during extraction. In this study, a flood assessment function was developed using the training datasets (70% of the total data), and the assessment power was evaluated by comparing the assessed value calculated by applying the developed function to the test dataset and the actual value of the measured data.

In this study, a flood assessment model was developed using DT, RF, NB, and SVM models.

In the DT, the decision-making rules were expressed in tree types for specific items, and based on the rules, data were classified into groups of similar data, and the classification continued until the final classification criteria were satisfied. This method has advantages in terms of objectivity and ease of interpretation compared with other machine learning methods (Breiman and Ihaka 1984). With DT, the space of each independent variable is repeatedly split to find the rule that best explains the dependent variable; therefore, it is possible to present the classification criteria of each impact variable used as an independent variable to assess the occurrence of floods. In this study, the ctree() function of the party package of R programming was used. The ctree() function uses the unbiased recursive partitioning based on the permutation test method, and because it determines the variables to be pruned based on significance determined by the p test, there is no risk of overfitting or bias, and no additional separate pruning is required. Additionally, by applying a procedure that considers multiple testing, the problem of repeated splitting of nodes in DT is reduced because the node splitting stops at an appropriate time.

An RF model generates multiple training data from one dataset and generates multiple DTs via multiple training, and the assessment performance is improved by aggregating the decisions of these trees. Hence, the model resolves the over-fitting problem, ordinarily a weakness of the DT model, by introducing maximum randomness. It also has high predictive power and strength of an ensemble model. In this study, a model was built using the randomforest() function of R programming, and the importance of each variable was derived using the importance() function, which was applied as a weight. Because RF uses

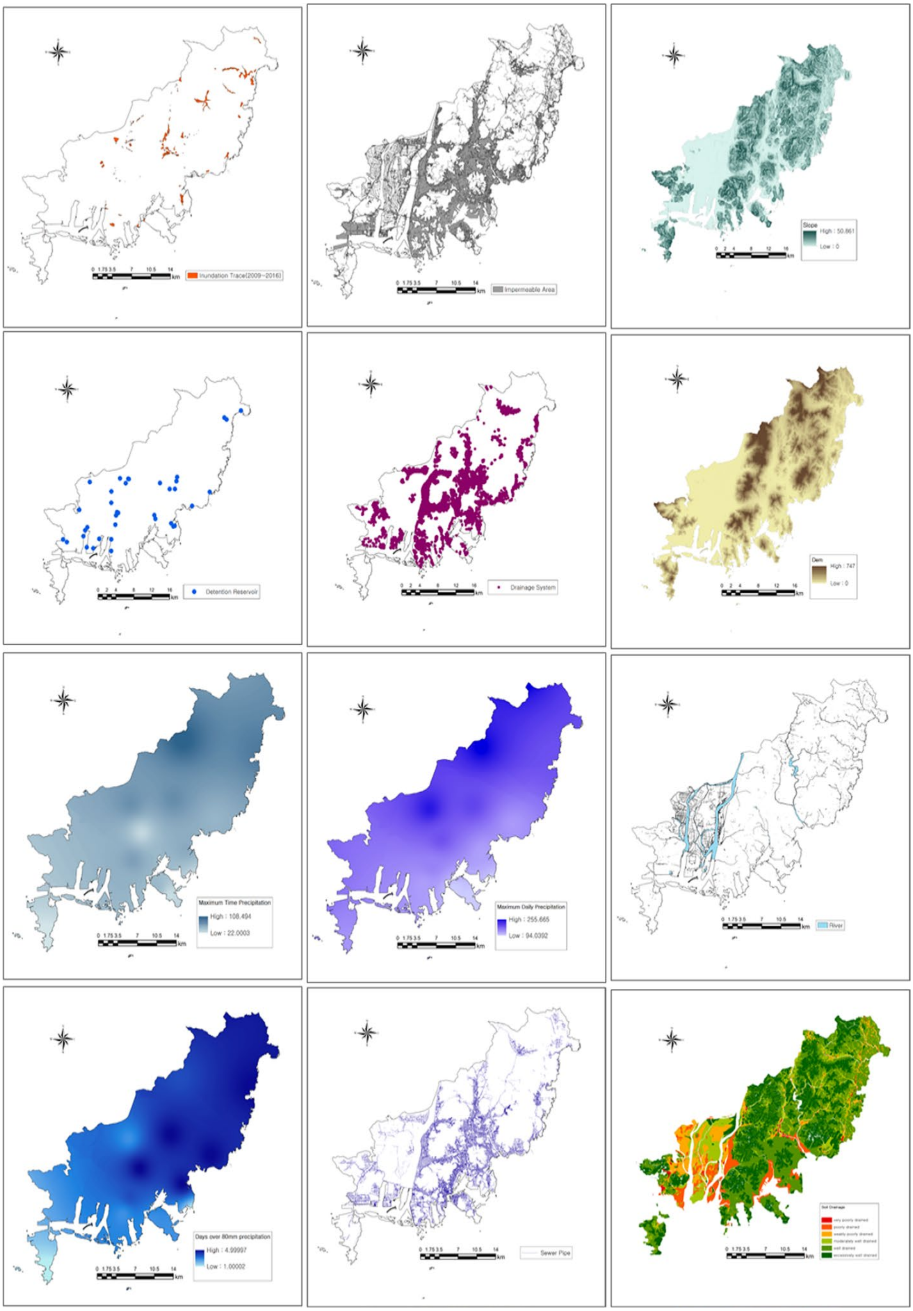


Fig. 2 Spatial visualization of analysis variables for assessment of flood-risk areas

bootstrap via random sampling, the value is not constant during repeated extractions. Thus, the seed value is set for extraction using the same value as the one used for extraction using the set.seed() function. Another point unlike the DT model is that the RF model does not

