Earth Science Informatics (2021) 14:2317–2328
https://doi.org/10.1007/s12145-021-00700-8

RESEARCH ARTICLE

Prediction of urban water accumulation points and water accumulation process based on machine learning

Hongfa Wang[1] · Yajuan Zhao[2] · Yihong Zhou[1] · Huiliang Wang[1]

Received: 21 April 2021 / Accepted: 2 September 2021 / Published online: 18 September 2021

Abstract

With the development of urbanization, global warming, rain island effect and other factors, cities around the world are facing more frequent and intense flood events. In order to deal with the damage caused by urban flood effectively, it is increasingly important to accurately predict and characterize the information of the flood in cities. In recent years, the rise of machine learning methods provides a new technical means for flood prediction. In this study, Naive Bayes (NB) and Random Forest (RF) algorithm were used to forecast the waterlogging point and the waterlogging process at the waterlogging point respectively to achieve the goal of predicting the whole process of urban waterlogging. Compared with the actual result, the four evaluation indexes (P, R, A and F_1) of the NB classification models are 91%, 90.5%, 98.9% and 90.7% respectively, and the three regression indexes (MAE, MRER and RMSE) of the RF regression model were respectively 0.95%, 9.53% and 1.21%. The results demonstrated that the prediction result of NB model for waterlogging point is reliable, and the process of waterlogging predicted by RF model is also consistent with the actual situation, which verify the validity and applicability of the NB model and RF model. This research is expected to provide scientific guidance and theoretical support for urban flood disaster mitigation and relief work.

Keywords Urban flood · Naive Bayes classification model · Random forest regression model · Waterlogging points prediction · Real-time depth prediction

Introduction

Extreme weather and climatic events, particularly flooding, have caused a huge impact on people's life and property and social development (Jamshed et al. 2021). In the twentieth century, the number of deaths caused by catastrophic floods has ranged from 100,000 to 1.4 million, according to published national statistics (Hajat et al. 2003). Recent studies have reported that floods, one of the natural disasters caused by extreme weather and climate events, are becoming more frequent and intense (Hirabayashi et al. 2013; IPCC 2014). It is estimated that from 2000 to 2020, flood events have caused economic losses of more than $537 billion globally, affecting the normal life of 1.6 billion people (EM-DAT 2020). With increasing impervious cover in urban areas driving dramatic changes in rainfall infiltration and storage capacity (Mu et al. 2020), which lead that urban flood appear sudden and frequent (Ward 1978), posing severe challenges to urban flood control and drainage. Cities gather a large number of talents, creating an economy that occupies an absolute advantage in the overall economic proportion, which leads to urban floods affect a large number of people worldwide, causing human fatalities and significant damages (Rahmati et al. 2020). After the Louisiana floods in 2016, the floods in Shouguang and Zhengzhou in China in 2018, and the floods in Iran on March 25 in 2019, as examples, these heavy rains and floods caused considerable economic losses and casualties, and have become prominent bottlenecks affecting the healthy development of cities (Yazdi et al. 2019). The main reason for urban flood is that urbanization increases hardened area, reduces infiltration, increases runoff and triggers higher and faster peak water flow (Loperfideo et al. 2014;

Communicated by H. Babaie

✉ Huiliang Wang
wanghuiliang@zzu.edu.cn

[1] College of Water Conservancy Engineering, Zhengzhou University, Zhengzhou 450001, Henan, People's Republic of China

[2] Yongcheng Vocational College, Yongcheng 476600, Henan, People's Republic of China

Ferreira et al. 2016). These changes have a considerable impact on the hydrological process when rainfall occurs, resulting in a large and rapid runoff generation, coupled with the failure of storm drainage system (GebreEgziabher and Demissie 2020), resulting in a higher probability of urban flood occurrence and a higher recurrence rate (Braud et al. 2013; Miller et al, 2014; Jongman 2018).

To assist decision makers in anticipating potential flooded and preemptively taking measures to lleviate the pressure brought by urban floods, promote the steady development of cities and ensure the safety of people's lives and property, researchers and practitioners have done a lot of research on urban flood prediction (Bhan and Team 2001; Diaz-Nieto et al. 2012; Gain and Hoque 2013; Kong et al. 2017). Hydrological and hydrodynamic models and data-driven models are the most popular and widely used tools in the research of early warning and forecast of urban flood information (White and Greer 2006; Bubeck et al. 2016).

Hydrological and hydrodynamic models are based on hydrological characteristics, which can physically describe runoff confluence by combining the physical laws of mass momentum and energy conservation (Vojinovic and Tutulic 2009). SWMM (Zhao et al. 2009; Huong and Pathirana 2013), Mike (Zoppou 2001; Zolch et al. 2017) and InfoWorks (Schmitt et al. 2004) are widely used hydrological and hydrodynamic models in flood prediction. Zhang et al. built an urban flood model based on SWMM to predict the flood disaster and pipeline drainage process under different types of designed rainfall, based on the data of topographic map underground drainage network, urban land use and rainfall. The results prove the applicability of SWMM in urban rainstorm flood simulation and drainage analysis of pipe network (Zhang and Li 2019). Wu et al. (2017) established a two-dimensional hydrodynamic inundation model through the coupling of SWMM and LISFlood-FP model, and on this basis revealed the evolution law of the inundation of Shiqiaoxi District (SCD) of Dongguan City under different scenarios of sea level rise and subsidence under heavy rain. Patro et al. (2009) took the data results of MIKE11 as the input of the two-dimensional model MIKE 21, coupled the MIKE11 model and the MIKE 21 model laterally to form the two-dimensional flood inundation simulation MIKE flood model in the study area, and carried out numerical simulation on the flood inundation range and flood inundation depth. Bisht et al. (2016) used the two-dimensional (2D) MIKE model to overcome the limitations of the one-dimensional (1D) SWMM model in simulating the flood range and flood inundation, and simulated the flood in a small urbanized area in West Bengal, India. The InfoWorks ICM 2D hydrodynamic model is utilized for simulating historical and designed rainfall events, which is carried out in the "Sponge City Construction" pilot area of Jinan City. The simulated water depth and flow velocity are recorded for flood risk zoning and the result shows that the InfoWorks ICM 2D model performed well (Cheng et al. 2017).

The data-driven intelligent model does not need to consider the specific process of the model. It is mainly manifested as the analysis and learning of the existing observation data, so as to establish the mapping relationship between input and output, so as to predict the specific variables (Nourani et al. 2009; Jhong et al. 2016). Ding et al. (2020) proposed an explicable spatiotemporal attention long—short memory model (STA-LSTM) based on LSTM and attention mechanism, and established the model using dynamic attention mechanism and LSTM method to make explicable analysis of flood prediction. Granata et al. (2016) predicted the runoff due to rainfall through support vector regression (SVR) and compared the results with those of the SWMM model. The results of SWMM overestimated the runoff compared to those of SVR. Kim and Han (2020) established flood prediction models for various basins by introducing nonlinear autoregressive model and self-organizing map (NARX-SOM), and carried out flood prediction for the extremely heavy rainstorm in Seoul, South Korea in 2010 and 2011, with high prediction ability. She and You (2019) combined the architectural advantages of radial basis function neural network (RBFNN) and nonlinear autoregressive and exogenous input neural network (NARXNN) and proposed the RBFM prediction model to predict the urban drainage system flow, which proved the great potential of RNFM in urban runoff prediction and management. Wu et al. (2020a) established a real-time prediction model of flood depth based on waterlogging point by using GBDT algorithm based on multi-factor analysis, and verified the validity and applicability of the model for real-time prediction of waterlogging process. However, the model that Wu used only be predicted when rainfall occurs, and cannot predict the flood depth after rainfall.

