

Novel Bayesian Additive Regression Tree Methodology for Flood Susceptibility Modeling

Janizadeh, Saeid; Vafakhah, Mehdi; Kapelan, Zoran; Dinan, Naghmeh Mobarghaee

DOI

[10.1007/s11269-021-02972-7](https://doi.org/10.1007/s11269-021-02972-7)

Publication date

2021

Document Version

Accepted author manuscript

Published in

Water Resources Management

Citation (APA)

Janizadeh, S., Vafakhah, M., Kapelan, Z., & Dinan, N. M. (2021). Novel Bayesian Additive Regression Tree Methodology for Flood Susceptibility Modeling. *Water Resources Management*, 35(13), 4621-4646. <https://doi.org/10.1007/s11269-021-02972-7>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

1 **Novel Bayesian Additive Regression Tree methodology for flood susceptibility**
2 **modeling**

3 Saeid Janizadeh¹, Mehdi Vafakhah^{1*}, Zoran Kapelan² and Naghmeh Mobarghaee Dinan³

4 ¹⁻ Department of Watershed Management Engineering and Sciences, Faculty in Natural Resources and Marine
5 Science, Tarbiat Modares University, Tehran, 14115-111 Iran, janizadehsaeid@modares.ac.ir

6 ¹⁻ Department of Watershed Management Engineering and Sciences, Faculty in Natural Resources and Marine
7 Science, Tarbiat Modares University, Tehran, 14115-111 Iran,* Correspond author: vafakha@modares.ac.ir

8 2-Department of Water Management, Delft University of Technology, Delft, The Netherlands, z.kapelan@tudelft.nl.

9 3- Department of Environmental Planning and Design, Environmental Sciences Research Institute, Shahid Beheshti
10 University, 1983969411 Tehran, Iran. n_mobarghaee@sbu.ac.ir

11 **Abstract**

12 Identifying areas prone to flooding is a key step in flood risk management. The purpose of this
13 study is to develop and present a novel flood susceptibility model based on Bayesian Additive
14 Regression Tree (BART) methodology. The predictive performance of the new model is assessed
15 via comparison with the Naïve Bayes (NB) and Random Forest (RF) based methods that were
16 previously published in the literature. All models were tested on a real case study based in the Kan
17 watershed in Iran. The following fifteen climatic and geo-environmental variables were used as
18 inputs into all flood susceptibility models: altitude, aspect, slope, plan curvature, profile curvature,
19 drainage density, distance from river distance from road, stream power index (SPI), topographic
20 wetness index (TPI), topographic position index (TPI), curve number (CN), land use, lithology
21 and rainfall. Based on the existing flood field survey and other information available for the
22 analyzed area, a total of 118 flood locations were identified as potentially prone to flooding. The

23 data available were divided into two groups with 70% used for training and 30% for validation of
24 all models. The receiver operating characteristic (ROC) curve parameters were used to evaluate
25 the predictive accuracy of the new and existing models. Based on the area under curve (AUC) the
26 new BART (86%) model outperformed the NB (80%) and RF (85%) models. Regarding the
27 importance of input variables, the results obtained showed that the location's altitude and distance
28 from the river are the most important variables for assessing flooding susceptibility.

29

30 **Keywords:** Flood susceptibility mapping; Bayesian; Regression Tree; Ensemble model; Bayesian
31 Additive Regression Tree (BART);

32 **1. Introduction**

33 Any unforeseen natural occurrence that weakens or destroys economic, social and physical
34 capacity, such as loss of life and finances, destruction of infrastructure, economic resources and
35 areas of employment is defined as a natural disaster. Examples include earthquakes, floods,
36 drought, seawater, volcanoes, landslides, hurricanes and natural pests (Vetrivel et al. 2018).
37 Flooding is one of the most dynamic and disruptive natural events that puts human life and property
38 and social and economic conditions at greater risk than any other natural disaster (Rahmati et al.
39 2016; Yariyan et al. 2020). This phenomenon causes damage to human achievements at all times
40 (Woodward et al. 2014; Darabi et al. 2019; Vafakhah et al. 2020). The highest risk of flooding and
41 corresponding damage is in the populated, i.e. urban areas. In recent years, the increase in urban
42 flood hazards, particularly along the river banks, has resulted in the risk of flooding for residents
43 and movable property (Choubin et al. 2019). Due to the varying climate, unpredictable
44 temperatures and rainfall in many of Iran's watersheds, several floods occur every year (Tehrany
45 et al. 2014). Limiting environmental resources, reducing and destroying them as a result of the

46 expansion of human activities, poses many challenges for today's society and the next generation.
47 The Kan watershed is affected by flooding events annually and this vulnerability has been
48 documented (Hooshyaripor et al. 2020). Seven important flood events were recorded in this
49 watershed since ..., causing damage to industrial, residential, agricultural land use, and fatalities,
50 according to the available information.

51 Reducing human casualties as well as damage to property and the environment is a key objective
52 shared by countries most often impacted by natural disasters. They are increasingly conducting
53 feasibility studies with economic analysis to mitigate the effects of these disasters (Molinos-
54 Senante et al. 2011). Although flooding cannot be prevented, the damage can be mitigated through
55 appropriate analysis and forecasting techniques (Heidari 2014). The first step is to identify flood-
56 prone areas (Janizadeh et al. 2019; Hosseini et al. 2020). One way to prevent and reduce flood
57 damage is to provide people with reliable information through flood hazard zoning maps (Cook
58 and Merwade 2009). The modelling of flood hazards, which may involve multi-temporal data sets,
59 is required. Recently, machine learning methods have been successfully applied to assess flood
60 risk with higher accuracy (Ngo et al. 2018; Talukdar et al. 2020). However, there is still no
61 agreement on which method or set of methods can provide the best predictions (Kalantar et al.
62 2021; Costache et al. 2021).

63 Rapid access to satellite imagery based on remote sensing data has increased the use of geographic
64 information systems in the preparation of flood susceptibility maps. A wide range of modelling
65 techniques has been proposed and used in natural disaster assessment including AI based
66 techniques (Sayers et al 2014). In recent years, Bayesian methods, partly because of their over-
67 resistance to the presence of small sample sizes and ability to deal with missing or incomplete data,
68 have been developed recently to model flood sensitivity. These include Naïve Bayes models (Liu

69 et al. 2016; Pham et al. 2020b; Tang et al. 2020) and regression tree models such as Random Forest
70 (RF) models (Arabameri et al. 2020; Chen et al. 2020; Vafakhah et al. 2020), Decision Tree models
71 (Khosravi et al. 2018; Costache 2019; Janizadeh et al. 2019; Pham et al. 2020a), Logistic
72 Regression models (Shafapour Tehrany et al. 2017; Al-Juaidi et al. 2018; Tehrany and Kumar
73 2018). These regression tree models have become popular in the research environment due to their
74 capability to model nonlinear phenomena such as floods.

75 Machine learning algorithms by default usually present point estimates only, and so decisions are
76 made ignoring the uncertainty surrounding these estimates. In recent years, the use of ensemble
77 models has attracted the attention of researchers in various fields as ensemble models benefit from
78 several individual models and therefore tend to have better performance than individual models
79 (Al-Abadi 2018; Tehrany et al. 2019a; Costache and Bui 2020; Shahabi et al. 2020). Bayesian
80 Additive Regression Tree is one of the new ensemble models that combines Bayesian and
81 Regression tree algorithms giving the access to the full posterior distribution of all unknown
82 parameters in the model. This can be useful to reduce the uncertainty.

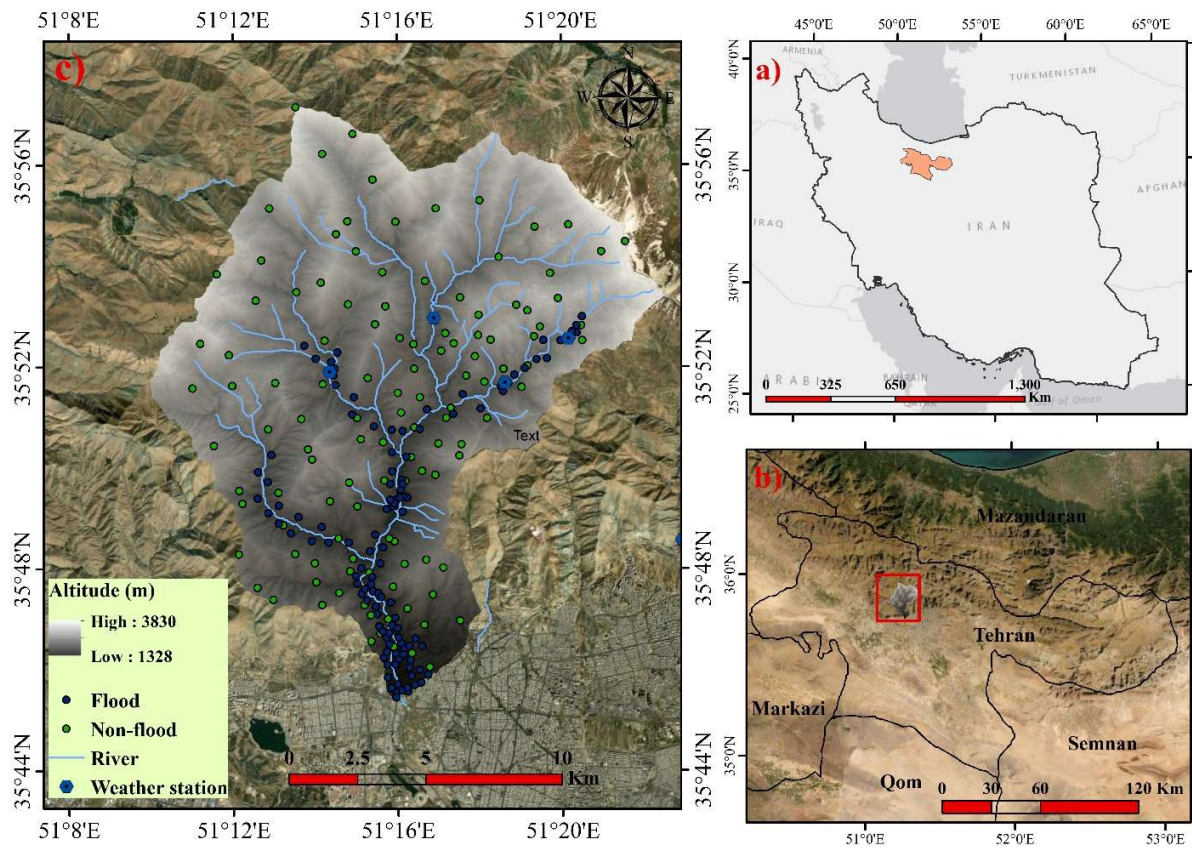
83 BART model has been used for modeling and predicting in different areas such as ecological
84 processes (Plant et al. 2021) and gully erosion (Chowdhuri et al. 2020). Due to the fact that the
85 flood is a non-linear phenomenon and has a lot of the uncertainty, use of appropriate models that
86 have the ability to predict this phenomenon and reduce uncertainty is essential in the management,
87 planning and prevention of flood risk. In the field of flood hazard modeling so far, very little
88 attention has been paid to the role of hybrid Bayesian and Decision Tree algorithms. Therefore,
89 the purpose of this study is to develop and present a new flood susceptibility model based on the
90 ensemble type Bayesian Additive Regression Tree (BART) method. The new method will be

91 compared with the Naïve Bayes (Bayesian type) and Random Forest (regression tree type) based
92 models to evaluate the predictive performance of the new method.

93

94 **1.2. Study area**

95 The Kan River watershed is 200 km² and is located northwest of Tehran, Iran. This watershed is
96 located between latitudes 51° 10' and 51° 23' east and 35° 46' and 35° 58' north (Fig. 1). The
97 average height of the watershed is 2428 meters, the average slope of the whole watershed is 43.4%
98 and the most important river in this mountainous region is the Kan river. The study area is located
99 in the southern margin of the central Alborz region in terms of geological status and has a
100 mountainous climate with the average annual rainfall of 414.13 mm. The average annual discharge
101 of the Kan River is 2.2 m³/s and its annual water flow is about 70 million m³/year. Seven important
102 flood events have been reported in the Kan watershed since ..., which have caused damage to
103 commercial and residential facilities, agricultural land and even caused casualties in the region
104 (Delkash et al. 2014).