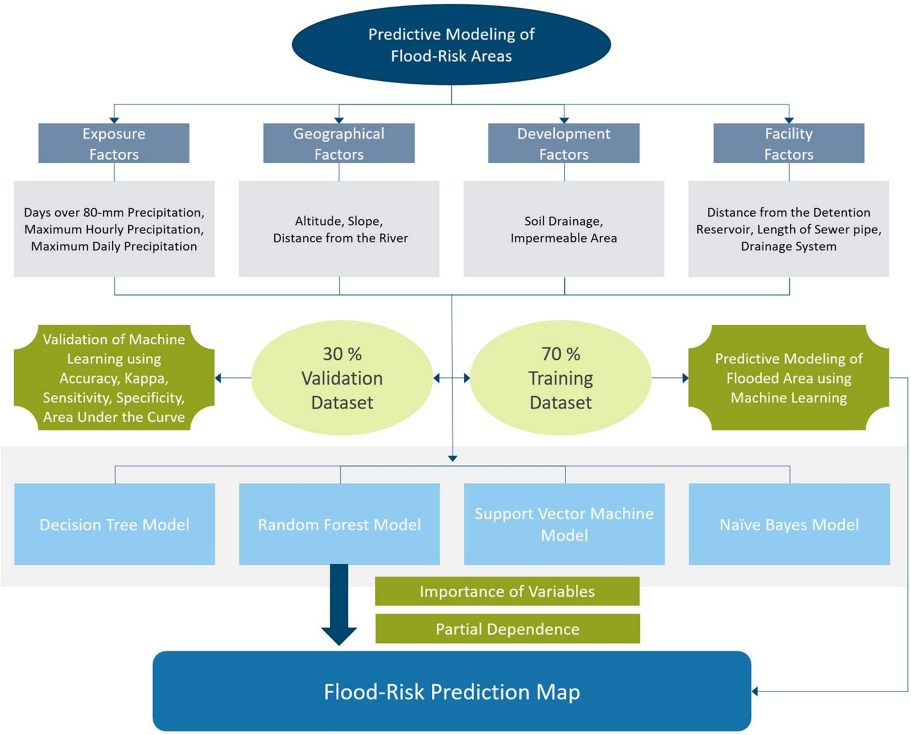

Predictive Modeling of Flood-Risk Areas
Exposure Factors
Geographical Factors
Development Factors
Facility Factors
Days over 80-mm Precipitation, Maximum Hourly Precipitation, Maximum Daily Precipitation
Altitude, Slope, Distance from the River
Soil Drainage, Impermeable Area
Distance from the Detention Reservoir, Length of Sewer pipe, Drainage System
Validation of Machine Learning using Accuracy, Kappa, Sensitivity, Specificity, Area Under the Curve
30 % Validation Dataset
70 % Training Dataset
Predictive Modeling of Flooded Area using Machine Learning
Decision Tree Model
Random Forest Model
Support Vector Machine Model
Naïve Bayes Model
Importance of Variables
Partial Dependence
Flood-Risk Prediction Map

Fig. 3 Methodological flowchart of the research process

need to divide the data into training and testing datasets for analysis. For a more accurate comparison with other models, the training and testing datasets are divided at a ratio of 7:3. Additionally, to select the optimal model, the performance between the RF models was comparatively evaluated using the out-of-bag (OOB) error rate. The OOB error rate is as accurate as using the same number of test datasets as the training datasets in the study of error measurement of bagged classifiers (Breiman 1996).

In the NB model, which is based on Bayes' theorem, the probability of each classification of the target for classification is measured, and the target is classified into a group having a larger value of calculated probability. In this method, the calculation is simplified by assuming conditional independence when calculating the posterior probability. The NB model is mainly suitable for problems requiring consideration of multiple attribute information to estimate the overall probability of a result. In this case, if all events are independent, it is impossible to assess one event by observing another. In this study, the naiveBayes() function of the e1071 package of R programming was used for analysis.

The SVM model is a non-probabilistic binary classification method that performs classification by minimizing the error for training data through support vectors. It finds a line (or a plane) that maximizes the distance between data belonging to a different class, and data are classified based on the line (plane) (Lee HH et al., 2016). SVM is widely considered the best method among classification methods and shows good performance in various data distributions because it has superior accuracy and a smaller probability of overfitting compared with other classification methods (Choi et al. 2013). When making assessments on new data, SVM measures the distance between the data and each support vector, the classification decision is based on the distance to the support vector, and the importance of the support vector is learned during training. In this study, analysis was performed using the ksvm() function of the kernlab package in R. With kernlab, a kernel-based machine learning algorithm is implemented in R, and the function has the advantage that users can easily extend functions without modifying the C++code.

2.4 Methods of assessment performance evaluation

There are several machine learning models; thus, the optimal model for flood assessment was selected by comparing different ones. Because the flood assessment model in this study is a classification model that assesses whether an area is flooded, to evaluate the assessment performance of the model, accuracy, kappa, sensitivity, specificity, and AUC were used. Accuracy is the ratio of correctly classified data among total data, indicating how accurately the model performs classification. It adjusts the accuracy by explaining the probability of making an accurate assessment of classification by chance. Kappa is particularly important because when there is a highly imbalanced dataset, high accuracy can be easily achieved by predicting only the most frequent values. Sensitivity and specificity can capture the trade-offs between the two. Sensitivity measures the percentage of true positives showing accurate classification, and specificity measures the percentage of true negatives showing accurate classification. In this study, sensitivity means the probability that an actual non-flooded area is assessed not to be flooded, and specificity means the probability that an actual flooded area is assessed to be flooded. Finally, AUC indicates the area under the ROC curve, and because the classification threshold does not change, AUC can evaluate the assessment quality of the model regardless of the selection of the classification threshold value. Therefore, a machine learning model should not be evaluated with only one of the above values; it is important to find an appropriate balance between them.

2.5 Partial dependence by variable

Even when a machine learning model with high assessment performance is derived, it is difficult to determine whether an independent variable has a positive (+) or negative (-) effect on the dependent variable using the model alone. Calculating the partial dependence for the independent variables addresses this problem, and the partial dependence represents the average marginal effect of an independent variable on the dependent variable (Liaw and Wiener 2002). Generally, a partial dependence plot is used for visualization, and it derives the functional relationship between a specific independent variable and the assessment of the model, showing how the assessment of the dependent variable is partially affected by the value of the independent variable of interest. The partial dependence for each variable is calculated as shown in Eq. 1:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_1\left(x, x_{ic}\right)}{p_0\left(x, x_{ic}\right)}\right). \tag{1}$$

A large partial dependence value indicates that when the values of the other independent variables, x_c, are constant, the probability of classification of the dependent variable in a specific value of the independent variable of interest, x, is relatively high.

3 Results

3.1 Prediction results of machine learning models

3.1.1 Prediction by DT

The results of evaluating the relationship between the flooded area in 2014 and the 11 influencing factors used with the DT model are presented in Fig. 4; the accuracy of the model was 98.64%. The root node located at the top of Fig. 4 is the maximum hourly precipitation (maxT_14), and the results show that the maximum hourly precipitation had the greatest effect on flood occurrence. Nodes were split based on a maximum hourly precipitation of 71 mm, which indicates that the flood occurrence was affected depending on whether the maximum hourly precipitation was more or less than 71 mm. For the distance from the river (river_dis), 1,323 m was the threshold value for classification, and when examining the area assessed for flood occurrence, the result was reclassified with 82 mm of maximum hourly precipitation as the threshold. When the maximum hourly precipitation was between 71 mm and ~ 82 mm, the impermeable area (impermeable_area) was analyzed as a factor having a large impact, followed by *slope* and *altitude* variables. When the maximum hourly precipitation was 82 mm or more, the maximum daily precipitation (maxT_14) had a large impact on flood occurrence, followed by the impermeable area and the distance from the detention reservoir.