The above studies have achieved good results in the field of flood prediction. However, the current research results still focus on the prediction of a single aspect of the depth range of urban flood and the duration of water retention, which leads to the failure to make appropriate decisions in time to avoid the damage caused by the flood disaster (Yazdi and Neyshabouri 2012; Wu et al. 2020b). Moreover, studies on the spatial flood prediction for large urban basins are not sufficient. As such, this paper intends to use the Naive Bayes algorithm and random forest algorithm in machine learning to forecast the information of urban waterlogging generated by rainfall. The specific contents are as follows: 1. According to the rainfall data information, a classifier model based on Naive Bayes algorithm is constructed to analyze and predict the urban waterlogging point; 2. Construct a regression prediction model based on random forest algorithm, and use real-time rainfall and water accumulation information to make short real-time prediction of water accumulation process at waterlogging points. From the determination of the waterlogging point to the prediction of the water level at the waterlogging point, the prediction research of the whole process of urban waterlogging is realized, which provides technical support for urban flood control management.

Materials and methods

Study area

Zhengzhou, the capital of Henan Province in Central China, covers an area of approximately 7446 km^2 (Fig. 1). Its permanent resident population reached 10.352,000 by the end of 2019, ranking 14th in China. Among them, the urban population was 7.721 million, with an urbanization rate of 74.6%. As an important hub city on the "new Silk Road" in Europe and Asia, Zhengzhou's the total GDP (Gross Domestic Product) reached 177.3 billion dollars in the same year. Zhengzhou's geographical location (34°16′–34°58′ N; 112°42′–114°14′ E) in the continental monsoon climate allows 60% of its 524.1 mm annual average rainfall to occur during the summer months from June to September, when there is an increased risk of urban flood. For example, heavy rains on August 19, 2018 and August 1, 2019 caused widespread flooding in city; some waterlogging prone points have serious water accumulation, compromising the regular traffic operation (Figs. 2, 3).