105

106

Fig. 1. Location of case study a) country of Iran b) Tehran Province and c) Kan watershed

107

2. Material and methods

108

2.1. Flood Inventory Data Preparation

109

In order to prepare a flood susceptibility map it is necessary to analyze the historical floods. The

110

Kan watershed has been severely affected by dangerous floods in recent decades, causing

111

extensive damage and casualties. According to historical floods recorded by the Regional Water

112

Company of Tehran Providence (1954/8/27, 1955/6/9, 1978/3/7, 1981/7/25, 1986/2/2, 1995/4/23,

113

1996/4/3), field visits and interviews with locals on 2019/10/5 to 2019/10/9 and the identification

114

of flood-affected areas by GPS equipment (Fig. 2), 118 flooding locations are identified in the

115

area. In addition to this, further 118 non-flood points were randomly placed in the inter-fluvial

116 area, or within very steep altitude where the flood phenomenon is almost impossible in the case
117 study area. The position of all 236 locations are presented in Fig. 1. The data were divided into
118 two categories of training and validation for modeling, so that 70% of the data were used for
119 training and 30% for validation (Ahmadlou et al. 2019; Choubin et al. 2019). The flowchart of
120 research methodology is given in Fig. 3.

121

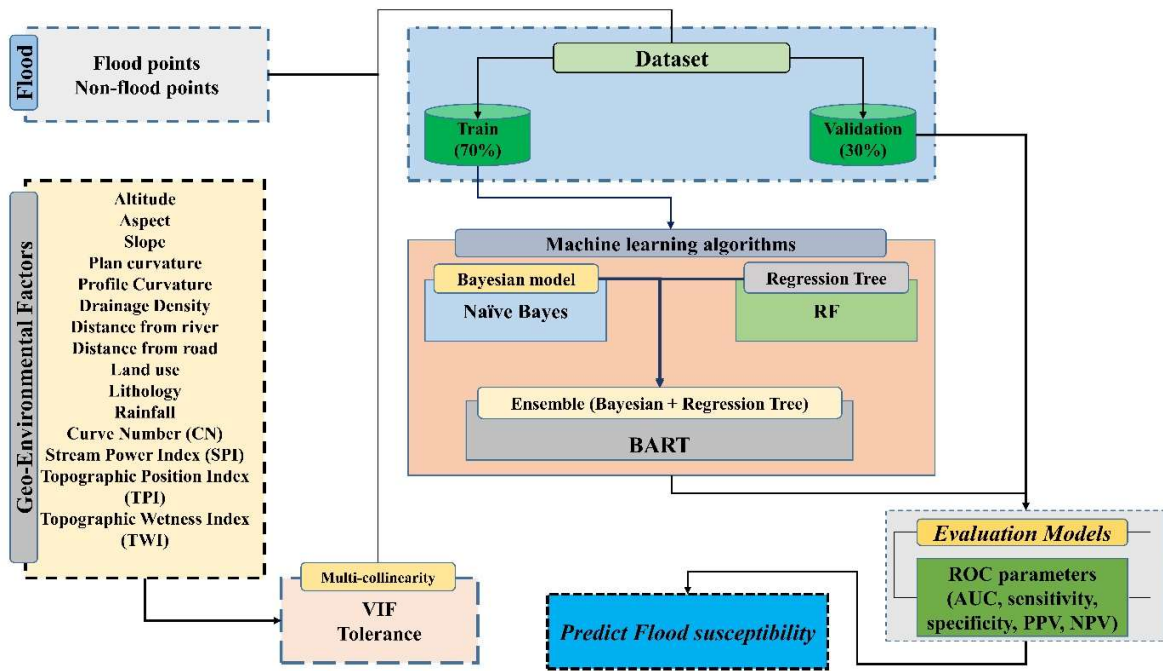


122

123

Fig. 2. Example of a flood location in the Kan watershed

124



125

126

Fig. 3. Research Methodology

127

128 2.2. Spatial Data Preparation

129 Floods are one of the natural phenomena and are affected by various climatic and geo-
 130 environmental factors. In this study, the following 15 climatic and geo-environmental variables
 131 are used as potential explanatory factors for flood susceptibility at a given location: altitude, aspect,
 132 slope, plan curvature, profile curvature, drainage density, distance form river distance from road,
 133 stream power index (SPI), topographic wetness index (TWI), topographic position index (TPI),
 134 curve number (CN), land use, lithology and annual rainfall (Ngo et al. 2018; El-Magd et al. 2021).

135 The above 15 factors (i.e. potential flood susceptibility model independent variables) were
 136 confirmed as significant by using the multi-collinearity analysis. The multi-collinearity analysis
 137 evaluates the intensity of multiple correlations between considered variables by calculating the

138 variance inflation factors (VIFs). The higher the value of the VIF the more likely it is that that
139 variable does not play a significant role in flood susceptibility prediction (Miles 2014). In this
140 study, the threshold of 5 was used for VIF to identify significant independent variables (Tehrany
141 et al. 2019a; Hosseini et al. 2020). VIFs were estimated using the USDMM package in R software.
142 The analysis has shown that all fifteen variables shown here have VIF values below the above
143 threshold (see section 4.1) hence they have all been used a potential explanatory factors for
144 predicting the flooding susceptibility.

145 The values of above 15 variables were prepared based on previous studies (see Fig 4, 5 and 6). For
146 this purpose, the digital elevation model (DEM) of the study area with resolution of 12.5×12.5 m
147 was developed with elevation data obtained using the type L-band Synthetic Aperture Radar
148 (PALSAR) (<https://vertex.daac.asf.alaska.edu/#>). The aspect map was prepared based on DEM at
149 nine class in the ArcGIS 10.5 software (Choubin et al. 2019; Janizadeh et al. 2019). The ground
150 slope is one of the important factors in the occurrence of floods in watersheds (Tehrany et al. 2015;
151 Chapi et al. 2017). The slope map was prepared based on the DEM in ArcGIS 10.5 software
152 (Khosravi et al. 2018).

153 The plan and profile curvature are the spatial parameters used in the preparation of flood maps of
154 watersheds. These variables were prepared in ArcGIS 10.5 software using a DEM (Rahmati et al.
155 2016; Hong et al. 2018). Drainage density of the study area in ArcGIS 10.5 environment was based
156 on line density extension (Mahmoud and Gan 2018; Zhao et al. 2019). Distance from rivers is one
157 of the most important factors affecting flooding of lands along the rivers (Tehrany et al. 2014;
158 Khosravi et al. 2016, 2018). This map was prepared using the Euclidean order in ArcGIS 10.5
159 software (Khosravi et al. 2018). Distance from the road is also a factor affecting flooding. This

160 variable was prepared using the 1:50,000 road map of Tehran province, the ArcGIS10.5 software
161 and the Euclidean extension, to determine distance from the road (Shafapour Tehrany et al. 2017).

162 The stream power index (SPI) is one of the important parameters for flooding in watersheds and
163 the following relationship is defined here (Tehrany et al. 2014; Shafizadeh-Moghadam et al. 2018):

$$164 \quad SPI = Catchment Area * \tan(slope) \quad (1)$$

165 System for Automated Geoscientific Analyses Geographic Information System (SAGA GIS 2.6)
166 software was used to prepare this variable (Tehrany et al. 2014).

167 Topographic position index (TPI) indicates the topographic status of the area, with positive values
168 indicating high altitudes and negative values indicating low altitudes such as valleys (Papaioannou
169 et al. 2015). Due to the role of topographic shape in the formation of floods, this index is considered
170 as one of factors affecting floods and this variable was prepared using the SAGAGIS 2.6 software.
171 TWI measures the effect of local topography on runoff production and shows the long-term
172 moisture content of a landscape (Hong et al. 2018; Khosravi et al. 2019), hence this indicator is
173 one of the influential variables in flood risk assessment in watersheds. This variable was obtained
174 based on the following (Khosravi et al. 2019) in SAGAGIS 2.6 software:

$$175 \quad TWI = \ln(Catchment Area / \tan(slope)) \quad (2)$$

176 Lithology is one of the important factors in watershed flooding due to its direct effect on the level
177 of permeability and surface runoff (Rahmati et al. 2016). The geological map of the Kan watershed
178 was prepared based on the 1:100,000 geological map of the Iranian National Cartographic Center
179 (NCC) and then turned into a raster layer with a resolution of 12.5 m. The lithology map of the
180 study area was divided into seven different classes. The soil type map was also prepared using the
181 data from the Administration of Natural Resources of Tehran Province and the vector file of this

182 map was created with a raster format with pixel size of 12.5 meters using the ArcGIS 10.5 software
183 (Tehrany et al. 2014).

184 Land use is the result of the interrelationships of socio-cultural parameters and the potential of the
185 land (Rahmati et al. 2016; Bui et al. 2018). Changes in land use and land cover can have significant
186 impact on flooding in watersheds (Khosravi et al. 2018). This map was prepared using images of
187 Landsat 8 satellite imagery OLI sensors in 2019 and using the maximum likelihood algorithm and
188 supervised classification in the ENVI 5.1 software and divided into four classes: orchard,
189 rangeland, residential and rocky lands.

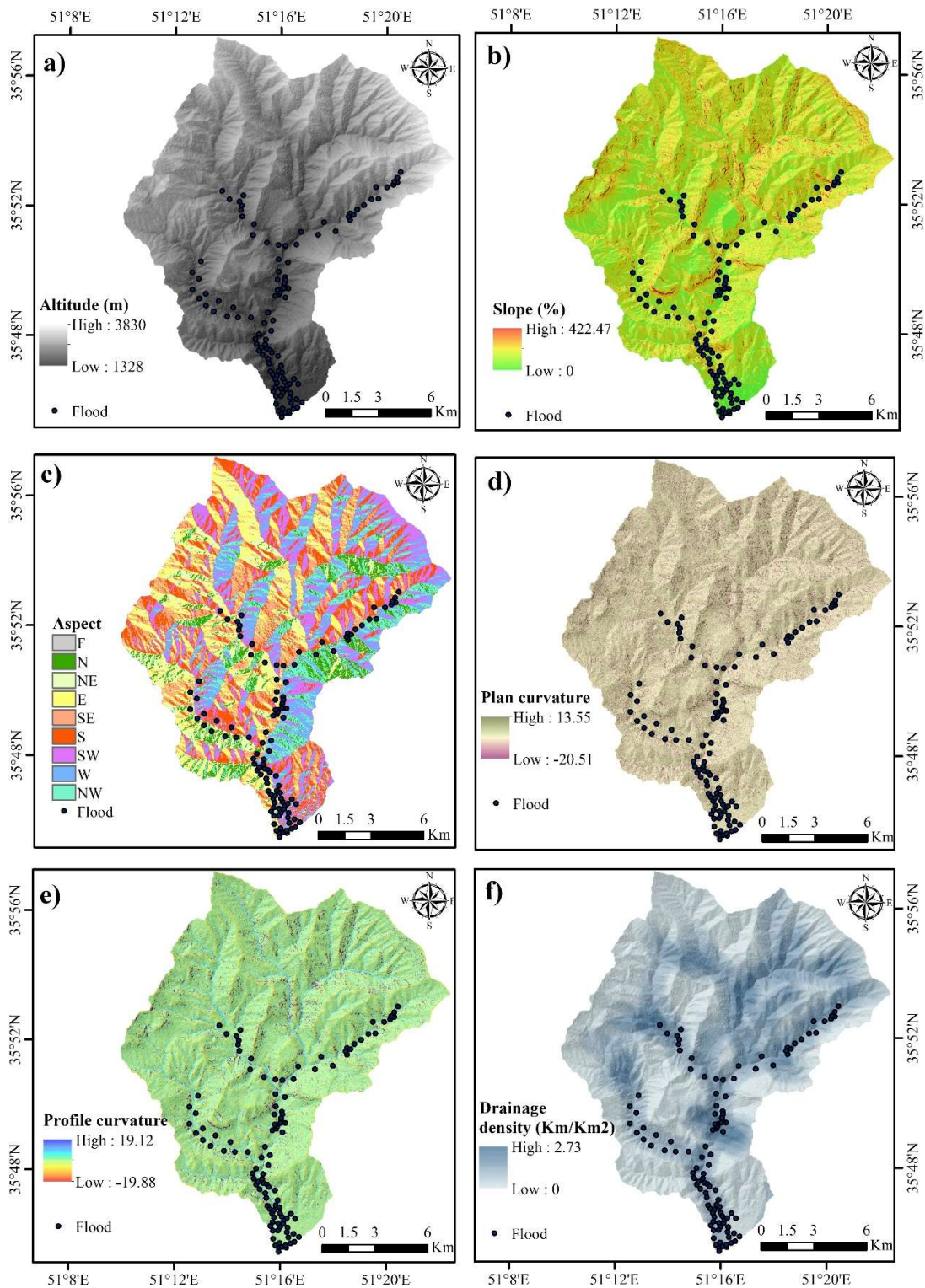
190 In order to prepare the annual rainfall map, the rainfall data of 7 gauge stations (inside and outside
191 the watershed) were used in the period 1994-2019. After carefully examining the various
192 interpolation methods in the ArcGIS 10.5 software, the distribution of annual rainfall in Kan
193 watershed was prepared based on the ordinary Kriging method.