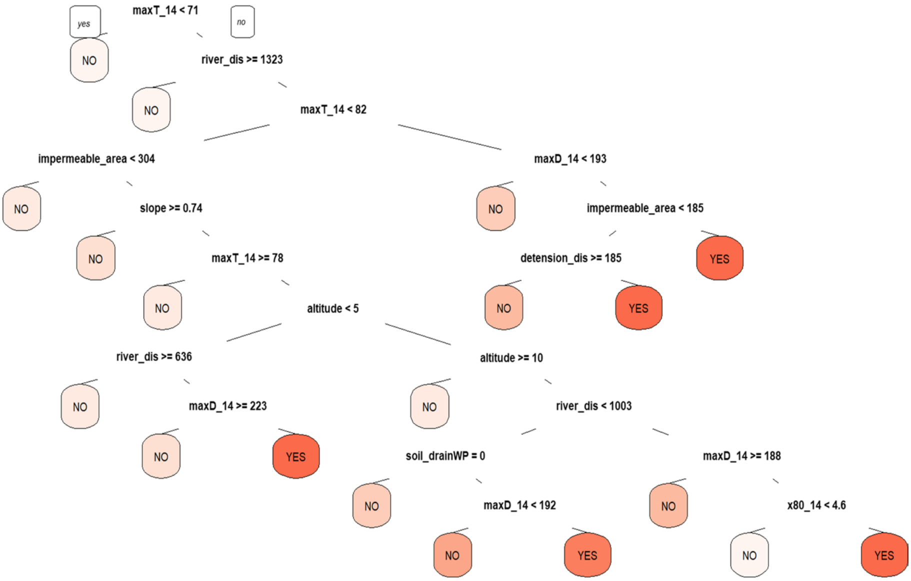


Fig. 4 2014 decision tree

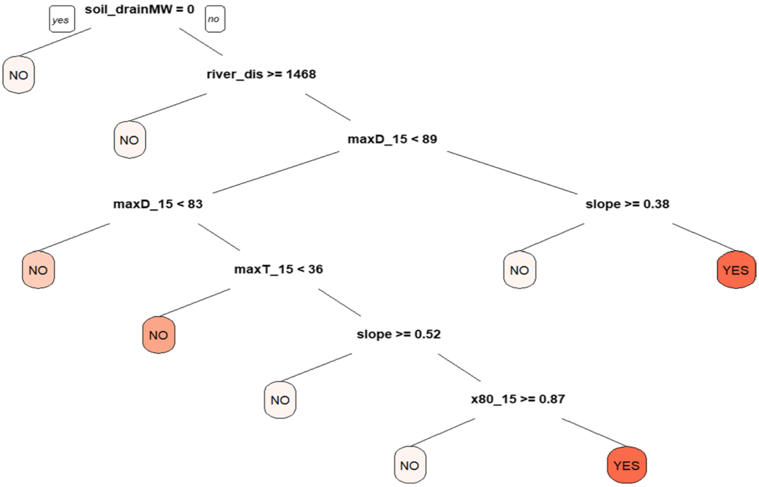


Fig. 5 2015 decision tree

The accuracy of the DT model for the 2015 data was 99.96%, and the model is illustrated in Fig. 5. Soil drainage (soil_drain) had the greatest impact on flood occurrence, and the nodes were split with reference to the soil drainage class of moderately well-drained (MW). This indicates that when the soil drainage class is MW, W, or EW, there is no occurrence of flood, and when the soil drainage class is WP, P, or VP, the flood occurrence probability is high. Then, the flood occurrence was classified with reference to the distance from the river at 1,468 m, and when examining the area assessed for flood occurrence, the result was reclassified with 89 mm maximum daily precipitation as the threshold. When the maximum daily precipitation was between 83 mm and ~ 89 mm, the influencing factors were in the order of *maximum hourly precipitation*, *slope*, and *days over 80 mm precipitation*, and when the maximum daily precipitation was 89 mm or greater, the slope was included as the influencing factor.

The results of the DT model with 2016 data are shown in Fig. 6, and the accuracy of the model was 99.92%. The length of the sewer pipe (sewer_pipe) had the greatest impact on flood occurrence, and the nodes were split based on the length at 0.21 m. When the length of the sewer pipe in the area (900 m^2) was less than 0.21 m, no occurrence of flood was assessed, and this is thought to be because in areas without human population such as mountains or rivers, there are no sewer pipes. Hence, flooding will not occur for sewer pipe reasons. Next, classification was made based on the distance from the river at 5191 m for both cases of a distance less than or greater than 5191 m; the distance from the detention reservoir was identified as an important influencing factor.

3.1.2 Prediction by RF model

In this study, to evaluate the factors affecting flood occurrence, the importance of 11 variables was evaluated using the RF method. When the RF model was implemented without parameter tuning, the number of trees to grow (ntree) was set to 500, and the number of variables randomly sampled as candidates at each split (mtry) was set to 2.