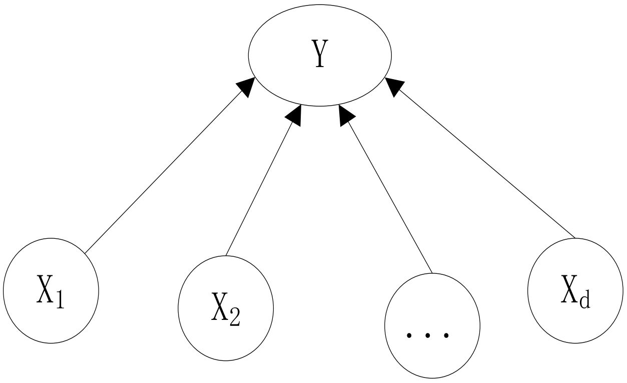


Fig. 2 The structure of the Naïve Bayes

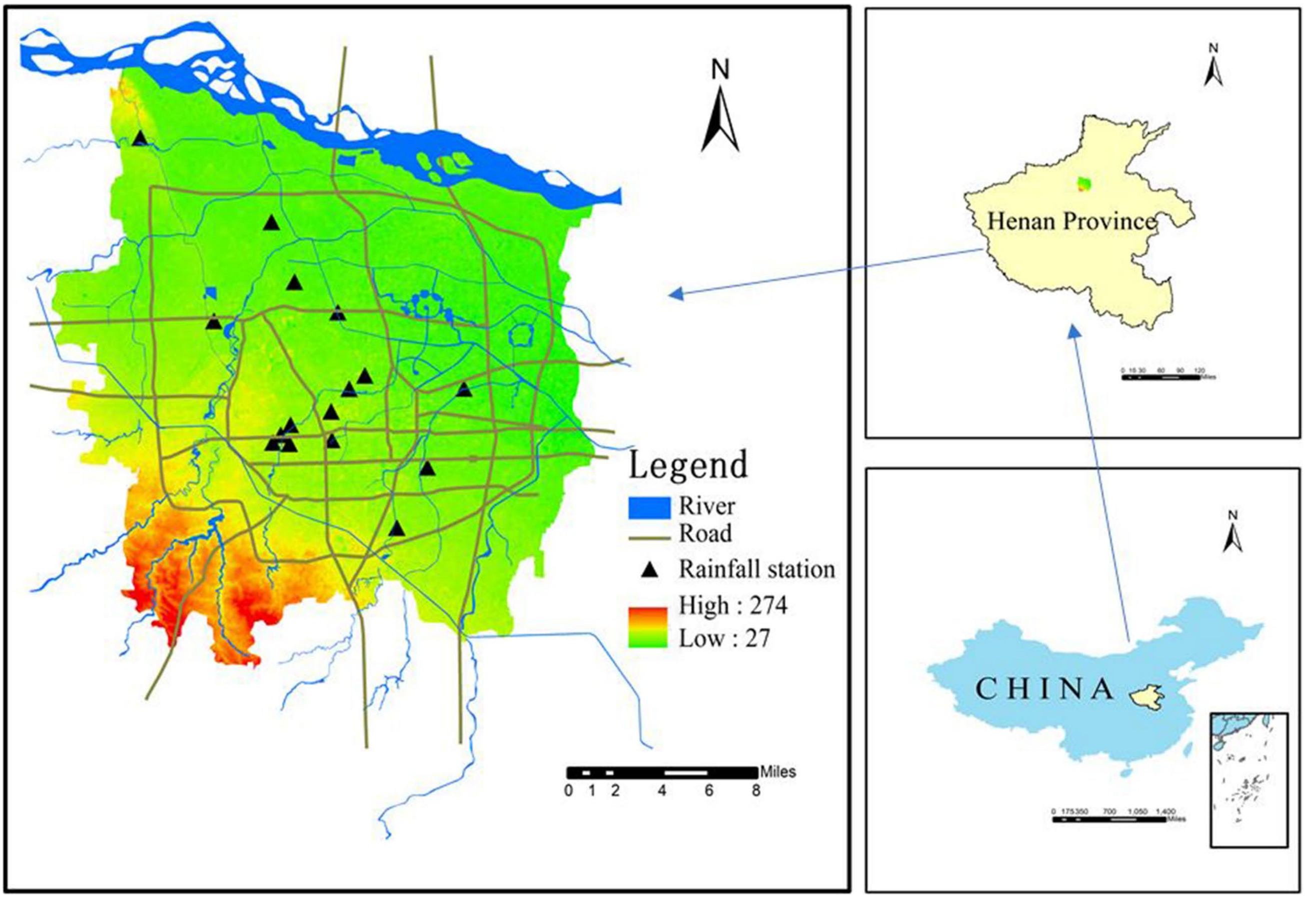


Fig. 1 Location of the study area

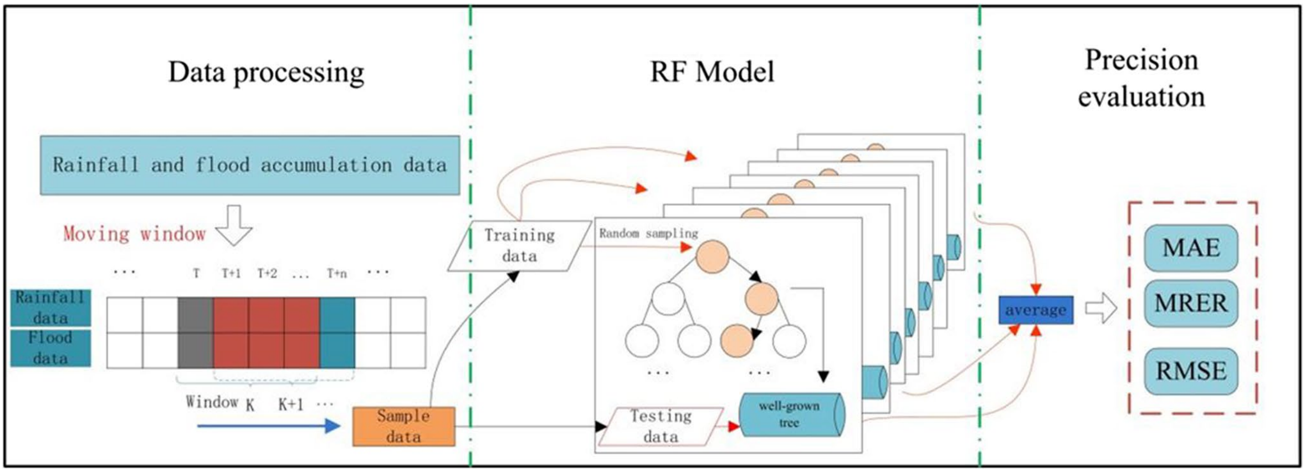


Fig. 3 Model construction of urban flood water accumulation process prediction based on RF regression algorithm

Date and material

Through the analysis of the hydrological process of the accumulation and confluence of water, it is found that both rainfall and geographical factors play an indispensable role in the formation of water. Ignoring any one of these factors may lead to the distortion and deviation of the predicted results. Thus, considering the usefulness of machine learning algorithm for multidimensional data and in combination with previous research results (Choubin et al. 2019; Vafakhah et al. 2020) and the data available in the study area, three main prediction features, namely, geographical characteristics, rainfall characteristics and flood characteristic are selected for training and verification the model. Geographical characteristics describe land use (the proportion of roads, woodlands, grasslands and building) and geographical structure (permeability, catchment area, and slope), which were obtained from the maps extracted of Pleiades Satellite in May 2014 with the 0.5 m high spatial resolution. Rainfall characteristics include three rainfall indexes, namely, rainfall, rainfall duration and peak rainfall, which were obtained from the Henan Meteorological Service. Because the occurrence of rainfall, the characteristics values of rainfall in different parts of Zhengzhou urban area are various, the data of the rainfall was processed by using the Kriging method of space interpolation to refine the rainfall data and increase the diversity of rainfall intensity. For flood characteristic, locations and depths information of flooded urban areas were included, which were collected from the monitoring equipment at each intersection administered by the Zhengzhou Municipal Urban Management Bureau.

Naive Bayes (NB) algorithm

Naive Bayes classifier is one of the few classification algorithms based on probability theory of the classical machine learning algorithms (Perez et al. 2009). It does not need to consume a lot of time for calculation like k-nearest neighbor, support vector machine and other methods, nor does it need to determine and input any parameters (Patil and Atique 2020). Therefore, the time of training model and model test is relatively fast, which is an outstanding advantage to provide sufficient time for urban flood control work to deal with the damage caused by urban flood. And it outperformed five other classifiers, including decision tree, logistic regression, k-nearest neighbor, support vector machine with polynomial kernel, and support vector machine with radial basis function (Lou et al. 2014). NB classifier predicts the probability of a class membership, that is to say the probability that a given set of variables (features) belongs to a particular class (Omran and El Houby 2020). The NB classifier works as shown in the following steps:

The NB classifier predicts the Y_i of classes that X belongs to, based on the highest posteriori probability of the class conditioned on X, which means that:

$$P\left(Y_i|X\right) > P\left(Y_j|X\right) \quad \text{for} \quad 1 \le j \le m, j \ne i \tag{1}$$

Based on Bayes' theorem, P (Y|X) can be written as formula:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{2}$$

For a given sample, the P (X) is independent of the class tag and same for all classes, so P (Y |X) is only related to P (Y) and P (X |Y). Based on the assumption that each network characteristic attribute independently has an attribute influence on the prediction results, he formula can be rewritten as:

$$h_{nb}(X) = \arg\max P(Y) \prod_{i=1}^{d} P\left(x_i | Y\right) \tag{3}$$

Then, the training set was used to set the value for P(xi|Yi). Finally, the model takes the category with the highest probability as the optimal output result.

Random Forest (RF) algorithm

Random Forest algorithm is an ensemble machine learning algorithm for performing classification or regression (Prajwala 2015; Kabir et al. 2018), which was first introduced by Breiman (Breiman 2001) and has been widely used in Geography (Gislason et al. 2006; Guo et al. 2020), Bioecology (Parkhurst et al. 2005; Smith et al. 2010), Medicine (Chen and Liu 2006; Lee et al. 2010) and so on recent years. RF is the algorithm of tree class structure, which combines multiple decision trees to generate corresponding prediction results for different characteristics of the same phenomenon. Compared with various current machine learning models, the RF algorithm has the following three obvious advantages (Malekipirbazari and Aksakalli 2015; Li et al. 2020): 1. RF can deal with high latitude independent variable problems. 2. able to fit and predict nonlinear problems. 3. the learning process is fast, and I can deal with a large amount of data efficiently. The important steps to implement the RF regression algorithm are presented below:

1. K data sets are extracted in the way of Bootstrap sampling with random from the input data sets. The data amount of the K data sets is the same as the original data amount and the composition of the data can be repeated. This step is the first "Random" in the RF model.
2. Assuming that the number of variables in a data set is M, the M_{try} variables are randomly selected from each node of each regression tree as alternative branching variables, and then the optimal branching is selected according to the branching excellence criterion. This step is the second "Random" in the RF model.
3. K_{tree} decision trees are constructed and trained by using the select data from the step 1 and 2. Each decision tree grows as much as possible without pruning, and then K decision trees are formed to form a random forest. This step is the "Forest" in the RF model.
4. The result of the prediction for a new sample is obtained by averaging the predictions from all the individual well-grown regression trees in the RF regression model:

$$f = \frac{1}{K_{tree}} \sum_{i=1}^{K_{tree}} f_i\,(x) \tag{4}$$

where K_{tree} is the total number of trees and $f_i\,(x)$ is the prediction from each individual well-grown regression tree by using the training data set training.