194 One of the most important factors in the occurrence of floods is soil condition and different land
195 uses, which directly affects the amount of water infiltration into the land. In other words, the curve
196 number (CN) at the level of each area indicates the hydrological behavior of that area and its
197 discharge regime during rainfall. In order to determine the CN map the land use map and the
198 hydrological soil groups map were combined in the ArcGIS software environment. Then, based
199 on the tables related to the CN for different land uses of watersheds and according to hydrological
200 soil groups map, the value of CN was determined in the case of previous average humidity
201 (Mahmoud and Gan 2018; Tang et al. 2018).

202 The data summary information of all independent variables is shown in Table 1.

203 Table 1. Information of independent variables

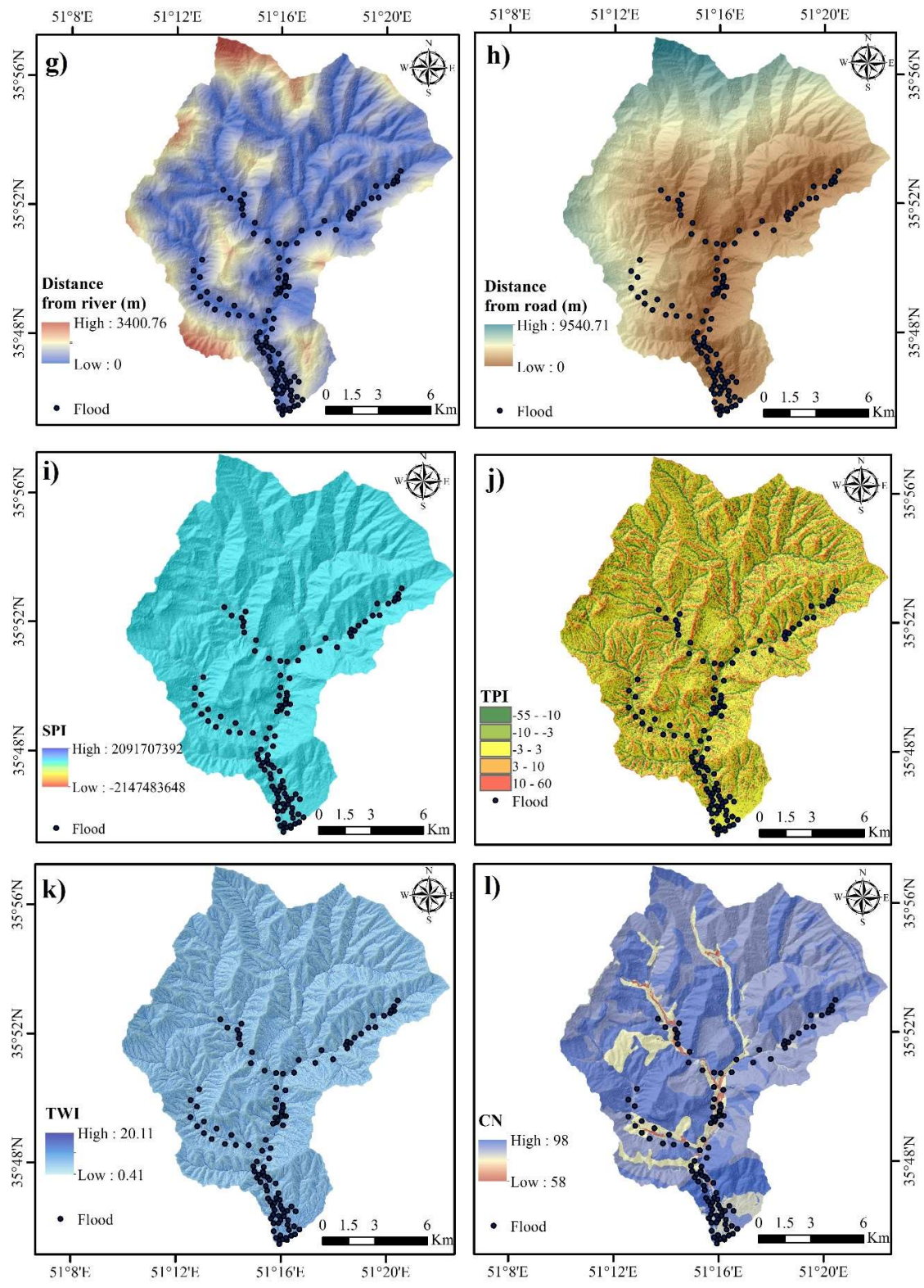
Variables	Data Type	Data Source	Data resolution
Elevation	Raster Grid	ALOS PALSAR DEM, (Alaska Satellite Facility)	12.5 m* 12.5 m resolution
Aspect			
Slope			
Plan Curvature			
Profile Curvature			
Drainage Density			
SPI			
TWI			
TPI			
Distance from River	Line and polygon coverage	Administration of Natural Resources, Department Tehran Province.	1:50000
Distance from Road	Line and polygon coverage		1:50000
LULC	Spatial/Raster grid	Landsat 8 OLI (USGS)	30 m spatial resolution
Lithology	Line, point and polygon coverage	Geological Map by country's mapping organization (Iran)	1: 100000
Rainfall	Station specific information	25 Years information of rain gage stations	Interpolation with same spatial resolution with other parameters
CN	Raster Grid	LULC and hydrological soil groups map	



205

206
207

Fig 4. Flood conditioning factors: a) altitude, b) aspect, c) slope, d) plan curvature, e) profile curvature, f) drainage density

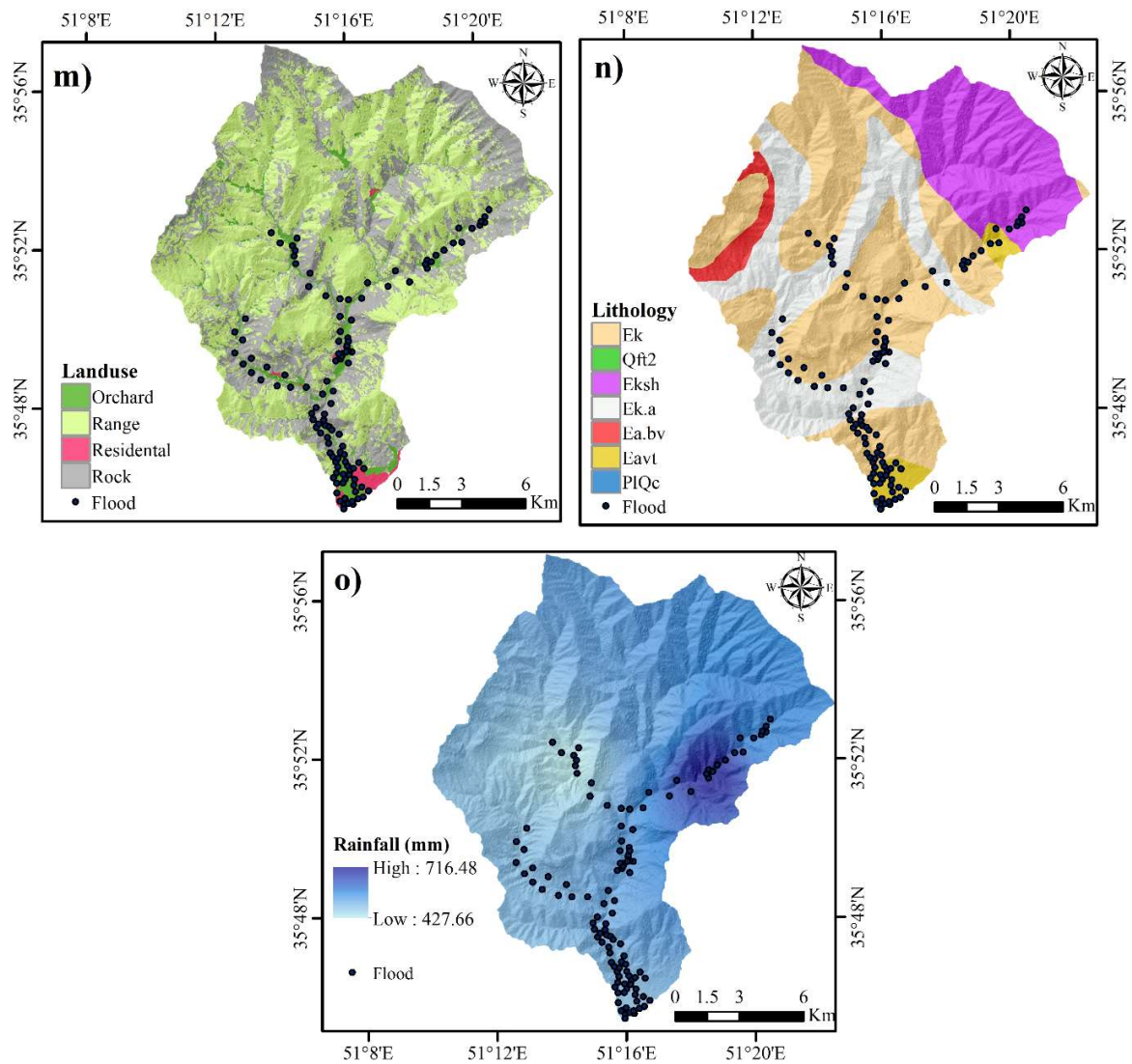


208

209 **Fig 5. Flood conditioning factors: g) distance from river, h) distance from road, i) SPI, j) TWI, k) TPI, l) CN**

210

211



212

213

Fig. 6. Flood conditioning factors: m) land use, n) lithology and o) annual rainfall

214 **2.3. Flood susceptibility models**

215 This section describes three different models for predicting flood susceptibility: BART, NB and
216 RF. All models are based on different machine learning methods that predict the flood
217 susceptibility defined as the probability of flood occurrence at a given location of the analyzed

218 watershed. All three models have the same set of input variables, the fifteen explanatory /
219 independent variables described in section 2.3. These model inputs were determined in all cases
220 using correlation and multi-collinearity analysis (see next section). Finally, all models are trained
221 and tested using the data described in the next section.

222

223 **2.3.1. Naïve Bayes Model**

224 The Bayesian method is a way of classifying phenomenon based on the probability of that
225 phenomenon occurring or not occurring. Based on the inherent characteristics of probability
226 (especially probability division), Naive Bayes method offers good results after receiving the initial
227 practice (Rish and others 2001). Learning method in the simplest way, the base is the type of
228 learning with the supervisor. Bayes suggests a way to calculate the posterior probability, $P(c | x)$,
229 from $P(c)$, $P(x)$ and $P(x | c)$. The Naive Bayes classifier assumes that the effect of the predictor
230 cost (x) on a given category (c) of the different predictor values is neutral. This assumption is
231 known as conditional independence:

$$232 \quad P(c|x) = \frac{P(c|x)*P(c)}{P(x)} \quad (3)$$

$$233 \quad P(c|X) = P(x_1|c) * P(x_2|c) * ... * P(x_n|c) \quad (4)$$

234 where $P(c|x)$ is posterior probability of target, $P(c)$ is prior probability of class and $P(x)$ is the
235 prior probability of predictor (Zhang 2004). The e1071 package in R software was used for Naïve
236 Bayes modeling.