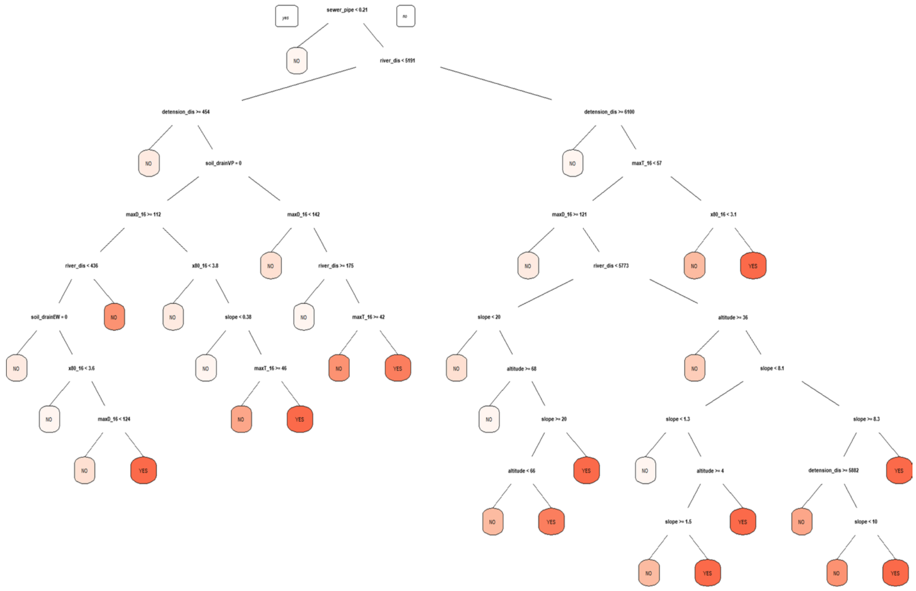


Fig. 6 2016 decision tree

Table 2 Optimal parameters for random forest model

(mtry, ntree)	2014 OOB	2015 OOB	2016 OOB
(2, 400)	0.4	0.02	0.08
(3, 400)	0.37	0.01	0.08
(4, 400)	0.36	0.01	0.07
(2, 500)	0.4	0.02	0.08
(3, 500)	0.37	0.01	0.08
(4, 500)	0.36	0.01	0.08
(2, 600)	0.4	0.02	0.08
(3, 600)	0.37	0.01	0.08
(4, 600)	**0.36**	**0.01**	**0.07**

Note: “OOB” = “out-of-box”

However, to improve model performance and increase accuracy, the values of these parameters must be tuned. Therefore, cross-validation was performed to determine appropriate parameter values. By varying the values of ntree to 400, 500, and 600 and mtry to 2, 3, and 4, a total of nine combinations were comparatively analyzed. As a result, as shown in Table 2, the combination showing the lowest OOB error rate for each year was (4, 600) with OOB error rates of 0.36, 0.01 in 2015 and 0.07 in 2016. Hence, the value of mtry was tuned to 4 and that of ntree was tuned to 600 for analysis.

In the RF model, explanatory power is the determination of the relative importance of explanatory variables affecting the target variable, which can be examined using two types of indicators. First, the mean decrease accuracy (MDA) determines the importance

Table 3 Performance comparison between models (second decimal place)

	2014 year				2015 year				2016 year			
	DT	RF	NB	SVM	DT	RF	NB	SVM	DT	RF	NB	SVM
Accuracy	0.99	1.00	0.94	0.99	1.00	1.00	0.96	1.00	1.00	1.00	0.97	1.00
Kappa	0.17	0.86	0.15	0.03	0.43	0.95	0.02	0.20	0.16	0.36	0.01	0.00
Sensitivity	1.00	1.00	0.95	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.97	1.00
Specificity	0.10	0.83	0.42	0.02	0.29	0.90	0.99	0.11	0.09	0.22	0.19	0.00
AUC	0.84	1.00	0.90	0.96	0.99	1.00	0.99	1.00	0.30	1.00	0.83	0.93

"DT" = "decision tree"; "RF" = "random forest"; "NB" = "Naïve Bayes"; "SVM" = "support vector machine"; and "AUC" = "area under the curve."

of variables using the concept of accuracy, and the difference in the reduced accuracy when the accuracy of the developed tree is reconstructed after removing a specific variable represented in terms of the average value for each variable. That is, the greater the variable having had a significant impact on increasing the classification accuracy, the greater the decrease in accuracy when the variable is removed. Using another method, when the RF model is constructed, the importance of a variable can also be judged by measuring the amount of impurity reduction in the selected variable whenever each tree branches out. This method is called "mean decrease Gini" (MDG). A high MDG value indicates that when classification is performed with the corresponding variable, and the impurity is reduced, which affects the grouping into the same categories. Therefore, the greater the values of both MDA and MDG, the higher is the importance of the variable.

A graph evaluating the variables of the model using the two indicators described above is presented in Fig. 7. By combining the two indicators, the results show that the variables of high importance in 2014 were *days over 80 mm precipitation*, *maximum hourly precipitation*, *maximum daily precipitation*, *slope*, and *distance from the detention reservoir*. For

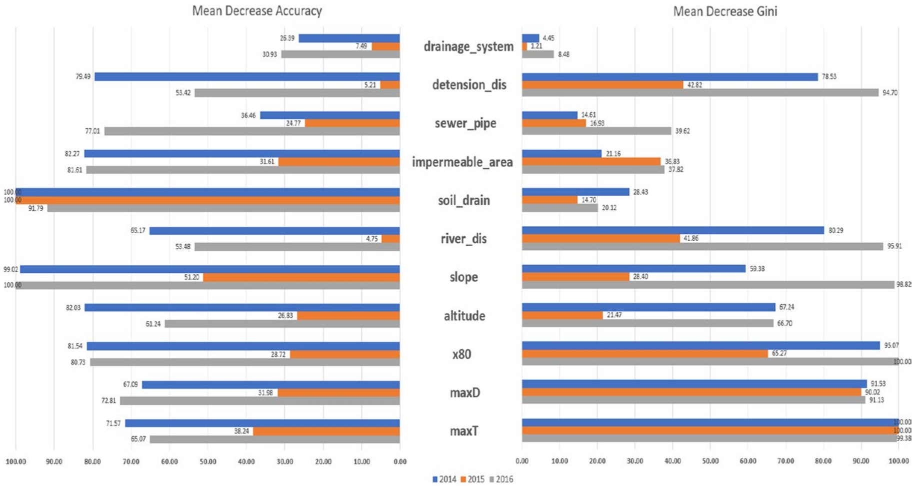


Fig. 7 Importance of variables for random forest model

2015, the important variables were *maximum hourly precipitation*, *maximum daily precipitation*, *soil drainage*, *days over 80 mm precipitation*, and *slope*. For 2016, *slope*, *days over 80 mm precipitation*, *maximum hourly precipitation*, *maximum daily precipitation*, and *distance from the detention reservoir* were identified as important.

3.1.3 Prediction by NB model

As a result of implementing the NB model for flood assessment, the accuracy was 93.87% in 2014, 95.7% in 2015, and 96.5% in 2016. As a result of comparing the assessed results of flood occurrence using the NB model with the test dataset, out of 264,482 total test datasets, flood occurrence in 1,615 areas was assessed using 3,819 areas with actual flood occurrence in 2015 data, flood occurrence in 147 out of 149 areas was assessed using 2015 data, and flood occurrence in 45 out of 235 areas was assessed using 2016 data. Although the overall accuracy was relatively high for the NB model, Kappa was 0.1463 in 2014, 0.0241 in 2015, and 0.0079 in 2016. Generally, if the kappa value was zero, the assessment was not in agreement with the actual measurements, and when $0 \leq \text{kappa} \leq 0.2$, it was analyzed to show slight agreement. That is, the model was considered unsuitable for flood occurrence assessment because the assessment result was at the level of agreement by chance.