What can be captured from the above modeling steps is that the diversity of the system in RF model is be improved, which can effectively avoid overfitting and improve the predictive performance of the model (Table 1).

Evaluation of model accuracy

Model evaluation is an important step in the modeling and prediction process, which represents accuracy of the results obtained by the model and the degree of people's trust that can be placed in the model. For the prediction of flood susceptibility in waterlogging points based on NB theory, Precision, Recall, Accuracy and F_1score are used as indicators for evaluation of model performance (Table 2).

For the short real-time prediction of flood process based on RF algorithm, Mean Absolute Error (MAE), Mean Relative Error Ratio (MRER) and Root Mean Square Error (RMSE) are used as indicators for evaluation of model performance, which are calculated by the following formula:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| \left(y_{si} - y_{oi}\right) \right| \tag{5}$$

Table 1 The impact factors and flood data in the study

Category	Indicator	Description
Geographical characteristic	Land use	Contains four parameters such as the proportion of roads, woodlands, grasslands and building
	Geographical structure	Contains three parameters such as permeability, catchment area, and slope
Rainfall characteristic	Rainfall	Denotes the total amount of rainfall for a given rainfall
	Rainfall duration	Represents the duration of a rainfall from the beginning to the end
	Peak rainfall	Denotes the highest point on the rainfall intensity process line
Flood characteristic	Flood level	The recorded value of the change in water level over time

Table 2 Categories of result and evaluation indexes of NB classification model

Indicator	Abbreviation	Description
True positive	TP	A positive sample predicted by the model to be positive
False positive	FP	A negative sample predicted by the model to be positive
True negative	TN	A negative sample predicted by the model to be negative
False negative	FN	A positive sample predicted by the model to be negative
Precision	P	P = TP/(TP + FP)
Recall	R	R = TP/(TP + FN)
Accuracy	A	A = (TP + TN)/(TP + FP + TN + FN)
F1score	F_1	$F_1 = 2PR/(P+R)$

*Positive sample: water accumulation occurred in the area; Negative sample: no water accumulation occurred in the area

$$MRER = \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\left|y_{si}-y_{oi}\right|}{y_{oi}}\times 100\% \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_{si}-y_{oi}\right)^{2}} \tag{7}$$

where y_{si} and y_{oi} is the simulated value and the measured value of the flood at the point i, respectively.

The closer the index (P, R, A and F_1) value is to 1, the more accurate the NB model is in predicting the waterlogging point. And the smaller the value of these three indicators (MAE, MRER and RMSE) is, the more the prediction flood depth result of the RF model is in line with the actual situation.

Results and discussions

Predictive analysis of flood susceptibility of urban waterlogging points based on Naive Bayes classification model

Previous studies and transport project appraisal (Dalziell and Nicholson 2001; Chang et al. 2010; Pregnolato et al. 2017) have shown that when the depth of the flood is 3-5 cm, urban vehicles can pass normally without being affected, so the threshold for determining the flood is 5 cm in this study. When the maximum depth of flood in a waterlogging area is greater than the threshold value 5 cm, it is considered that flood will occur in this area, namely the positive sample above; if not, it is considered that there is no need to worry about the occurrence of floods. There are 10 historical rainfalls and corresponding floods depth data available, which happened specifically on July 26th, 2011; August 2nd, 2012; May 26th, 2013; June 9th and 19th, 2014; July 22th, 2015; June 11th,July 19th and August 5th, 2016; July 20th, 2017. SQL Server Data Tools was used to process diversified Data of geographical characteristics, rainfall characteristics and flood characteristic and build database. The geographical feature information of Zhengzhou city and the information of the first 7 rainfall floods were used as training data set to train the model, and the remaining 3 rainfall (August 2nd, 2012; June 19th, 2014 and July 19th, 2016) information was used to verify the model (Table 3).

Short real-time prediction of water accumulation process based on random forest regression algorithm

A waterlogging point in the city was randomly selected after obtaining the flood susceptibility analysis and waterlogging point results by using NB model, and data of 6 rainfall-water events occurred before about the waterlogging point was collected (Fig. 4). By the data preprocessing of linear interpolation, rainfall data and water accumulation data are unified into the same time scale, and the time granularity is 1 min.

The time series of rainfall and water accumulation at this waterlogging prone point are divided and the data set is constructed by using the moving window method, which is a common method for constructing datasets (Wang et al. 2005; Jing et al. 2020). A moving-window of 2 × w grids rolls through the rainfall-flood data grids with size of

Table 3 The prediction result by the NB model (one of three validation events)

The label number of the sample points	Predicted results	Actual situation	Category
1	N	N	TN
2	N	N	TN
3	P	P	TP
4	N	P	FN
/	/	/	/
3322	N	N	TN
3323	P	P	TP
3324	N	N	TN