237 **2.3.2. Random Forest Model**

238 Random Forest (RF) method is a relatively complex method in which several decision trees are
239 trained in order to increase the predictive accuracy of the model. The result is a prediction of a
240 group of decision trees. In the random forest learning method, each decision tree is taught using a
241 random sample selected from the training data set. The total selection of predictive variables used
242 to divide nodes is also random. In the random forest method, the two properties *mtry* and *ntree* are
243 determined for the number of auxiliary variables used in each subset and the number of trees used
244 in the forest, respectively. One of the advantages of a random forest is that it can be used for both
245 classification and regression type models. Random forest has parameters similar to the decision
246 tree or "Bagging Classifier". Random forest adds randomness to the model as trees grow. Instead
247 of searching for the most important features when dividing a "node", this algorithm looks for the
248 best features among a random set of features. This leads to more variety and ultimately a better
249 model. Therefore, in a random forest, only one subset of features is considered by the algorithm to
250 divide a node. By adding a random threshold for each attribute, instead of searching for the best
251 possible threshold, trees can be made even more random (Liaw et al. 2002). The randomForest
252 package in R software was use for the RF modeling here.

253 **2.3.3. Bayesian Additive Regression Tree (BART) Model**

254 BART is a Bayesian approach to non-parametric output estimation using regression trees. The
255 regression trees are relying on the return of the binary division of the predictive space into a set of
256 superconductors to approximate certain unknown functions. The predictive space has dimensions
257 corresponding to the number of variables. Tree-based regression models are capable of generating
258 plenty of interaction and nonlinearity (Hill et al. 2020). Models consisting of a number of
259 regression trees are more capable of capturing interaction and nonlinearity than single trees, as are
260 additives in f.

261 BART can be considered a general collection of trees with a new estimation method based on a
262 complete Bayesian probability model. The BART model can be expressed as follows:

$$263 \quad P(Y = 1|X) = \varphi(\tau_1^N(X) + \tau_2^N(X) + \dots + \tau_n^N(X)) \quad (5)$$

264 where φ denotes the cumulative density attribute of the prevalent regular distribution. In this
265 formulation, the sum-of-trees model serves as an estimate of the conditional probit at x which can
266 be besides issues modified into a conditional threat estimate of $Y = 1$ (Kapelner and Bleich 2013).
267 The `bartMachine` package in R software was use for BART modeling.

268 **2.3.4. Model Validation and Performance Assessment**

269 The ROC curve characterizes the relative performance of each model. The ROC curve is a graph
270 in which the true positive (or specificity value) is shown in the vertical axis whilst the false positive
271 (or sensitivity) is shown on the vertical axis (Frattoni et al. 2010). For the sensitivity or a proportion
272 of occurrence pixels that have been correctly predicted, the larger this value the more accurate the
273 model is in determining the occurrence points. Also, the feature means a ratio of non-occurring
274 pixels that the model correctly predicted. The area under the curve (AUC) measures one aspect of
275 performance. The value of AUC varies from 0 to 1, where the value of 0.5 denotes the random
276 prediction and 1 denotes the perfect prediction (Yesilnacar and Topal 2005). In this study, the
277 following equations have been used to calculate true positive rate (TPR), true negative rate (TNR),
278 specificity, sensitivity and AUC:

$$279 \quad TPR = \frac{TP}{(TP+FN)} \quad (6)$$

$$280 \quad TNR = \frac{TN}{(TN+FP)} \quad (7)$$

281
$$Sensitivity = \frac{Number\ of\ positives}{(Number\ of\ positives + Number\ of\ false\ positives)} \quad (8)$$

282
$$Specificity = \frac{Number\ of\ true\ negatives}{(Number\ of\ true\ negatives + Number\ of\ false\ negatives)} \quad (9)$$

283
$$AUC = \frac{\sum TP + \sum TN}{(P + N)} \quad (10)$$

284 where, TP (true positive) and TN (true negative) are truly classified pixel numbers, while FP (false
 285 positive) and FN (false negative) are falsely classified pixel numbers; P is the total number of
 286 floods and N is the total number of non-floods (Choubin et al. 2019; Khosravi et al. 2019).

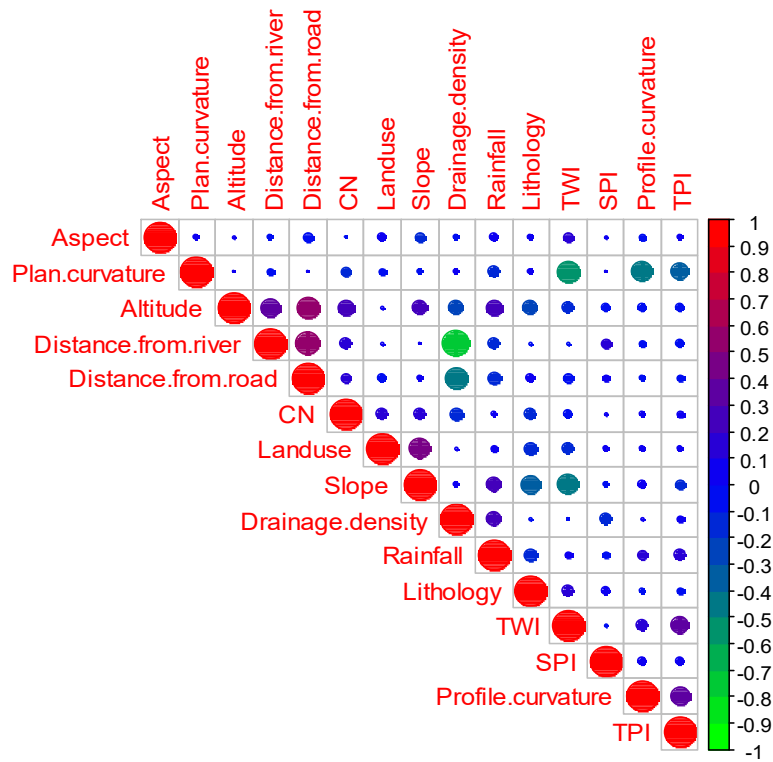
287

288 **3. Results**

289 **3.1. Analysis of Independent Variables**

290 In order to build a flood susceptibility model, potential model input variables are first analyzed for
 291 independence (via correlation) and linearity (via multi-collinearity analysis).

292 The results of the correlation study of the variables used in flood susceptibility modelling based
 293 on Spearman correlation test are shown in Fig.7. As it can be seen from this figure, the analyzed
 294 variables have a relatively low correlation with each other hence these were all selected for further
 295 analysis.



296

297

Fig. 7. Correlation analyses between independent variables

298

299 In order to determine the appropriate inputs for flood susceptibility modelling, multiple
 300 multiplexing and tolerance tests were used using *usdm* package (in the R software environment).

301 In order to investigate the linearity of the VIF range, all variables with VIF value smaller than 5
 302 were considered.

303 The results of multi-colinearity and tolerance analyses are shown in Table 2. The study of the
 304 linearity of the variables shows that all analyzed variables have a VIF value smaller than 5. The
 305 highest linearity was obtained for distance from the river with VIF equal to 2.39 and the tolerance
 306 equal to 0.42. The smallest linearity was obtained for the aspect variable with VIF of 1.07 and

307 tolerance of 0.93. Based on this, all variables shown in Table 2 are selected as potential inputs into
 308 the flood susceptibility model.

309 **Table 2. Multi-collinearity analysis base on VIF and Tolerance to determine the linearity of the independent**
 310 **variables**

Variables	VIF	Tolerance
Altitude	2.09	0.48
Aspect	1.07	0.93
Slope	1.57	0.64
Plan curvature	1.9	0.53
Profile Curvature	1.47	0.68
Drainage density	2.33	0.43
Distance from River	2.39	0.42
Distance from road	2.09	0.48
SPI	1.09	0.92
TPI	1.37	0.73
TWI	2.01	0.50
CN	1.35	0.74
Land use	1.29	0.77
Lithology	1.27	0.79
Rainfall	1.46	0.68

311

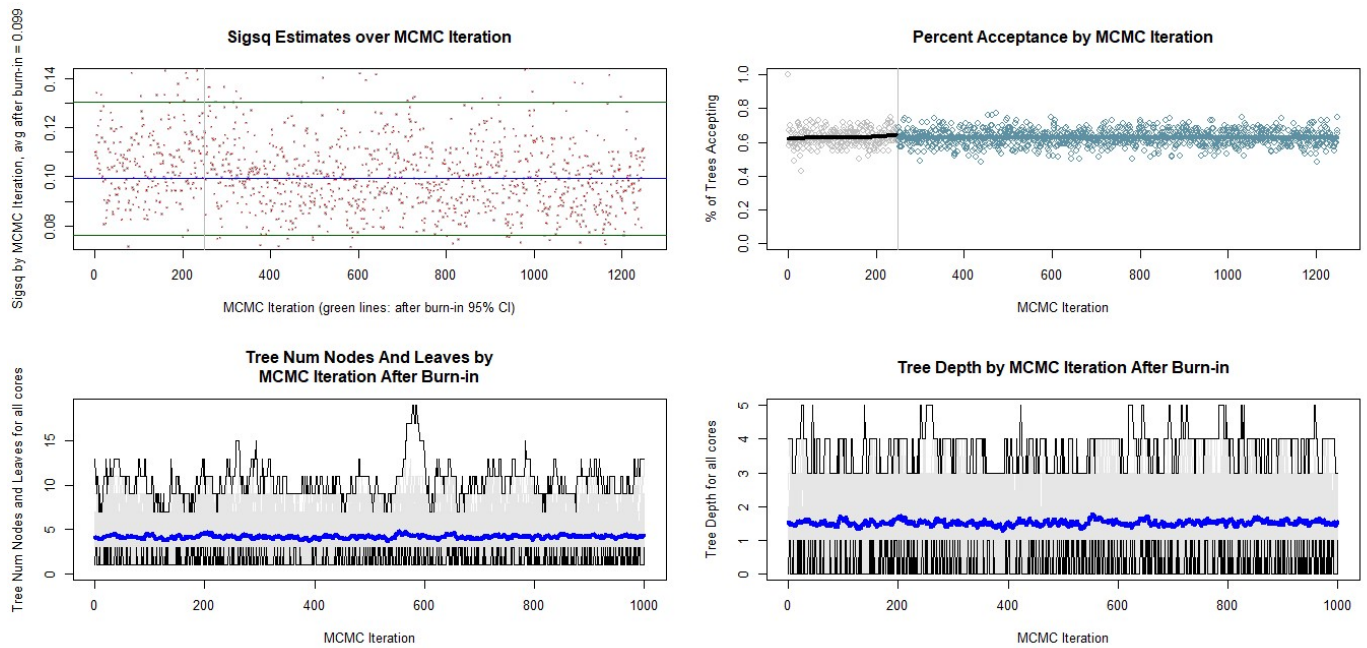
312 **3.2. Tuned parameters**

313 The tuned parameter values for the BART model are shown in Table 3 and Figure 8.

314 **Table 3. Tune parameters in BART model**

Parameters	Tuned value
Number of trees	100
Number burn in	500
Number iteration after burn in	1000
Alpha	0.95
Beta	2
K	2
Q	0.9

315



316

317

Fig. 8. The result of the BART model for flood susceptibility

318 3.3 Model Validation

319 ROC curves parameters include sensitivity, specificity, NPV, PPV and area under curve (AUC).

320 These parameters were used to evaluate the efficiency of Naïve Bayes, RF and BART models. The

321 corresponding results for the training and testing stages of these models are shown in Figs. 9 and

322 10 and Table 4.

323 According to the results obtained in the training phase, the sensitivity statistics in NB, RF and

324 BART models are equal to 0.76, 0.99 and 0.99, respectively. This shows the high sensitivity of the

325 three models and their accuracy. The specificity statistics for the NB, RF and BART models are

326 equal to 0.89, 0.95 and 0.90, respectively. The PPV statistics of 0.74, 0.95 and 0.91 and the NPV

327 statistics of 0.77, 0.99 and 0.98 were obtained for the NB, RF and BART models, respectively.