3.1.4 Prediction by SVM model

The parameters of the SVM model include gamma and cost. Gamma is a parameter required for all kernels except for the linear one, and cost indicates the cost of violation of the margin. In this study, gamma was one/(data dimension), the value of 0.1 was used, and 10 was used as the cost parameter. As a result of implementing the SVM model for flood assessment, the accuracy was 98.57% for 2014 data, 99.95% for 2015 data, and 99.91% for 2016 data showed relatively high accuracy. However, as a result of examining the assessed data, it was found that in 2016, assessment was only possible when there was no flooding. Therefore, the SVM model was judged to be unsuitable for assessing and evaluating flood occurrences. Additionally, as a result of comparison with the test dataset, out of 264,482 total test datasets from the 3,819 areas with actual flood occurrence in 2014, 66 areas showed agreement with the assessment. For 2015 data, 17 out of 149 areas showed agreement, and for 2016 data, zero out of 235 areas showed agreement. The model was determined to be an extreme type with sensitivity in the range of 99.98–100% and a specificity of 0–11.41%, which shows that the model assessed most data as the areas of no flood occurrence, indicating the problem in the classification performance of the model.

3.2 Selection of optimum model through performance comparison

A confusion matrix is normally used to evaluate the performance of the classification model, the practical significance of the classification result of the model, and the precision and accuracy in the classification of the model. Among the indicators used in the confusion matrix, accuracy is the most representative, and it shows the accuracy of the model classification in terms of the percentage of correctly classified data out of the total data. A comparison of the average accuracy of the four models analyzed in this study is made (i.e., DT, RF, NB, and SVM), the accuracy of the DT model was 99.85%, the RF model was

99.51%, the NB model was 95.36%, and the SVM model was 99.48%; thus, the RF model was judged to be the optimal model for assessing flood occurrences (Table 3).

In addition to accuracy, the RF model showed the highest values of kappa, sensitivity, and specificity. In the case of kappa used as a measure of agreement in categorical data, the values were DT 0.25, RF 0.72, NB 0.06, and SVM 0.08, respectively, and the values of sensitivity representing the rate of true positives (i.e., the area without flood occurrence was correctly classified as the area without flood occurrence) were DT = 0.99, RF = 0.99, NB = 0.96, and SVM = 0.99. The values of specificity representing the rate of true negatives were DT = 0.16, RF = 0.65, NB = 0.53, and SVM = 0.04. In terms of AUC, the value of RF was the highest at 0.99, whereas for the other models, the values were DT at 0.71, NB at 0.91, and SVM at 0.96.

The values of the four indicators of kappa, sensitivity, specificity, and AUC ranged from zero to one, with one indicating that the model's assessed value and actual value were in perfect agreement. The closer the value is to one, the better the model performance. Categorical classification is subjective in terms of interpretation, but generally when the value is less than 0.2, it indicates almost no agreement. When the value is between 0.2 and 0.4, it indicates slight agreement, and when the value is between 0.4 and 0.6, it indicates moderate agreement. When the value is between 0.6 and 0.8, it indicates good agreement, and when the value is between 0.8 and 1.0, it indicates very good agreement. Therefore, the RF model can be interpreted as having more than good agreement, and it was judged to be suitable as a flood assessment model.

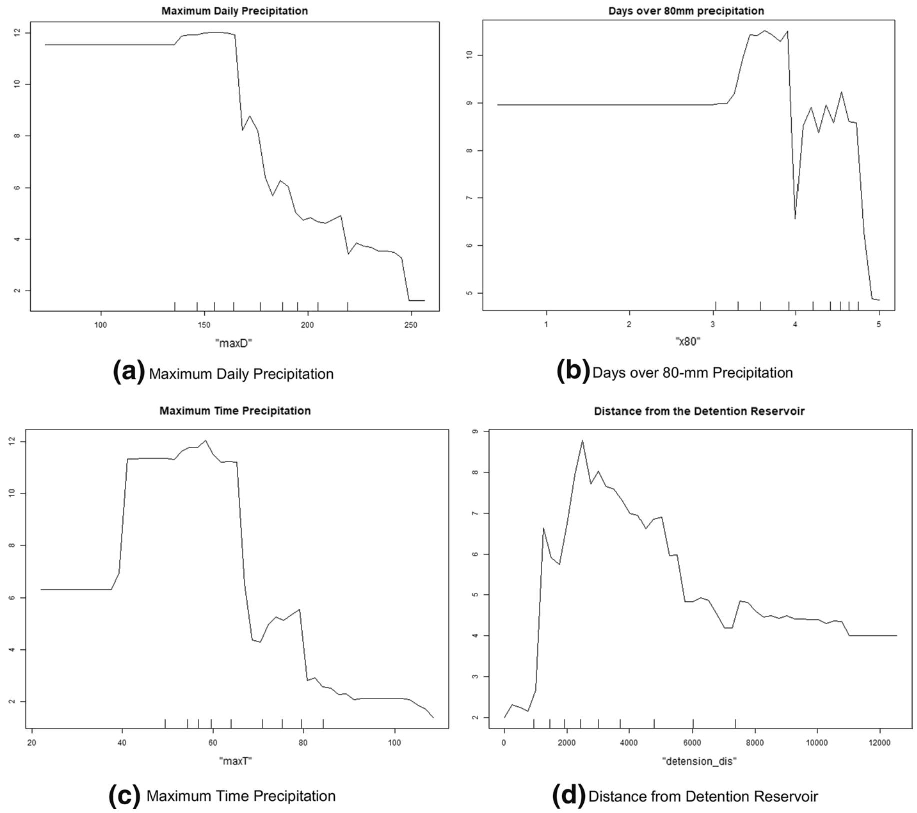
As a result of calculating the average importance of the variables for each year for the RF model selected as the optimal model for flood occurrence assessment in this study, the order of the importance of the variables was as follows: *maximum hourly precipitation*, *days over 80 mm precipitation*, *maximum daily precipitation*, *slope*, *soil drainage*, *distance from the detention reservoir*, *distance from the river*, *altitude*, *impermeable area*, *length of sewer pipe*, and *number of drainage systems*.

3.3 Partial dependence by independent variables in the random forest assessment model

In the analysis results of the RF model selected as the optimal model for flood assessment, the importance of independent variables for flood assessment was presented, but the method of the respective independent variables specifically impacting flood occurrence was not represented. In this regard, the final assessment model was constructed by synthesizing flood data from 2014, 2015, and 2016, and based on the model, detailed relationships between independent variables were examined using partial dependence.