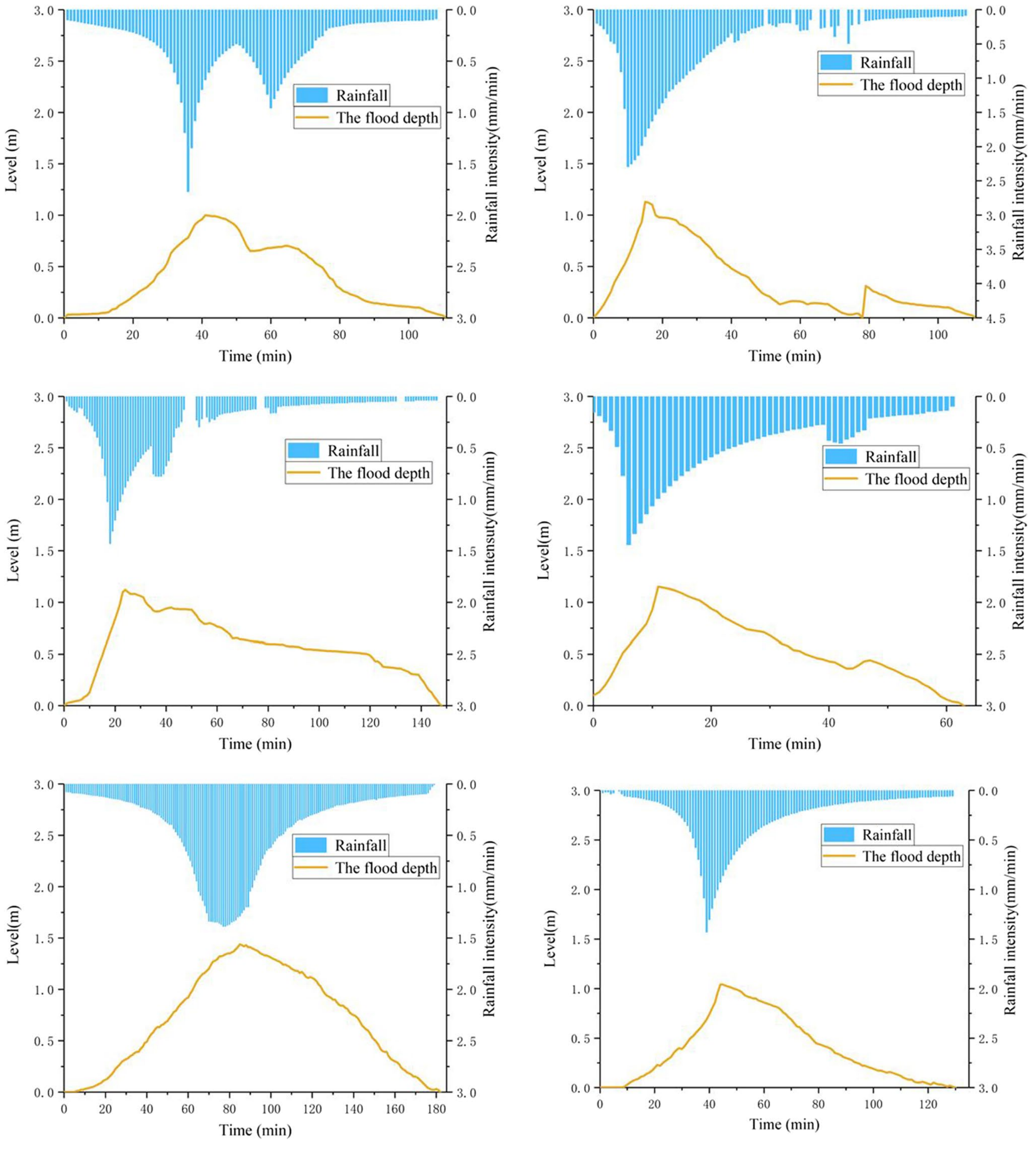


Fig. 4 Time series diagram of rainfall intensity and flood depth

2×J(rows×columns) at a step of 1. The number of the input variables (w of the moving-window) is the most important task in RF model development. For determining the value of w, samples of 12 different combinations of input data were arranged as provided in Table 4. Figure 5 shows the results of training the RF regression model by input different models, which shows that after A9, the model's OOBS (out_of_bag score) increases by less. Thus, considering both the accuracy of the model and the complexity of the model input, the width of the moving window is set as 9, that is,

Table 4 Model structure with a different input combination

Model	Input combination
A1	$X_i = f(y_{i-1}, q_{i-1})$
A2	$X_i = f(y_{i-1}, y_{i-2}, q_{i-1}, q_{i-2})$
A3	$X_i = f(y_{i-1}, y_{i-2}, y_{i-3}, q_{i-1}, q_{i-2}, q_{i-3})$
/	/
A10	$X_i = f(y_{i-1}, y_{i-2}, \ldots, y_{i-8}, q_{i-1}, q_{i-2}, \ldots, q_{i-10})$
A11	$X_i = f(y_{i-1}, y_{i-2}, \ldots, y_{i-9}, qi\text{-}1, q_{i-2}, \ldots, q_{i-11})$
A12	$X_i = f(y_{i-1}, y_{i-2}, \ldots, y_{i-10}, q_{i-1}, q_{i-2}, \ldots, q_{i-12})$

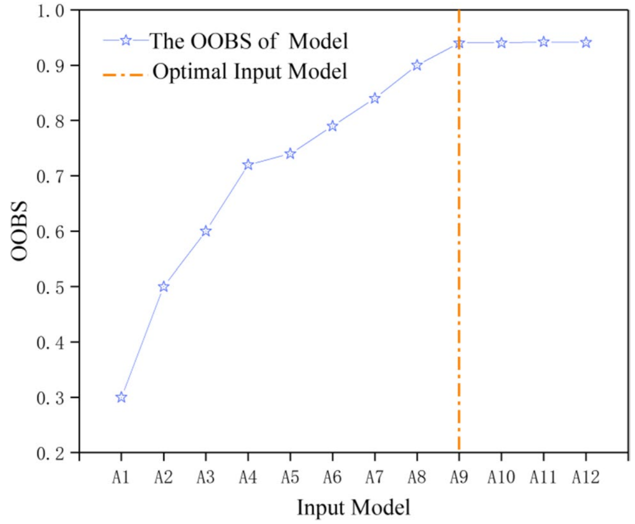


Fig. 5 The OOBS of RF regression model with 12 different input combination models

Table 5 Parameters optimization results of the RF model

Number of the tree (K_{tree})	Maximum depth	Sampling ratio (M_{try}/M)
20	10	0.7

each data set contains respectively 9 rainfall and water accumulation data recorded successively. And the predicted time step is set as 5 min here.

Table 6 Simulation result of the last rainfall event by RF model

Time	Simulated depth of flooded/m	Measured depth of flooded/m	Absolute error/m
0	0	0	0
1	0	0.01	− 0.01
2	0.012	0.03	− 0.018
/	/	/	/
60	0.741	0.74	0.001
61	0.712	0.71	0.002
62	0.71	0.69	0.02
/	/	/	/
120	0.002	0.02	− 0.018
121	0.035	0.01	0.025
122	0	0	0

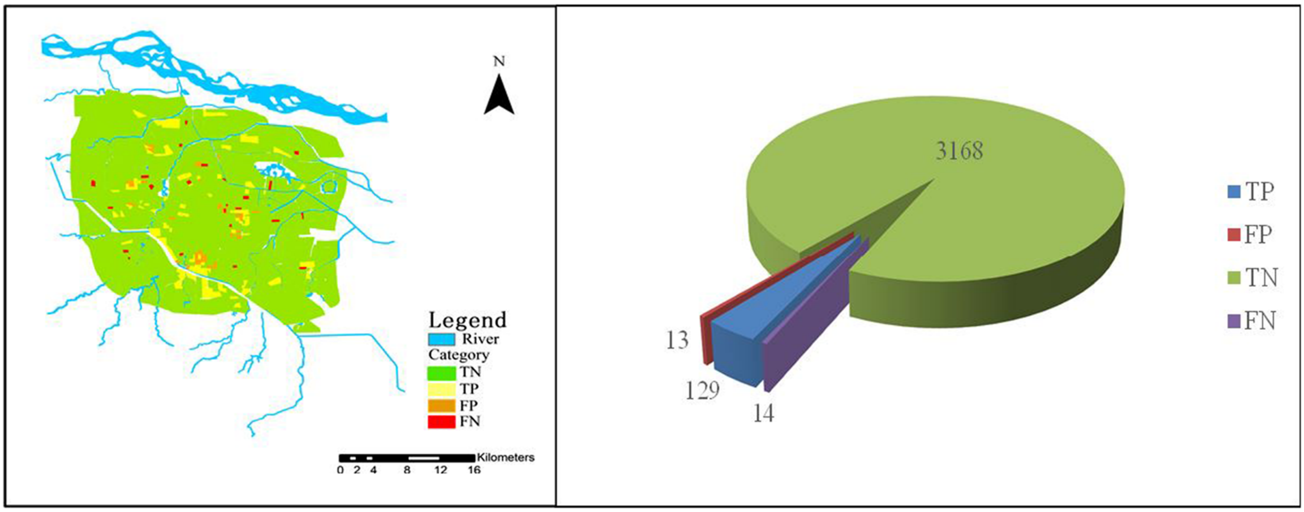
Table 7 Performances of the NB classification prediction model