328 This shows the high accuracy of these models in predicting the non-occurrence points. The results

329 of model evaluation based on the AUC show that the accuracy of NB, RF and BART models is

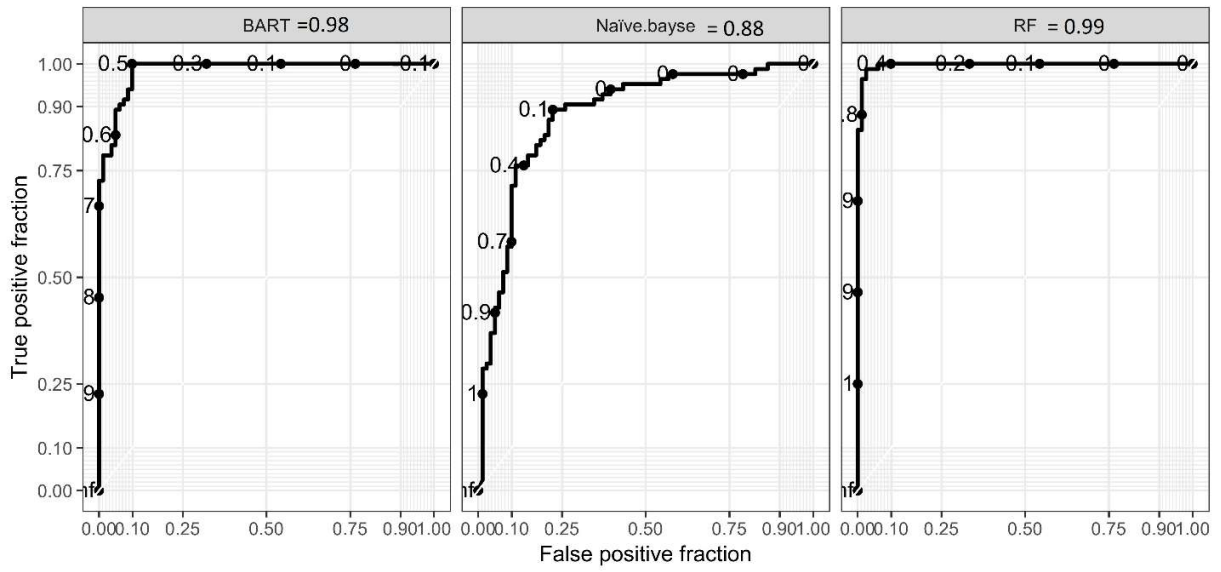
330 0.88, 0.99 and 0.89, respectively. Therefore, all three models have high predictive accuracy at the
 331 training stage.

332 Evaluation of the three models at the validation stage shows that the sensitivity statistics for NB,
 333 RF and BART models are equal to 0.76, 0.91 and 0.94, respectively. This shows the high
 334 sensitivity of these models in flood estimation. The specificity statistics in the NB, RF and BART
 335 models are equal to 0.75, 0.72 and 0.78, respectively. Evaluation of the same three models based
 336 on PPV and NPV statistics result in PPV values of 0.74, 0.75, 0.80, and NPV values of 0.77, 0.90,
 337 and 0.93 respectively, indicating high accuracy of these models when predicting non-flood points
 338 compared to flood points. For the overall evaluation of the models at the validation stage, the AUC
 339 statistic was used too and the values obtained for the NB, RF and BART models are equal to 0.81,
 340 0.85 and 0.89, respectively.

341 **Table 4. The results of evaluating the efficiency of Naïve Bayes, RF and BART models in train and validation**
 342 **stage**

Models	Stage	Parameters				
		Sensitivity	Specificity	PPV	NPV	AUC
Naïve	Train	0.76	0.89	0.87	0.78	0.88
Bayes	Validation	0.76	0.75	0.74	0.77	0.81
RF	Train	0.99	0.95	0.95	0.99	0.99
	Validation	0.91	0.72	0.75	0.90	0.85
BART	Train	0.99	0.90	0.91	0.98	0.98
	Validation	0.94	0.78	0.80	0.93	0.89

343

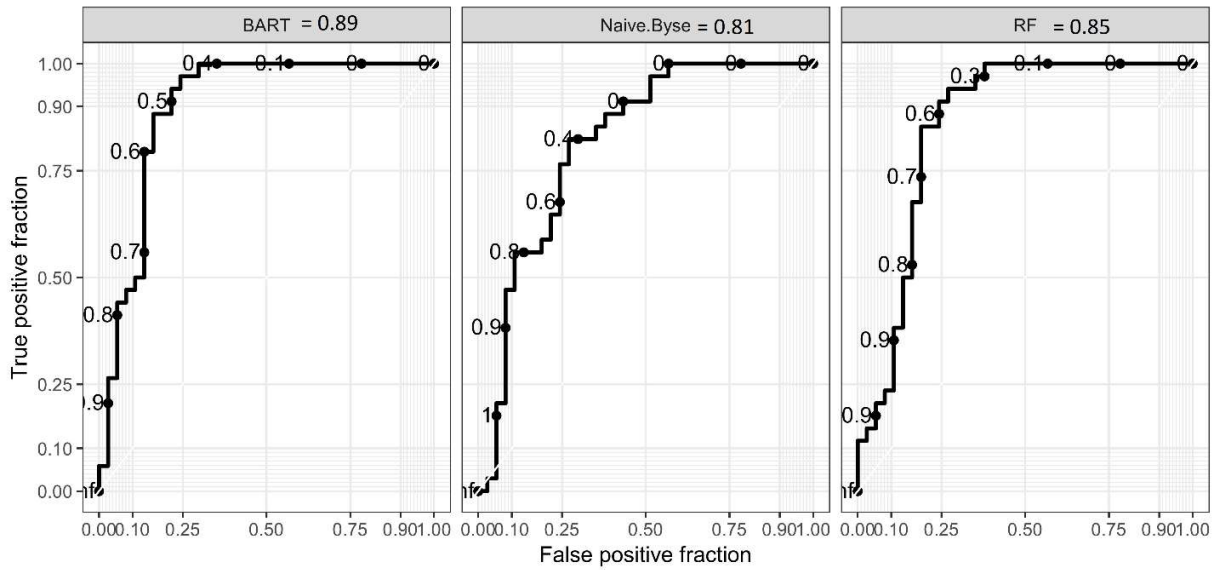


344

345

Fig. 9. The ROC curve analysis for Naïve Bayes, RF and BART models using the train dataset

346



347

348

Fig. 10. The ROC curve analysis for Naïve Bayes, RF and BART models using the testing dataset.

349

350

351

352

353 **3.4. Flood susceptibility modeling results**

354 After modelling the flood sensitivity using NB, RF and BART models and evaluating the
355 efficiency of these models, flood susceptibility was forecasted for the whole analyzed watershed.

356 The final map was divided into five flooding susceptibility classes (very low, low, moderate, high
357 and very high) by using the natural break algorithm (Fig. 11). According to the map obtained,
358 flooding susceptibility is the highest sensitivity around the main river and the areas near the outlet
359 of the watershed, which have a lower altitude. At the same time, most of the area analyzed, which
360 is generally high altitude, has a very low sensitivity.

361 The results of the area and percentage covered by each susceptibility class are shown in Table 5.

362 According to the results, the area of very high susceptibility class is equal to 22.11 km² (10.26%)
363 in the NB model, 21.23 km² (9.85%) in the RF model and 19.48 km² (9.04%) in the BART model.

364 However, the BART model, with 50.5 km² (23.5%) has predicted the largest area with very high
365 and high susceptibility classes.

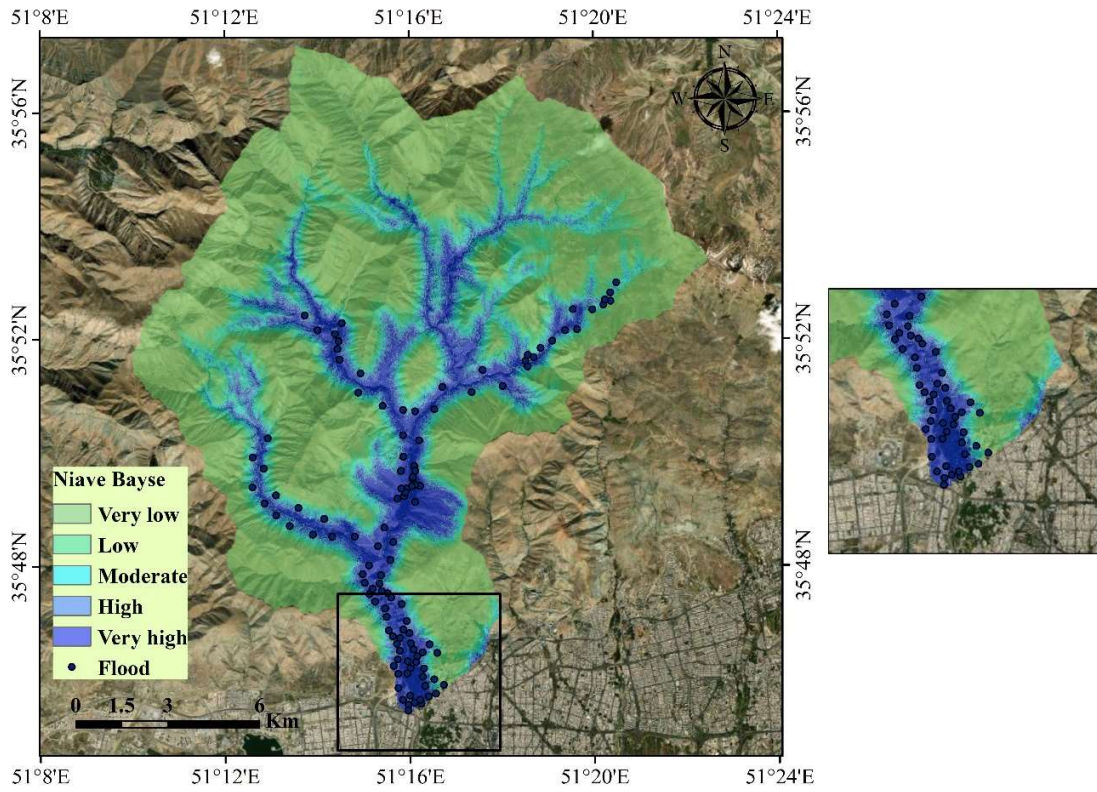
366 In order to evaluate the validity of the predicted flood susceptibility maps in relation to the
367 identified flood points in the study area, the frequency ratio (FR) approach was used (Fig. 9). As

368 it can be seen from Figure 12, the highest frequency ratio is in very high and high classes, which
369 indicates the appropriate prediction of the models used for flood-susceptibility areas. However,

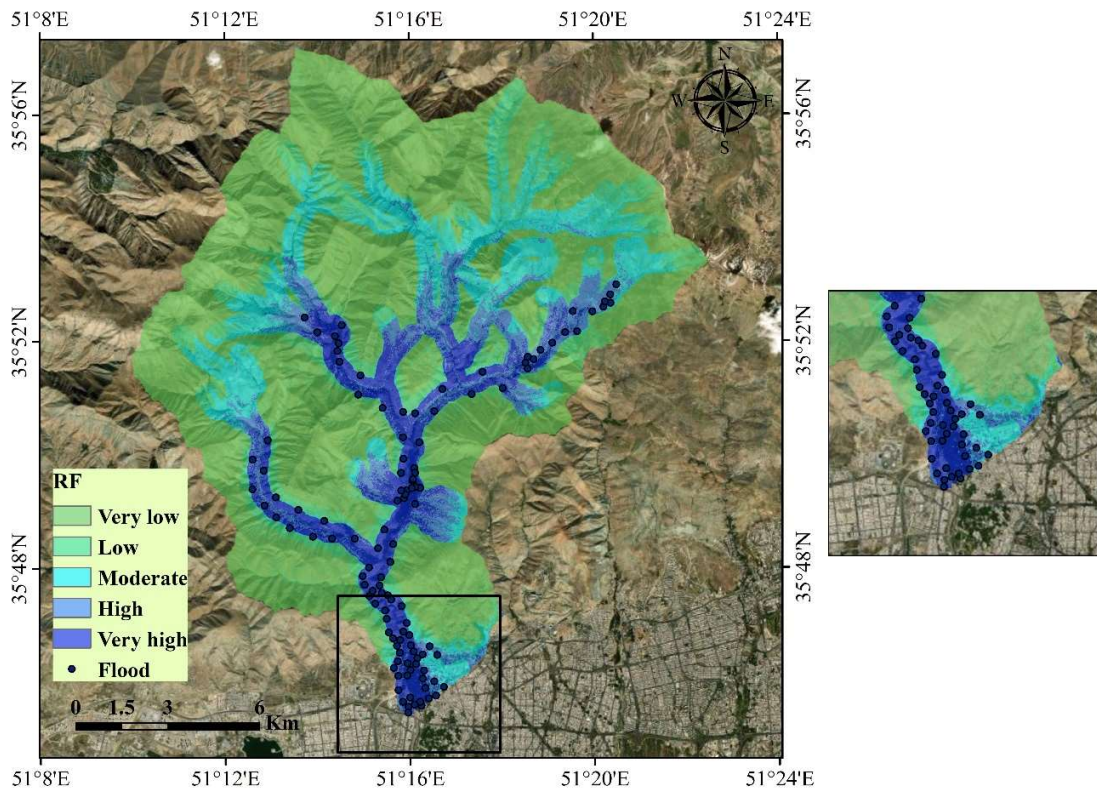
370 the predictions of the RF and BART models that are in the very high class are much higher than
371 the corresponding class predictions made by two other models, which indicates a more accurate

372 prediction of flood susceptibility in this area.