In the analysis of the importance of variables, partial dependence was calculated according to changes in four variables: *maximum daily precipitation*, *maximum hourly precipitation*, *days over 80 mm precipitation*, and *distance from the detention reservoir*. The results are shown in Fig. 8. The values of the range for each variable showing the highest probability of flood occurrence are presented in Table 4.

For *maximum daily precipitation*, the precipitation range of 135–165 mm had the highest probability of flood occurrence, and for *maximum hourly precipitation*, the precipitation range of 42–65 mm showed the highest flood occurrence probability. The *days over 80 mm precipitation* was 3.4–3.8 days for the highest flood occurrence probability, and for the *distance from the detention reservoir*, the distance range of 2,200–3,000 m had the



(a) Maximum Daily Precipitation

(b) Days over 80-mm Precipitation

(c) Maximum Time Precipitation

(d) Distance from Detention Reservoir

Fig. 8 Partial dependence by independent variables

Table 4 Partial dependence result (section having the highest risk of flood occurrence)

	Highest range
Maximum daily precipitation (mm)	135–165
Maximum time precipitation (mm)	42–65
Days over 80 mm precipitation (days)	3.4–3.8
Distance from the detention reservoir (m)	2200–3000

highest probability of flooding. Because the value range with the highest probability of flood occurrence can be interpreted as the point where the flood occurs (i.e., the starting point), information on the value range of each variable having the highest probability of flood occurrence can be used as reference data for the establishment of a disaster prevention policy or response strategy.

3.4 Flood risk assessment map using RF model

The RF model was selected as the optimal model for flood occurrence assessment, and the flood risk level was calculated by applying the average importance of each variable as a

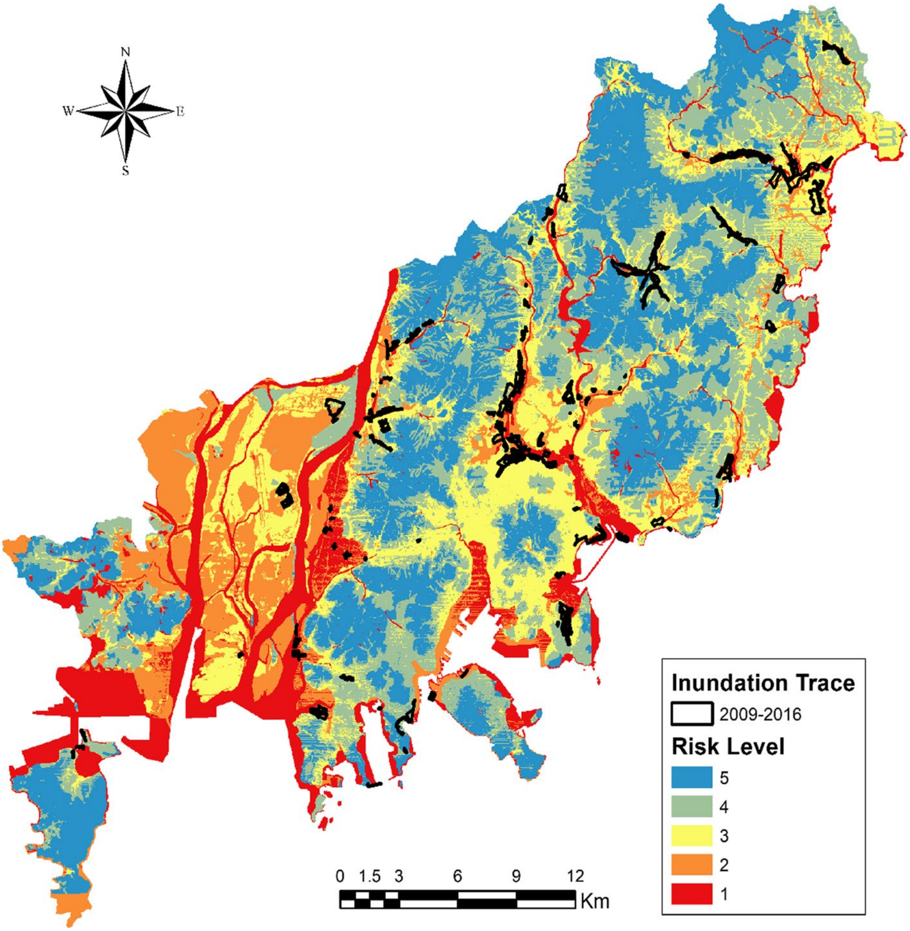


Fig. 9 Flood risk assessment map of Busan Metropolitan City

weight. When applying this weight, to find out whether the variable had a positive or negative (+, -) impact, the results of previous studies and those of the DT model were used as references. The derived flood risk levels are illustrated as a map of approximately 900,000 grid cells with dimension of 30×30 m for the area of Busan Metropolitan City (Fig. 9). Using the Jenks natural breaks classification (a classification technique of ArcMap), the flood risk level was classified into five classes for visualization. The Jenks natural breaks classification optimizes the order of data values into natural classes. Based on the average of all values within the same class, the average deviation was minimized, and the variance between each class was maximized. Thus, the method reduces variance within a class and maximizes the variance between classes, and it is mainly used when there is a considerable difference in data values, as in the case of the developed model (Jenks 1967).

In the flood risk assessment map of Busan Metropolitan City, the area with the flood risk class 1 level accounted for 122,580 grid cells representing an area of approximately 113; the area with the flood risk class 2 level accounted for 116,216 grid cells representing an area of approximately 105 km^2, the area with the flood risk class 3 level accounted for

177,970 grid cells representing an area of 160 km^2, the area with the flood risk class 4 level accounted for 229,487 grid cells representing an area of 207 km^2, and the area with the flood risk class 5 level accounted for 232,351 grid cells representing an area of 209 km^2. The class 1 and 2 areas with relatively high flood risk accounted for approximately 27.4% of the total area, and it can be seen that the distribution of risk areas was concentrated around the rivers.

By comparing the derived results with the actual areas of flood damage, it was confirmed that the actual flooded areas from 2009 to 2016 were distributed in the designated flood-risk areas. As can be seen from the comparison, because the analysis results of this study showed an overall similar pattern to the actual flood damage cases, it is considered that the developed flood risk assessment map has high applicability.

4 Conclusions and implications

The amount and scale of flood damage from urban floods have escalated, owing to the frequent occurrence of urban floods following extreme climate events caused by climate change in Korea. Hence, there is an urgent need for efficient flood risk maps that can be used to establish urban flood prevention measures. However, there are insufficient basic reference data available to assess areas subject to routine flooding in Korea. Japan and Europe have already mandated the creation of disaster maps.