Indicator	P (%)	R (%)	A (%)	F_1 (%)
Number	91	90.5	98.9	90.7

The RF model contains several built-in parameters, but there are three main parameters affecting the accuracy of the model, respectively: the number of trees, number of features considered at each split and maximum depth of each decision tree (Liu et al. 2020). Those three built-in parameters of the RF model are obtained by means of traversal search and tenfold cross validation. Those three built-in parameters of the RF model are selected and optimized, and the best parameter combination is obtained by means of the traversal search algorithm and tenfold cross validation (Table 5). The parameter Sampling ratio represents the proportion of predicted features of each selected sample. Each sample contains 18 prediction features, and the proportion value of 0.7 means that a sample 12 predicted features are selected. At the beginning and end of the record, rainfall and water values were replenished to 0 for input to the model. The collected data of the first five rainfall accumulation were used as training data set for training and learning of the model, and the last rainfall data was used to verify the prediction performance of the model (Table 6).

Evaluating the performance of the model

The accuracy of NB classifier was evaluated by using the difference between flood susceptibility and predictive classification of urban waterlogging points under real rainfall events (Table 7). In order to make the predicted results more intuitive, the prediction results of flood susceptibility of waterlogging prone points combined with geographic location information were introduced into GIS, and compared with the actual flood's location and results, the actual distribution diagram of the indicators of waterlogging prone points was obtained (Fig. 6). According to the indexes obtained from the results, the precision, recall, accuracy and F_1score all reached more than 90%, indicating that the analysis and prediction of flood susceptibility at urban waterlogging points are reliable. In case of rainfall, NB model can predict the area where urban flooding is likely to occur in Zhengzhou city. Provide reliable information support for city flood control workers. It can be seen from the figure that the waterlogging situation in



(a) (b)

Fig. 6 Distribution and number of waterlogging points of the rainfall event on June 19th, 2014

Table 8 Performances of the RF regression prediction model

Indicator	MAE (%)	MRER (%)	RMSE (%)
Number	0.95	9.53	1.21

Zhengzhou is not too serious compared with that in southern cities, and the area of flood waterlogging is relatively concentrated in the southwest of the city, which may be caused by the early construction of the drainage system in this area and the long-term failure of maintenance and repair.

The accuracy of the RF regression prediction model was assessed using the values of the MAE, MRER and RMSE between the simulated and measured value (Table 8). As shown in Table 8, the MAE, MRER and RMSE of the prediction results of water accumulation depth are 0.95%, 9.53% and 1.21% respectively, which indicates that the water depth predicted by RF model is close to the measured value and the RF prediction model is feasible in the prediction of water accumulation processes. In order to compare the difference between the predicted water level and the actual water level over time more intuitively, the regression curve of water level was fitted (Fig. 7).

It can be seen from the figure that the variation trend of the predicted water depth of the RF model is synchronized with the variation trend of the measured water depth. Combined with the data values of the three indexes (MAE, MRER and RMSE), there are sufficient reasons to prove the applicability of RF model in predicting the process of water accumulation.

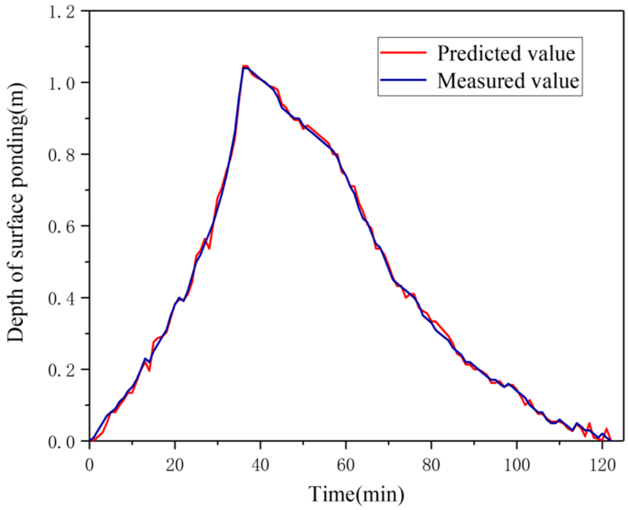


Fig. 7 Fitting curves between predicted and measured values

Conclusion

In this study, in order to achieve the goal of predicting the whole process of urban waterlogging, Naive Bayes and random forest algorithm were used to forecast the waterlogging point and the waterlogging process at the waterlogging point respectively. Four classification evaluation indexes (P, R, A and F_1) and three regression evaluation indexes (MAE, MRER and RMSE) were used to evaluate the prediction performance of the NB classification model and RF regression model.

The results show that NB modal predicted waterlogging point with good performance. Four classification evaluation indexes (P, R, A and F_1) are 91%, 90.5%, 98.9% and 90.7% respectively. These findings demonstrate the validity of the model for the predicting the water accumulation points, when rainfall specific information is available. Therefore, under the background of relatively accurate rainfall forecast information, NB classification algorithm can be used to predict waterlogging points, so as to give urban flood control workers more sufficient time to respond to urban waterlogging. The input data set of RF model is constructed by using sliding window. By comparing the OOBS obtained from 12 different input models, the optimal input model of RF model was determined as A9. The first 5 rainfalls data were used for the training of the model, and the last rainfall was simulated and predicted, and the three regression indexes (MAE, MRER and RMSE) were respectively 0.95%, 9.53% and 1.21%, which demonstrates the validity of the RF regression model for the predicting the water accumulation process of the water accumulation point.

From the results, NB model and RF model can be used to predict the flood and waterlogging information under urban rainfall, which provide effective technical support for urban flood control and forecasting and allow the city's flood control work to have enough time and accurate flood information to prevent and make decisions on the damage caused by the flood in advance.

Acknowledgements The authors thank the anonymous reviewers for their valuable comments. They declare that there is no conflict of interest regarding the publication of this paper.