373

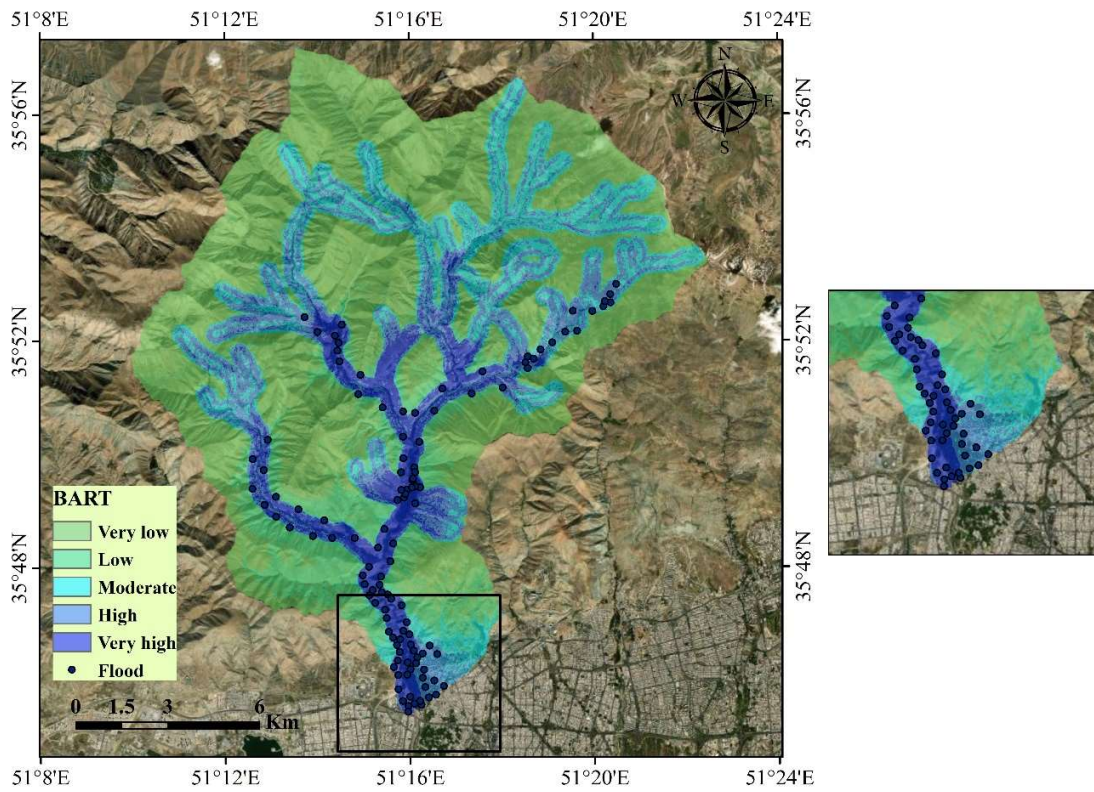


374



375

376



377

378

Fig. 11. Flood susceptibility map using the Naïve Bayes, RF and BART models

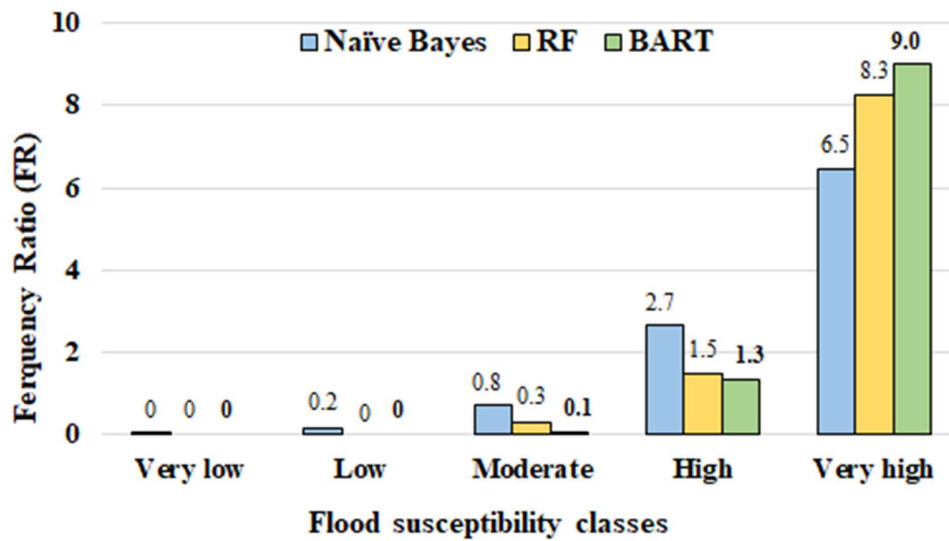
379

Table 5. The watershed area (in km² and %) in each flood susceptibility class

Susceptibility class	NB model		RF model		BART model	
	Area (km ²)	Area (%)	Area (km ²)	Area (%)	Area (km ²)	Area (%)
Very low	121.85	56.52	112.32	52.10	106	49.17
Low	31.53	14.62	33.24	15.42	28.39	13.17
Moderate	21.67	10.05	31.04	14.40	30.65	14.22
High	18.44	8.55	17.77	8.24	31.08	14.42
Very High	22.11	10.26	21.23	9.85	19.48	9.04

380

381



382

383 Fig. 12. Analysis of the frequency of floods on the flood susceptibility maps predicted using the FR method

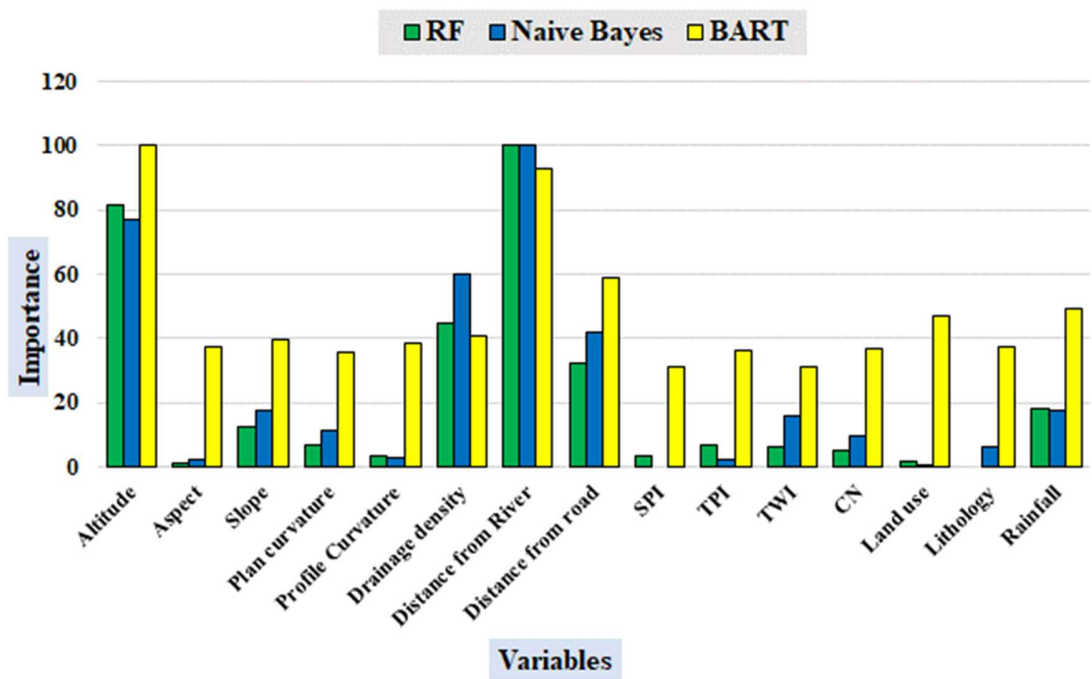
384 **3.5. Explanatory Variable Importance**

385 The results of the importance of the independent (i.e. input) variables used to model the flood
 386 susceptibility using the three models are shown in Fig. 13. It is clear that in the three models used
 387 different input variables have different effects on determining the flood susceptibility. It is also
 388 clear that altitude and distance from the river are more important than other variables in all three
 389 models.

390 Due to the importance of 4 variables (altitude, distance from the river, distance from the road and
 391 rainfall) on flood susceptibility in the BART model, these 4 variables were further investigated
 392 (Fig. 14). As it can be seen from Fig. 14, the flood susceptibility decreases with increasing altitude,
 393 with highest sensitivity to floods being at an altitude of 1400 meters (which is close to the altitude
 394 of the outlet of the watershed). This indicates the inverse relationship between the altitude and the
 395 flooding susceptibility. Further, a study of the distance from the river shows that locations with
 396 distances smaller than 500 meters have a high susceptibility to flooding whilst locations with