Therefore, this study developed a machine learning flood assessment model that leverages DT, RF, NB, and SVM models for Busan Metropolitan City. Each model was applied to a test dataset, and their performances were evaluated based on accuracy, sensitivity, specificity, and AUC. The evaluation results showed that the RF model was optimal, reflecting the importance of influencing factors based on the number of days having greater than 80 mm precipitation, maximum hourly precipitation, maximum daily precipitation, ground slope, distance from detention reservoirs, soil drainage, distance from rivers, altitude, impermeable areas, length of sewer pipes, and number of drainage systems. According to the results of partial dependence analysis, maximum daily precipitation in the range of 135–165 mm was used to assess the highest probability of flood occurrence, and these values were 7.04–37.04 mm higher than the 127.96 mm average value (1995–2014) of maximum daily precipitation in the Korean Peninsula. 80 mm precipitation in the range of 3.4–3.8 days was used to assess the highest probability of flood occurrence, which is 0.5–0.8 days longer than the 2.9 average days over 80 mm precipitation in the last 10 years (2001–2010) in Busan. For maximum hourly precipitation, the highest probability of flood occurrence was assessed in the range of 42–65 mm. Considering that the 1-h design rainfall with 10-, 20-, and 30-year return periods in Busan City is 77.4, 89.4, and 96.5 mm, respectively, the results indicate that flooding may now occur with less rainfall than that of the 10-year return period.

The flood risk level was calculated using the RF model, and it resulted in a grid map of Busan Metropolitan City with approximately 900,000 grid cells at a scale of 30×30 m. The Jenks natural breaks classification technique of ArcMap was used, and the flood risk level was classified into five grades and visualized in the map. As a result of comparing areas having flood risk level classes of 1 and 2 (i.e., high flood risk), the actual flooded areas were correlated.

The proposed flood assessment model and risk levels are expected to be utilized as an authoritative data source for establishing disaster prevention measures, including determining priority areas and remedial actions to prevent flooding damage, which is expected to increase with climate change. It is also expected that the findings of this study can be used in the development of design standards for flood prevention facilities with comprehensive consideration of the size and importance of the facilities, their climate exposure, and the topography of the area.

This study is significant in that machine learning techniques employed in various other fields were adopted for flood assessment in Korea, and the superior performance of the RF model was confirmed through comparative analysis with the results of other models. However, because this study was localized to the region of the Busan Metropolitan City, different results may be derived according to different regional characteristics. Thus, these findings may not be globally generalizable. Additionally, in some areas, damage information was excluded because of the objection of residents. Nevertheless, this is one of the few studies both in Korea and in abroad that investigated the assessment of flood occurrence and estimation of flood-risk areas using machine learning methods. As a result, it is considered that further machine learning approaches will be usefully applied to this field. Therefore, we recommend that these techniques be applied to various other areas of disaster prevention in future studies.

Acknowledgements This study was partially supported by Korea Ministry of Land, Infrastructure and Transport (MOLIT), as Innovative Talent Education Program for Smart City.

References

Ahmad N, Hussain M, Riaz N, Subhani F, Haider S, Alamgir KS, Shinwari F (2013) Flood prediction and disaster risk analysis using GIS based wireless sensor networks, a review. J Basic Appl Sci Res 3(8):632–643

Alexander D (1993) Natural disasters. UCL Press, London

Birkmann J (2006) Measuring vulnerability to natural hazards: towards disaster resilient societies. United Nations University Press, Tokyo

Breiman L (1996) Out-of-bag estimation. Technical Report, Statistics Department, University of California Berkeley, Berkeley, California 94708, https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf.

Breiman L, Ihaka R (1984) Nonlinear discriminant analysis via scaling and ACE. Department of Statistics, University of California

Büchele B, Kreibich H, Kron A, Thieken A, Ihringer J, Oberle P, Nestmann F (2006) Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks. Nat Hazard 6(4):485–503

Bui DT, Pradhan B, Nampak H, Bui QT, Tran QA, Nguyen QP (2016) Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibility modeling in a high-frequency tropical cyclone area using GIS. J Hydrol 540:317–330

Cançado VL, Brasil LS, Guerra A, Nascimento NDO (2008) Análise de vulnerabilidade à inundação: estudo de caso da cidade de Manhuaçu, Minas Gerais. XVII Simpósio Brasileiro de Recursos Hídricos 3:1–16

Carlos E, Tucci, M (2007) Urban Flood Management. Hydraulic Research Institute, Federal University of Rio Grande, Rio Grande

Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, Khosravi K (2017) A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ Model Softw 95:229–245

Chen W, Li Y, Xue W et al (2020) Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. Sci Total Environ 701:134979

Chen YR, Yeh CH, Yu B (2011) Integrated application of the analytic hierarchy process and the geographic information system for flood risk assessment and flood plain management in Taiwan. Nat Hazards 59(3):1261–1276

Choi CH, Kim JS, Kim DH, Lee JH, Kim DK, Kim HS (2018) Development of heavy rain damage prediction functions in the seoul capital Area using machine learning techniques. J Korean Soc Hazard Mitig 18:435–447

Choi HS, Park HW, Park CY (2013) Support vector machines for big data analysis. J Korean Data Inf Sci Soc 24(5):989–998

Dutta D, Herath S (2001) GIS based flood loss estimation modeling in Japan. In: Proceedings of the US-Japan 1st workshop on comparative study on urban disaster management, pp 151–161

Falconer RH, Cobby D, Smyth P, Astle G, Dent J, Golding B (2009) Pluvial flooding: new approaches in flood warning, mapping and risk management. J Flood Risk Manag 2(3):198–208

Golding BW (2009) Long lead time flood warnings: reality or fantasy? Meteorol Appl 16:3–12

Han WS, Park TS (2014) Diagnosis of problems in the urban flood disaster prevention system and policy direction. KRIHS POLICY BRIEF 470:1–6

Herath S, Musiake K, Hironaka S (2003) A simulation study of infiltration facility impact on the water cycle of an urban catchment. Int Assoc Hydrol Sci 281:294–302

Hong H, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W (2018) Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Sci Total Environ 625:575–588

Jang DW, Kim BS, Yang DM, Kim BK, Seoh BH (2009) A development of GIS based excess flood protection system-using decision support methods. Korean Soc Civil Eng 2:630–634

Jenks GF (1967) The data model concept in statistical mapping. Int Yearbook Cartogr 7:186–190

Kang JE, Lee MJ (2012) Assessment of flood vulnerability to climate change using fuzzy model and GIS in Seoul. J Korean Assoc Geogr Inf Stud 15(3):119–136

Kang JE, Lee MJ (2015) Analysis of urban infrastructure risk areas to flooding using neural network in Seoul. J Korean Soc Civil Eng 35(4):997–1006