Funding The study was funded by the Key Project of National Natural Science Foundation of China (No: 51739009), Natural Science Foundation of China (51879242), Science and Technology Innovation Talents Project of Henan Education Department of China (21HASTIT011), Young backbone Teachers Training Fund of Henan Education Department of China (2020GGJS005), and Excellent Youth Fund of Henan Province of China(212300410088).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bhan SK, Team F (2001) Study of floods in West Bengal during september, 2000 using indian remote sensing satellite data. J Indian Soc Remote 29:1–2. https://doi.org/10.1007/bf02989907

Bisht DS, Chatterjee C, Kalakoti S, Upadhyay P, Sahoo M, Panda A (2016) Modeling urban floods and drainage using SWMM and MIKE URBAN: a case study. Nat Hazards 84(2):749–776. https://doi.org/10.1007/s11069-016-2455-1

Braud I, Breil P, Thollet F, Laagouy M, Branger F, Jacqueminet C, Kermadi S, Michel K (2013) Evidence of the impact of urbanization on the hydrological regime of a medium-sized periurban catchment in France. J Hydrol 485:5–23. https://doi.org/10.1016/j.jhydrol.2012.04.049

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Bubeck P, Aerts JCJH, de Moel H, Kreibich H (2016) Preface: flood-risk analysis and integrated management. Nat Hazard Earth Syst Sci 16(4):1005–1010. https://doi.org/10.5194/nhess-16-1005-2016

Chang H, Lafrenz M, Jung IW, Figliozzi M, Platman D, Pederson C (2010) Potential impacts of climate change on flood-induced travel disruptions: a case study of Portland, Oregon, USA. Annal Assoc Am Geogr 100(4):938–952. https://doi.org/10.1080/00045608.2010.497110

Chen XW, Liu M (2006) Prediction of protein-protein interactions using random decision forest framework. Bioinformatics 21(24):4394–4400. https://doi.org/10.1093/bioinformatics/bti721

Cheng T, Xu ZX, Hong SY, Song SL (2017) Flood risk zoning by using 2D hydrodynamic modeling: a case study in Jinan City. Math Probl Eng 2017:5659197. https://doi.org/10.1155/2017/5659197

Choubin B, Moradi E, Golshan M, Adamowski J, Sajedi-Hosseini F, Mosavi A (2019) An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci Total Environ 651:2087–2096. https://doi.org/10.1016/j.scitotenv.2018.10.064

Dalziell E, Nicholson A (2001) Risk and impact of natural hazards on a road network. J Transp Eng 127(2):159–166. https://doi.org/10.1061/(ASCE)0733-947X(2001)127:2(159)

Diaz-Nieto J, Lerner DN, Saul AJ, Blanksby J (2012) GIS water-balance approach to support surface water flood-risk management. J Hydrol Eng 17(1):55–67. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000416

Ding YK, Zhu YL, Feng J, Zhang PC, Cheng ZR (2020) Interpretable spatio-temporal attention LSTM model for flood forecasting. Neurocomputing 403:348–359. https://doi.org/10.1016/j.neucom.2020.04.110

EM-DAT 2020. Disaster profiles. https://www.emdat.be/emdat_db/. Accessed 9 March 2020

Ferreira CSS, Walsh RPD, Shakesby RA, Keizer JJ, Soares D, Gonz á lez-Pelayo O, Coelho COA, Ferreira AJD (2016) Differences in overland flow, hydrophobicity and soil moisture dynamics between Mediterranean woodland types in a peri-urban catchment in Portugal. J Hydrol 533:473–485. https://doi.org/10.1016/j.jhydrol.2015.12.040

Gain AK, Hoque MM (2013) Flood risk assessment and its application in the eastern part of Dhaka City, Bangladesh. J Flood Risk Manag 6(3):219–228. https://doi.org/10.1111/jfr3.12003

GebreEgziabher M, Demissie Y (2020) Modeling urban flood inundation and recession impacted by manholes. Water 12(4):1160. https://doi.org/10.3390/w12041160

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27(4):294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Granata F, Gargano R, Marinis G (2016) Support vector regression for rainfall-runoff modeling in urban drainage: a comparison with the EPA's storm water management model. Water 8(3):69. https://doi.org/10.3390/w8030069

Guo XJ, Zhang CC, Luo WR, Yang J, Yang M (2020) Urban impervious surface extraction based on multi-features and random forest. IEEE Access 8:226609–226623. https://doi.org/10.1109/ACCESS.2020.3046261

Hajat S, Ebi KL, Kovats S, Menne B, Edwards S, Haines A (2003) The human health consequences of flooding in Europe and the implications for public health. Appl Environ Sci Public Health 1(1):13–21

Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H, Kanae S (2013) Global flood risk under climate change. Nat Clim Chang 3(9):816–821. https://doi.org/10.1038/nclimate1911

Huong HTL, Pathirana A (2013) Urbanization and climate change impacts on future urban flooding in Can Tho city, Vietnam. Hydrol Earth Syst Sci 17(1):379–394. https://doi.org/10.5194/hess-17-379-2013

IPCC (2014). In: Core Writing Team, Pachauri RK, Meyer LA (eds) Climate Change 2014: synthesis report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland, p 151

Jamshed A, Birkmann J, McMillan JM, Rana IA, Feldmeyer D, Sauter H (2021) How do rural-urban linkages change after an extreme flood event? Empirical evidence from rural communities in Pakistan. Sci Total Environ 705:141462. https://doi.org/10.1016/j.scitotenv.2020.141462

Jhong BC, Wang JH, Lin GF (2016) Improving the long lead-time inundation forecasts using effective typhoon characteristics. Water Resour Manag 30(12):4247–4271. https://doi.org/10.1007/s11269-016-1418-3

Jing WL, Zhang PY, Zhao XD, Yang YP, Jiang H, Xu JH, Yang J, Li Y (2020) Extending GRACE terrestrial water storage anomalies by combining the random forest regression and a spatially moving window structure. J Hydrol 590:125239. https://doi.org/10.1016/j.jhydrol.2020.125239

Jongman B (2018) Effective adaptation to rising flood risk COMMENT. Nat Commun 9:1986. https://doi.org/10.1038/s41467-018-04396-1

Kabir E, Guikema S, Kane B (2018) Statistical modeling of tree failures during storms. Reliab Eng Syst Safe 177:68–79. https://doi.org/10.1016/j.ress.2018.04.026

Kim HI, Han KY (2020) Data-driven approach for the rapid simulation of urban flood prediction. KSCE J Civ Eng 24(6):1932–1943. https://doi.org/10.1007/s12205-020-1304-7

Kong FH, Ban YL, Yin HW, James P (2017) Modeling stormwater management at the city district level in response to changes in land use and low impact development. Environ Modell Softw 95:132–142. https://doi.org/10.1016/j.envsoft.2017.06.021

LA Lee S, Kouzania AZ, Hu EJ (2010) Random forest based lung nodule classification aided by clustering. Comput Med Imag Grap 34(7):535–542. https://doi.org/10.1016/j.compmedimag.2010.03.006

Li W, Niu L, Chen H, Wu H (2020) Robust downscaling method of land surface temperature by using random forest algorithm. J Geo-Inf Sci 22(8):1666–1678. https://doi.org/10.12082/dqxxkx.2020.190142

Liu J, Sun SQ, Tan ZL, Liu Y (2020) Nondestructive detection of sunset yellow in cream based on near-infrared spectroscopy and interval random forest. Spectrochim Acta A 242:118718. https://doi.org/10.1016/j.saa.2020.118718

Loperfideo JV, Noe GB, Jarnagic ST, Hogan DM (2014) Effects of distributed and centralised stormwater best management practices and land cover on urban stream hydrology at the catchment scale. J Hydrol 519:2584–2595. https://doi.org/10.1016/j.jhydrol.2014.07.007