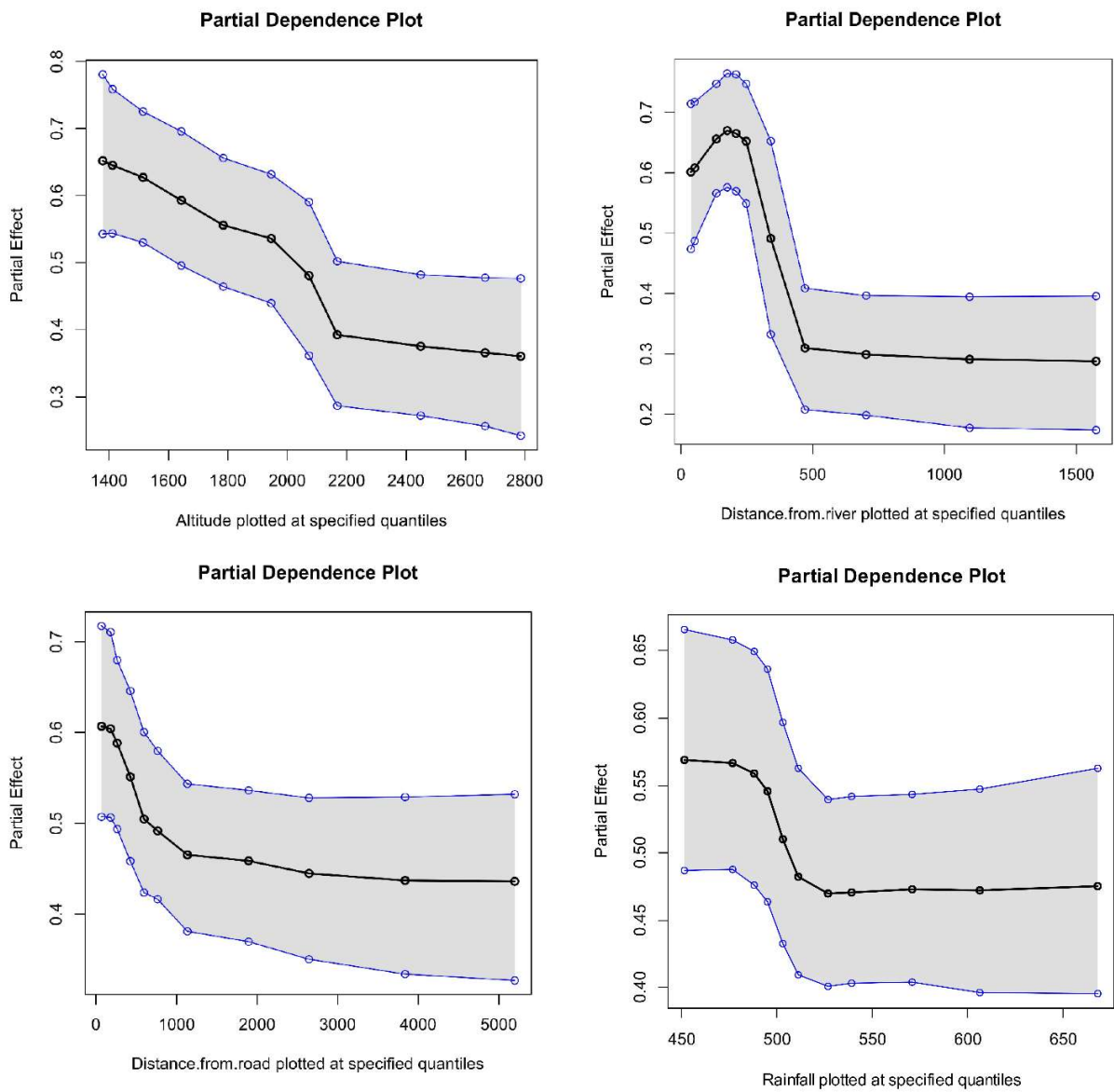
397 distances larger than 500 meters from the river have a decreasing flooding susceptibility which
 398 stabilizes around a low value for the distances of 1000-1500 meters. Regarding the distance from
 399 the road, it can be noted from Fig 14 that the flooding susceptibility decreases with the increasing
 400 distance from the road with most sensitive areas being located less than 1000 meters from the road.
 401 Finally, a study of the effect of rainfall on flooding susceptibility shows that areas with 450 to 500
 402 mm of rainfall per year are more sensitive than the areas with higher rainfall (the susceptibility
 403 decreases so that from rainfall 550 to 650 mm it is low and constant).

404



405

406 **Fig. 13. Results of relative importance of independent variables in flood sensitivity modeling in Naïve Bayes,**
 407 **RF and BART models**



408

409 **Fig 14. Partial effect plot for four importance variable (altitude, distance from river, distance from road and**
 410 **rainfall)**

411

412 **4. Discussion**

413 In the present study, we developed and presented a novel flood susceptibility BART model that is
 414 based on machine learning and Bayesian approach. In addition, two existing models, NB and RF

415 were used for comparison. The results obtained showed that all three models have a high
416 performance in predicting the flooding susceptibility in the Kan watershed in Iran but, based on
417 the model performance criteria, the new BART model has outperformed the other two models. In
418 terms of input variable importance, the results obtained show that the altitude and distance from
419 the river are the most important variables for assessing flooding susceptibility in the study area.

420 One of the main objectives of this study was to apply the BART model and evaluate the efficiency
421 of this model in flood modeling in the study area. Performance evaluation of NB, RF, and BART
422 models shows that the BART model performed best in the validation stage in terms of predicting
423 flood susceptibility. The use of the Bart model in Natural Hazard studies and especially flood
424 sensitivity modeling has been reported rarely before. The efficiency of this model has been proven
425 in other fields such as forest science. Ahmadi et al. (2021) used BART model to mapping forest
426 stand characteristics and showed that this model has a high performance in comparison to other
427 models.

428 The BART model is a non-parametric Bayesian regression approach that uses consistent basic
429 random elements. Bayesian Additive Regression Trees (BART) provides a flexible way to fit a
430 variety of regression models while avoiding strong parametric assumptions (Hill et al. 2020). The
431 tree ensemble model is supported by an uncertainty framework in the Bayesian inferential
432 framework and provides a principled approach to regulation through previous specifications
433 (Pratola and Higdon 2016; Sparapani et al. 2016). This model uses a non-parametric tree
434 aggregation model to allow flexibility of the average structure of a regression. But it also has the
435 advantages of a Bayesian inferential framework given the amount of uncertainty and its regulation
436 through calibrated data locations (Sparapani et al. 2016; Hill et al. 2020; Prado et al. 2021; Wu et
437 al. 2021).

438 One of the main advantages of the BART model is the capacity to form inference on numerous
439 features of the survival distribution directly from the posterior samples. As a Bayesian model,
440 BART consists of a set of priors for the construction and the leaf parameters and a possibility for
441 data in the terminal nodes (Pratola and Higdon 2016; Sparapani et al. 2016). The object of the
442 priors is to afford regularization, limiting any single regression tree from dominating the total fit.
443 Many Machine learning (ML) models suffer from missing data problems. BART model has a
444 specialty that provides the user with the straight designation missing covariate data within the
445 BART structure. This method combines missing data indicators into the training data set and
446 supports for divisions on the missing indicators, guiding to raised efficiency under a pattern
447 ensemble model structure (Hill et al. 2020; Prado et al. 2021; Sparapani et al. 2021).

448 Determining the importance of independent variables in flood susceptibility modeling in the Kan
449 watershed showed that altitude, distance from river, distance from road and rainfall variables are
450 important factors affecting flood susceptibility in this region. A study of altitude variable shows
451 that low altitudes, which are often at the outlet of watersheds, are highly susceptible to flooding,
452 which is consistent with the findings of Khosravi et al. (2019), Pham et al., (2020a).

453 Distance from river is another important factor in flood susceptibility in the Kan watershed, and
454 the results indicate the sensitivity of areas close to the river. Ahmadlou et al., (2019) showed in
455 their studies that areas 500-1000 meters from the river are highly sensitive to flooding. Given that
456 the flood-prone areas are located near the river and the reason is due to rise of flow from the river
457 channels (Choubin et al. 2019; Darabi et al. 2019; Panahi et al. 2021), in the Kan watershed, due
458 to lack of observance of riverbed and river boundaries, several restaurants and villas have been
459 built in the areas near the river, and due to the presence of more orchard in the river area, has led
460 to the obstruction of flow in these areas and has increased the sudden release of flood current.

461 Invasion of the river boundaries and the create of orchard in it, in addition to causing financial
462 damage to the residents of the area, also by blocking the flow in sections such as tunnels, will
463 cause secondary floods and intensify the damage to the people and downstream areas.

464 Another factor affecting the flood susceptibility in the Kan watershed is the distance from road.
465 Construction and crate of communication roads will increase the runoff and runoff speed because
466 it will reduce the area of the existing surface to absorb rainfall and thus will increase the sensitivity
467 to flooding in these areas (Tehrany et al. 2019b; Zhao et al. 2019).

468 The study of the rainfall indicates that areas with less rainfall are highly susceptible to flood, which
469 are mainly areas close to the outlet of the Kan watershed. Due to the mountainous nature of the
470 region, most of the precipitation in the upstream areas of the Kan watershed is snow, so in these
471 areas the possibility of infiltration is higher. In addition, precipitation in the downstream areas is
472 in the form of storms and these storms are usually more severe in the autumn and causes the river
473 inundation and flooding.

474 In recent years, due to human interventions and the resulting climate and land-use changes, the
475 rate of flooding and the corresponding damages have increased significantly. Studies such as this
476 one allow managers to reduce flood risks through planning and flood susceptibility analysis.
477 Therefore, we are always looking for more accurate modeling approaches to reduce the bias in the
478 prediction of flood susceptibility. In the present study, we showed that BART model is an accurate
479 model that can be used for effective flood susceptibility modeling. This model can be applied in
480 the future along with other modes that have shown high ability in flood modeling studies.

481

482 **5. Conclusion**

483 Floods are one of the most frequent and destructive natural disasters that can cause a lot of damage.
484 In order to investigate and analyze the susceptibility of some are to flooding, different methods
485 have been developed by the researchers.

486 In this study, the Bayesian based model (Naïve Bayes), regression tree type model (Random
487 Forest) and ensemble type model (Bayesian Additive Regression Tree - BART) were developed
488 to predict flood susceptibility in the Kan watershed. A total of 15 explanatory (i.e. model input)
489 variables were used after multi-collinearity analyses as independent variables and 118 flood
490 locations and 115 non-flood locations after field surveys and the use of available information as a
491 dependent variable for flood modeling.

492 The validation results obtained for flood susceptibility modeling showed that the Naïve Bayes, RF
493 and BART models all have a good predictive performance. However, the new BART model has
494 the higher prediction accuracy than the Naïve Bayes and RF models. This is due to the fact that it
495 uses features of both methods in the ensemble setting.

496 The analysis of the importance of explanatory variables showed that the effect of independent
497 variables is different in each model. However, the altitude and distance from the river were more
498 important than other variables in all three models meaning that low-height areas and areas close to
499 the river are more susceptible to flooding.

500 The Kan watershed is close to the city of Tehran and the pleasant climate of this tourist area has
501 caused that its riverbanks are occupied with many constructions that have been carried out. These
502 areas receive a large number of tourists in spring and summer and hence are strongly affected by
503 the floods. It is therefore necessary to provide flood hazard maps for the region. The results of this

504 research can be used as a baseline map in development projects to determine areas susceptible to
505 flooding hence prevent the construction in these high-risk areas.

506 **Author Contributions:** Saeid Janizadeh acquired the data; Saeid Janizadeh and Mehdi Vafakhah
507 conceptualized and performed the analysis; Saeid Janizadeh wrote the manuscript and discussion,
508 and analyzed the data; Mehdi Vafakhah, Zoran Kapelan and Naghmeh Mobarghaee Dinan
509 provided technical sights, as well as edited, restructured, and professionally optimized the
510 manuscript. All authors discussed the results and edited the manuscript. All authors have read and
511 agreed to the published version of the manuscript.

512 **Ethical Approval:** We confirm that this manuscript has not been published elsewhere and is not
513 under consideration by another journal.

514 **Consent to Participate:** All authors have participated the manuscript and agree with submission
515 to Water Resources Management.

516 **Consent to Publish:** All authors have approved the publication of this manuscript in the Water
517 Resources Management Journal.

518 **Availability of data and materials:** We have no permission to release data and codes.

519

520 **Funding:** The authors received no specific funding for this work.

521 **Acknowledgments:** We acknowledge Tarbiat Modares University's support for this work.

522 **Conflicts of Interest:** The authors declare no conflict of interest.