Kelman I, Spence R (2004) An overview of flood actions on buildings. Eng Geol 73(3–4):297–309

Khosravi K, Nohani E, Maroufinia E, Pourghasemi HR (2016) A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. Nat Hazards 83(2):947–987

Khosravi K, Shahabi H, Pham BT et al (2019) A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. J Hydrol 573:311–323

Kia MB, Pirasteh S, Pradhan B, Mahmud AR, Sulaiman WNA, Moradi A (2012) An artificial neural network model for flood simulation using GIS: Johor River Basin. Malay Environ Earth Sci 67(1):251–264

Kim DH, Kim JM, Yoon BC, Chang EM, Choi YS (2015) Development plan of grid system utilizing spatial information. J Korea Spat Inf Soc 23(6):43–55

Kim MJ, Kim GS (2018) Analysis of flood damage type by climate change using stepwise regression model. In: Proceedings of the korea water resources association conference. Korea Water Resources Association, pp 394

Korea Meteorological Administration (2017) Busan Metropolitan City Climate Change Assessment Report.

Lee HH, Chung SH, Choi EJ (2016a) A case study on machine learning applications and performance improvement in learning algorithm. J Digit Converg 14(2):245–258

Lee SE, Kim CH, Park TS, Kim ME, Kim SY, Lee TS, Kim JH (2016) Development of the Urban Flooding Risk Prevention System (I). Korea Research Institute for Human Settlements

Lee SE, Kim JW, Han WS et al. (2018) Development of the Urban Flooding Risk Prevention System (III). Korea Research Institute for Human Settlements

Lee SE, Kim SH (2015) New water resource technology paradigm in the era of water welfare, securing water, securing diversity in water use and reducing water disasters. Water Futur 48(1):68–75

Lee SH, Kang JE (2018) Urban flood vulnerability and risk assessments for applying to urban planning. J Korea Plan Assoc 53(5):185–206

Lee SH, Kang JE, Park CS (2016c) Urban flood risk assessment considering climate change using bayesian probability statistics and GIS: a case study from Seocho-Gu, Seoul. J Korean Assoc Geogr Inf Stud 19(4):36–51

Lee SM (2017) Spatial Analysis of Flood and Landslide Susceptibility in Seoul using Random Forest and Boosted Tree Models. Dissertation, University of Seoul

Lee SS, Park GW, Jang SM, An BJ, Ha SM (2019) Change from scenario-based to real-time prediction for proactive urban flood response. Mag Korean Soc Agric Eng 61(3):38–43

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22

Lim KS, Kang SW, Hwang MH, Choi SJ (2010) Development of evaluation system for optimal flood protection plan. In: Proceedings of the Korea water resources association conference, Korea Water Resources Association, pp 822–826

Liu R, Chen Y, Wu J et al (2016) Assessing spatial likelihood of flooding hazard using naïve Bayes and GIS: a case study in Bowen Basin, Australia. Stoch Env Res Risk Assess 30(6):1575–1590

Marconi M, Gatto B, Magni M, Marincioni F (2016) A rapid method for flood susceptibility mapping in two districts of Phatthalung Province (Thailand): present and projected conditions for 2050. Nat Hazards 81(1):329–346

Nicholas J, Holt GD, Proverbs DG (2001) Towards standardising the assessment of flood damaged properties in the UK. Struct Surv 19(4):163–172

Ouma YO, Tateishi R (2014) Urban flood vulnerability and risk mapping using integrated multi-parametric AHP and GIS: methodological overview and case study assessment. Water 6(6):1515–1545

Park MJ, Jun HD, Shin MC (2007) Estimation of sediments in urban watersheds and relation analysis between sediments and inundation risk using GIS. J Korean Soc Civil Eng B 27(3B):277–287

Penning-Rowsell E (1999) Flood hazard assessment, modelling and management: results from the EUROflood project. Environments 27(1):79

Rahmati O, Pourghasemi HR (2017) Identification of critical flood prone areas in data-scarce and ungauged regions: a comparison of three data mining models. Water Resour Manag 31(5):1473–1487

Seekao C, Pharino C (2016) Assessment of the flood vulnerability of shrimp farms using a multicriteria evaluation and GIS: a case study in the Bangpakong Sub-Basin. Thailand Environ Earth Sci 75(4):308

Seo MW, Lee JS, Choi Y (2016) Estimation of the natural damage disaster considering the spatial autocorrelation and urban characteristics. J Korean Soc Civil Eng 36(4):723–733

Shon TS, Kang DH, Jang JK, Shin HS (2010) A study of assessment for internal inundation vulnerability in urban area using SWMM. J Korean Soc Hazard Mitig 10(4):105–117

Sim UB (2008) Characteristics of urban flood damage in Korea and response tasks. Water Future 41(9):41–46

Sim UB (2011) New paradigm and urban policy for urban prevention of natural disasters preparing to climate change. Issue Paper 29:1–58

Smith DI (1994) Flood damage estimation-a review of urban stage-damage curves and loss functions. Water Sea 20(3):231–238

Son MS, Park JY, Kim HS (2013) Urban environmental risk -evaluating flooding risk indices of Seoul. Seoul Stud 14(4):127–140

Son MW, Sung JY, Chung ES, Jun KS (2011) Development of flood vulnerability index considering climate change. J Korea Water Resour Assoc 44(3):231–248

Song YS (2012) Quantitative relationship between flood vulnerability and urban inundation characteristics. Dissertation, University of Hanseo

Tang Z, Zhang H, Yi S et al (2018) Assessment of flood susceptible areas using spatially explicit, probabilistic multi-criteria decision analysis. J Hydrol 558:144–158

Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 504:69–79

Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J Hydrol 512:332–343

Tehrany MS, Pradhan B, Mansor S et al (2015) Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. CATENA 125:91–101

Tsakiris G (2014) Flood risk assessment: concepts, modelling, applications. Nat Hazard 14(5):1361–1369

US Army Corps of Engineers (USACE) (1996) Risk-based analysis for flood damage reduction studies. Manual, Eng, Washington, D.C.

Vojtek M, Vojteková J (2019) Flood susceptibility mapping on a national scale in Slovakia using the analytical hierarchy process. Water 11(2):364

Youssef AM, Pradhan B, Sefry SA (2016) Flash flood susceptibility assessment in Jeddah city (Kingdom of Saudi Arabia) using bivariate and multivariate statistical models. Environ Earth Sci 75(1):12

Zhao G, Pang B, Xu Z et al (2018) Mapping flood susceptibility in mountainous areas on a national scale in China. Sci Total Environ 615:1133–1142