Lou WC, Wang XQ, Chen F, Chen YX, Jiang B, Zhang H (2014) Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and gaussian Naive bayes. PLoS ONE 9(1):e86703. https://doi.org/10.1371/journal.pone.0086703

Malekipirbazari M, Aksakalli V (2015) Risk assessment in social lending via random forests. Expert Syst Appl 42(10):4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Miller JD, Kim H, Kjeldsen TR, Packman J, Grebby S, Dearden R (2014) Assessing the impact of urbanization on storm runoff in a peri-urban catchment using historical change in impervious cover. J Hydrol 515:59–70. https://doi.org/10.1016/j.jhydrol.2014.04.011

Mu DR, Luo PP, Lyu J, Zhou MM, Huo AD, Duan WL, Nover D, He B, Zhao XL (2020) Impact of temporal rainfall patterns on flash floods in Hue City, Vietnam. J Flood Risk Manag 14(1):e12668. https://doi.org/10.1111/jfr3.12668

Nourani V, Alami MT, Aminfar MH (2009) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Eng Appl Artif Intell 22(3):466–472. https://doi.org/10.1016/j.engappai.2008.09.003

Omran S, El Houby EMF (2020) Prediction of electrical power disturbances using machine learning techniques. J Ambient Intel Hum Comp 11(7):2987–3003. https://doi.org/10.1007/s12652-019-01440-w

Parkhurst DF, Brenner KP, Dufour AP, Wymer LJ (2005) Indicator bacteria at five swimming beaches—analysis using random forests. Water Res 39(7):1354–1360. https://doi.org/10.1016/j.watres.2005.01.001

Patil HP, Atique M (2020) CDNB: CAVIAR-dragonfly optimization with Naive bayes for the sentiment and affect analysis in social media. Big Data 8(2):107–124. https://doi.org/10.1089/big.2019.0130

Patro S, Chatterjee C, Mohanty S, Singh R (2009) Flood inundation modeling using MIKE FLOOD and remote sensing data. J Indian Soc Remote 37(1):107–118. https://doi.org/10.1007/s12524-009-0002-1

Perez A, Larranaga P, Inza I (2009) Bayesian classifiers based on kernel density estimation: flexible classifiers. Int J Approx Reason 50(2):341–362. https://doi.org/10.1016/j.ijar.2008.08.008

Prajwala TR (2015) A comparative study on decision tree and random forest using R tool. IJARCCE 4(1):4

Pregnolato M, Ford A, Glenis V, Wilkinson S, Dawson R (2017) Impact of climate change on disruption to urban transport networks from pluvial flooding. J Infrastruct Syst 23(4):04017015. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000372

Rahmati O, Darabi H, Panahi M, Kalantari Z, Naghibi SA, Ferreira CSS, Kornegady A, Karimidastenaei Z, Mohamadi F, Stefanidis S, Bui DT, Haghighi AT (2020) Development of novel hybridized models for urban flood susceptibility mapping. Sci Rep 10(1):12937. https://doi.org/10.1038/s41598-020-69703-7

Schmitt TG, Thomas M, Ettrich N (2004) Analysis and modeling of flooding in urban drainage systems. J Hydrol Eng 299(3–4):300–311. https://doi.org/10.1016/j.jhydrol.2004.08.012

She L, You XY (2019) A dynamic flow forecast model for urban drainage using the coupled artificial neural network. Water Resour Manag 33(9):3143–3153. https://doi.org/10.1007/s11269-019-02294-9

Smith A, Sterba-Boatwright B, Mott J (2010) Novel application of a statistical technique, random forests, in a bacterial source tracking study. Water Res 44(14):4067–4076. https://doi.org/10.1016/j.watres.2010.05.019

Vafakhah M, Loor SMH, Pourghasemi H, Katebikord A (2020) Comparing performance of random forest and adaptive neuro-fuzzy inference system data mining models for flood susceptibility mapping. Arab J Geosci 13(11):417. https://doi.org/10.1007/s12517-020-05363-1

Vojinovic Z, Tutulic D (2009) On the use of 1D and coupled 1D–2D modelling approaches for assessment of flood damage in urban areas. Urban Water J 6(3):183–199. https://doi.org/10.1080/15730620802566877

Wang X, Kruger U, Irwin GW (2005) Process monitoring approach using fast moving window PCA. Ind Eng Chem Res 44(15):5691–5702. https://doi.org/10.1021/ie048873f

Ward R (1978) Floods: a geographical perspective. MacMillan, London

Wu XS, Wang ZL, Guo SL, Liao WL, Zeng ZY, Chen XH (2017) Scenario-based projections of future urban inundation within a coupled hydrodynamic model framework: a case study in Dongguan City, China. J Hydrol 547:428–442. https://doi.org/10.1016/j.jhydrol.2017.02.020

Wu ZN, Zhou YH, Wang HL (2020a) Real-time prediction of the water accumulation process of urban stormy accumulation points based on deep learning. IEEE Access 8:151938–151951. https://doi.org/10.1109/ACCESS.2020.3017277

Wu ZN, Zhou YH, Wang HL, Jiang ZH (2020b) Depth prediction of urban flood under different rainfall return periods based on deep learning and data warehouse. Sci Total Environ 716:137077. https://doi.org/10.1016/j.scitotenv.2020.137077

White MD, Greer KA (2006) The effects of watershed urbanization on the stream hydrology and riparian vegetation of Los Penasquitos Creek, California. Landsc Urban Plan 74(2):125–138. https://doi.org/10.1016/j.landurbplan.2004.11.015

Yazdi J, Neyshabouri SAAS (2012) A simulation-based optimization model for flood management on a watershed scale. Water Resour Manag 26(15):4569–4586. https://doi.org/10.1007/s11269-012-0167-1

Yazdi MN, Ketabchy M, Sample DJ, Scott D, Liao HH (2019) An evaluation of HSPF and SWMM for simulating streamflow regimes in an urban watershed. Environ Modell Softw 118:211–225. https://doi.org/10.1016/j.envsoft.2019.05.008

Zhang S, Li Z (2019) Simulation of urban rainstorm waterlogging and pipeline network drainage process based on swmm. J Phys 1213:052061. https://doi.org/10.1088/1742-6596/1213/5/052061

Zhao DQ, Chen JN, Wang HZ, Tong QY, Cao SB, Sheng Z (2009) GIS-based urban rainfall-runoff modeling using an automatic catchment-discretization approach: a case study in Macau. Environ Earth Sci 59(2):465–472. https://doi.org/10.1007/s12665-009-0045-1

Zolch T, Henze L, Keilholz P, Pauleit S (2017) Regulating urban surface runoff through nature-based solutions—an assessment at the micro-scale. Environ Res 157:135–144. https://doi.org/10.1016/j.envres.2017.05.023

Zoppou C (2001) Review of urban storm water models. Environ Modell Softw 16(3):195–231. https://doi.org/10.1016/S1364-8152(00)00084-0

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.