523 **References**

524 Ahmadlou M, Karimi M, Alizadeh S, et al (2019) Flood susceptibility assessment using integration of
525 adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization

526 (BBO) and BAT algorithms (BA). *Geocarto Int* 34:1252–1272

527 Al-Abadi AM (2018) Mapping flood susceptibility in an arid region of southern Iraq using ensemble
528 machine learning classifiers: a comparative study. *Arab J Geosci* 11:218

529 Al-Juaidi AEM, Nassar AM, Al-Juaidi OEM (2018) Evaluation of flood susceptibility mapping using
530 logistic regression and GIS conditioning factors. *Arab J Geosci* 11:765

531 Arabameri A, Saha S, Chen W, et al (2020) Flash flood susceptibility modelling using functional tree and
532 hybrid ensemble techniques. *J Hydrol* 125007

533 Bui DT, Panahi M, Shahabi H, et al (2018) Novel hybrid evolutionary algorithms for spatial prediction of
534 floods. *Sci Rep* 8:15364

535 Chapi K, Singh VP, Shirzadi A, et al (2017) A novel hybrid artificial intelligence approach for flood
536 susceptibility assessment. *Environ Model Softw* 95:229–245

537 Chen W, Li Y, Xue W, et al (2020) Modeling flood susceptibility using data-driven approaches of
538 naïve bayes tree, alternating decision tree, and random forest methods. *Sci Total Environ*
539 701:134979

540 Choubin B, Moradi E, Golshan M, et al (2019) An ensemble prediction of flood susceptibility using
541 multivariate discriminant analysis, classification and regression trees, and support vector machines.
542 *Sci Total Environ* 651:2087–2096

543 Chowdhuri I, Pal SC, Arabameri A, et al (2020) Implementation of artificial intelligence based ensemble
544 models for gully erosion susceptibility assessment. *Remote Sens* 12:3620

545 Cook A, Merwade V (2009) Effect of topographic data, geometric configuration and modeling approach
546 on flood inundation mapping. *J Hydrol* 377:131–142

547 Costache R (2019) Flash-flood Potential Index mapping using weights of evidence, decision Trees
548 models and their novel hybrid integration. *Stoch Environ Res Risk Assess* 33:1375–1402

549 Costache R, Arabameri A, Blaschke T, et al (2021) Flash-Flood Potential Mapping Using Deep Learning,
550 Alternating Decision Trees and Data Provided by Remote Sensing Sensors. *Sensors* 21:280.
551 <https://doi.org/10.3390/s21010280>

552 Costache R, Bui DT (2020) Identification of areas prone to flash-flood phenomena using multiple-criteria
553 decision-making, bivariate statistics, machine learning and their ensembles. *Sci Total Environ*
554 712:136492

555 Darabi H, Choubin B, Rahmati O, et al (2019) Urban flood risk mapping using the GARP and QUEST
556 models: A comparative study of machine learning techniques. *J Hydrol* 569:142–154

557 Delkash M, Al-Faraj FAM, Scholz M (2014) Comparing the export coefficient approach with the soil and
558 water assessment tool to predict phosphorous pollution: the Kan watershed case study. *Water, Air,
559 Soil Pollut* 225:2122

560 El-Magd SAA, Pradhan B, Alamri A (2021) Machine learning algorithm for flash flood prediction
561 mapping in Wadi El-Laqeita and surroundings, Central Eastern Desert, Egypt. *Arab J Geosci* 14:1–
562 14

563 Frattini P, Crosta G, Carrara A (2010) Techniques for evaluating the performance of landslide
564 susceptibility models. *Eng Geol* 111:62–72

565 Heidari A (2014) Flood vulnerability of the K arun R iver S ystem and short-term mitigation measures. *J*
566 *flood risk Manag* 7:65–80

567 Hill J, Linero A, Murray J (2020) Bayesian additive regression trees: A review and look forward. *Annu*
568 *Rev Stat Its Appl* 7:251–278

569 Hong H, Panahi M, Shirzadi A, et al (2018) Flood susceptibility assessment in Hengfeng area coupling
570 adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Sci Total*
571 *Environ* 621:1124–1141

572 Hooshyaripor F, Faraji-Ashkavar S, Koohyian F, et al (2020) Annual flood damage influenced by El Niño
573 in the Kan River basin, Iran. *Nat Hazards Earth Syst Sci* 20:2739–2751

574 Hosseini FS, Choubin B, Mosavi A, et al (2020) Flash-flood hazard assessment using ensembles and
575 Bayesian-based machine learning models: application of the simulated annealing feature selection
576 method. *Sci Total Environ* 711:135161

577 Islam ARMT, Talukdar S, Mahato S, et al (2021) Flood susceptibility modelling using advanced

578 ensemble machine learning models. *Geosci Front* 12:101075

579 Janizadeh S, Avand M, Jaafari A, et al (2019) Prediction Success of Machine Learning Methods for Flash
580 Flood Susceptibility Mapping in the Tafresh Watershed, Iran. *Sustainability* 11:5426

581 Kalantar B, Ueda N, Saeidi V, et al (2021) Deep Neural Network Utilizing Remote Sensing Datasets for
582 Flood Hazard Susceptibility Mapping in Brisbane, Australia. *Remote Sens* 13:2638

583 Kapelner A, Bleich J (2013) bartMachine: Machine learning with Bayesian additive regression trees.
584 arXiv Prepr arXiv13122171

585 Khosravi K, Pham BT, Chapi K, et al (2018) A comparative assessment of decision trees algorithms for
586 flash flood susceptibility modeling at Haraz watershed, northern Iran. *Sci Total Environ* 627:744–
587 755

588 Khosravi K, Pourghasemi HR, Chapi K, Bahri M (2016) Flash flood susceptibility analysis and its
589 mapping using different bivariate models in Iran: a comparison between Shannon’s entropy,
590 statistical index, and weighting factor models. *Environ Monit Assess* 188:656

591 Khosravi K, Shahabi H, Pham BT, et al (2019) A comparative assessment of flood susceptibility
592 modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods. *J Hydrol*
593 573:311–323

594 Liaw A, Wiener M, others (2002) Classification and regression by randomForest. *R news* 2:18–22

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

597 Mahmoud SH, Gan TY (2018) Multi-criteria approach to develop flood susceptibility maps in arid
598 regions of Middle East. *J Clean Prod* 196:216–229

599 Miles J (2014) Tolerance and variance inflation factor. *Wiley StatsRef Stat Ref Online*

600 Molinos-Senante M, Hernández-Sancho F, Sala-Garrido R (2011) Cost-benefit analysis of water-reuse
601 projects for environmental purposes: A case study for Spanish wastewater treatment plants. *J*
602 *Environ Manage* 92:3091–3097

603 Ngo P-T, Hoang N-D, Pradhan B, et al (2018) A Novel Hybrid Swarm Optimized Multilayer Neural

604 Network for Spatial Prediction of Flash Floods in Tropical Areas Using Sentinel-1 SAR Imagery
605 and Geospatial Data. *Sensors* 18:3704. <https://doi.org/10.3390/s18113704>

606 Panahi M, Dodangeh E, Rezaie F, et al (2021) Flood spatial prediction modeling using a hybrid of meta-
607 optimization and support vector regression modeling. *Catena* 199:105114

608 Papaioannou G, Vasiliades L, Loukas A (2015) Multi-criteria analysis framework for potential flood
609 prone areas mapping. *Water Resour Manag* 29:399–418

610 Pham BT, Avand M, Janizadeh S, et al (2020a) GIS Based Hybrid Computational Approaches for Flash
611 Flood Susceptibility Assessment. *Water* 12:683

612 Pham BT, Phong T Van, Nguyen HD, et al (2020b) A Comparative Study of Kernel Logistic Regression,
613 Radial Basis Function Classifier, Multinomial Naïve Bayes, and Logistic Model Tree for Flash
614 Flood Susceptibility Mapping. *Water* 12:239

615 Plant E, King R, Kath J (2021) Statistical comparison of additive regression tree methods on ecological
616 grassland data. *Ecol Inform* 61:101198

617 Prado EB, Moral RA, Parnell AC (2021) Bayesian additive regression trees with model trees. *Stat*
618 *Comput* 31:1–13

619 Prasad P, Loveson VJ, Das B, Kotha M (2021) Novel Ensemble Machine Learning Models in Flood
620 Susceptibility Mapping. *Geocarto Int* 1–22. <https://doi.org/10.1080/10106049.2021.1892209>

621 Pratola MT, Higdon DM (2016) Bayesian additive regression tree calibration of complex high-
622 dimensional computer models. *Technometrics* 58:166–179

623 Rahmati O, Pourghasemi HR, Zeinivand H (2016) Flood susceptibility mapping using frequency ratio and
624 weights-of-evidence models in the Golastan Province, Iran. *Geocarto Int* 31:42–70

625 Rish I, others (2001) An empirical study of the naive Bayes classifier. In: *IJCAI 2001 workshop on*
626 *empirical methods in artificial intelligence*. pp 41–46

627 Shafapour Tehrany M, Shabani F, Neamah Jebur M, et al (2017) GIS-based spatial prediction of flood
628 prone areas using standalone frequency ratio, logistic regression, weight of evidence and their
629 ensemble techniques. *Geomatics, Nat Hazards Risk* 8:1538–1561

630 Shafizadeh-Moghadam H, Valavi R, Shahabi H, et al (2018) Novel forecasting approaches using
631 combination of machine learning and statistical models for flood susceptibility mapping. *J Environ*
632 *Manage* 217:1–11

633 Shahabi H, Shirzadi A, Ghaderi K, et al (2020) Flood detection and susceptibility mapping using sentinel-
634 1 remote sensing data and a machine learning approach: Hybrid intelligence of bagging ensemble
635 based on k-nearest neighbor classifier. *Remote Sens* 12:266

636 Sparapani R, Spanbauer C, McCulloch R (2021) Nonparametric machine learning and efficient
637 computation with bayesian additive regression trees: the BART R package. *J Stat Softw* 97:1–66

638 Sparapani RA, Logan BR, McCulloch RE, Laud PW (2016) Nonparametric survival analysis using
639 Bayesian additive regression trees (BART). *Stat Med* 35:2741–2753

640 Talukdar S, Ghose B, Salam R, et al (2020) Flood susceptibility modeling in Teesta River basin,
641 Bangladesh using novel ensembles of bagging algorithms. *Stoch Environ Res Risk Assess* 34:2277–
642 2300

643 Tang X, Li J, Liu M, et al (2020) Flood susceptibility assessment based on a novel random Naⁱve
644 Bayes method: A comparison between different factor discretization methods. *Catena* 190:104536

645 Tang Z, Yi S, Wang C, Xiao Y (2018) Incorporating probabilistic approach into local multi-criteria
646 decision analysis for flood susceptibility assessment. *Stoch Environ Res risk Assess* 32:701–714

647 Tehrany MS, Jones S, Shabani F (2019a) Identifying the essential flood conditioning factors for flood
648 prone area mapping using machine learning techniques. *Catena* 175:174–192

649 Tehrany MS, Kumar L (2018) The application of a Dempster--Shafer-based evidential belief function in
650 flood susceptibility mapping and comparison with frequency ratio and logistic regression methods.
651 *Environ Earth Sci* 77:490

652 Tehrany MS, Kumar L, Shabani F (2019b) A novel GIS-based ensemble technique for flood
653 susceptibility mapping using evidential belief function and support vector machine: Brisbane,
654 Australia. *PeerJ* 7:e7653

655 Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-

656 of-evidence and support vector machine models in GIS. *J Hydrol* 512:332–343

657 Tehrany MS, Pradhan B, Mansor S, Ahmad N (2015) Flood susceptibility assessment using GIS-based
658 support vector machine model with different kernel types. *Catena* 125:91–101

659 Vafakhah M, Loor SMH, Pourghasemi H, Katebikord A (2020) Comparing performance of random forest
660 and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping.
661 *Arab J Geosci* 13:417

662 Vetrivel A, Gerke M, Kerle N, et al (2018) Disaster damage detection through synergistic use of deep
663 learning and 3D point cloud features derived from very high resolution oblique aerial images, and
664 multiple-kernel-learning. *ISPRS J Photogramm Remote Sens* 140:45–59

665 Woodward M, Kapelan Z, Gouldby B (2014) Adaptive flood risk management under climate change
666 uncertainty using real options and optimization. *Risk Anal* 34:75–92

667 Wu W, Tang X, Lv J, et al (2021) Potential of Bayesian additive regression trees for predicting daily
668 global and diffuse solar radiation in arid and humid areas. *Renew Energy*

669 Yariyan P, Janizadeh S, Van Phong T, et al (2020) Improvement of Best First Decision Trees Using
670 Bagging and Dagging Ensembles for Flood Probability Mapping. *Water Resour Manag* 1–17

671 Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and
672 neural networks methods in a medium scale study, Hendek region (Turkey). *Eng Geol* 79:251–266

673 Zhang H (2004) The Optimality of Naive Bayes, 2004. *Am Assoc Artif Intell* (www.aaai.org)

674 Zhao G, Pang B, Xu Z, et al (2019) Assessment of urban flood susceptibility using semi-supervised
675 machine learning model. *Sci Total Environ* 659:940–949

676