Natural Hazards (2021) 108:2335–2356
https://doi.org/10.1007/s11069-021-04782-x

ORIGINAL PAPER

Rapid forecasting of urban flood inundation using multiple machine learning models

Jingming Hou[1] · Nie Zhou[1] · Guangzhao Chen[1] · Miansong Huang[2] · Guangbi Bai[3]

Received: 3 January 2021 / Accepted: 3 May 2021 / Published online: 18 May 2021

Abstract
Urban flood inundation is worsening as the number of short-duration rainstorms increases, and it is difficult to accurately predict urban flood inundation over a long lead time; however, the traditional hydrodynamic-based urban flood models still have difficulty realizing real-time prediction. This study establishes a rapid forecasting model of urban flood inundation based on machine learning (ML) algorithms and a hydrodynamic-based urban flood model. The ML model is obtained by training the simulation results of the hydrodynamic model and rainfall characteristic parameters. Part of Fengxi New Town, China, was used to validate the forecasting model. A comparison of ML predictions and hydrodynamic model simulations shows that when using one ML algorithm (random forest (RF) or K-nearest neighbor (KNN)) for inundation prediction, the accuracy of the inundation water volume and area is insufficient, with a maximum error of 28.56%. Combining the RF and KNN models can effectively improve the prediction accuracy and overall stability, the mean relative errors of the inundation area and depth are less than 5%, and the mean relative errors of the inundation volume can control within 10%. The simulated time of a single rainfall event can be controlled within 20 s, which can provide sufficient lead time for emergency decision-making, thereby helping decision-makers to take more appropriate measures against inundation.

Keywords Urban inundation · Rapid forecasting · Machine learning · Random forest model · K-nearest neighbor model

1 Introduction

Flood disasters, one of the most frequent natural disasters, have a great impact on agriculture, transportation, and people's lives and property (Smith et al. 2014; Guha-Sapir et al. 2016). As the urban heat island effect worsens due to the acceleration of

✉ Nie Zhou
2190420070@stu.xaut.edu.cn

[1] State Key Laboratory Base of Eco-Hydraulic Engineering in Arid Area, Xi'an University of Technology, Xi'an 710048, Shaanxi, China

[2] Beijing Capital Co.Ltd., Bei'jing 100044, China

[3] Shaanxi Meteorological Service Center, Xi'an 710014, Shaanxi, China

urbanization and the increasing incidence of extreme weather events resulting from climate change, urban flood inundation disasters show an increasing tendency and cannot be effectively predicted in a short period of time (Xie et al. 2017; Wu et al. 2020).

Accurate forecasting models can effectively simulate the situation of urban flood inundation, which has great significance both for guiding timely warnings to alleviate the loss of lives and properties and assisting urban construction by setting up optimization plans (Xie et al. 2017). Therefore, it is necessary to establish urban flood inundation forecasting models. However, urban flood inundation simulation is a complicated process, and the influence of various factors, such as rainfall, soil moisture, river conditions, and landform features, needs to be considered. In addition, iteratively solving complex equations multiple times will take considerable time, so in general, such models cannot provide an effective decision basis for decision-makers when facing emergency cases.

Several studies have been devoted to developing simulation models for providing rapid prediction. such as using discrete Boltzmann equation (DBE) to bypass the complexity of the usual shallow water models (Rocca et al. 2020), accelerating parallel computing through graphics processing unit (Hu et al. 2018; Liang et al. 2016), optimizing the iterative format to reduce the computation time (Chew et al. 2020; Hou et al. 2015), or using unstructured grids to reduce the number of computing grids (Wu et al. 2018). However, these traditional studies mostly focused on algorithm optimization, though have certain benefits but cannot achieve breakthroughs because regardless of how the model is optimized, complex equations are still required due to the complexity of the physical mechanism.

In the last two decades, artificial intelligence algorithms have been effectively developed, of which machine learning (ML) is widely used in most aspects of social production. ML methods can summarize the rules between input parameters and output results with lower computational cost. The solution structure is simpler and more efficient than the physical model (Mekanik et al. 2013). Several methods are also being increasingly combined with ML: artificial neural networks (ANNs) for multi-step-ahead flood inundation forecasting (Chang et al. 2018), physical hybrid neural network models to forecast typhoon floods (Jhong et al. 2018), monthly runoff forecasting based on hybrid long short-term memory neural network and ant lion optimizer (LSTM-ALO) model (Yuan et al. 2018), support vector regression (SVR), multivariate adaptive regression spline (MARS) and M5 model tree (M5Tree) for river flow data forecasting in semiarid mountainous catchments (Yin et al. 2018), hybrid extreme learning machine-particle swarm optimization algorithms for flood forecasting (Anupam and Pani 2020), support vector machine (SVM)-based ML methods to predict water levels in rainwater pipe networks (Wang and Song 2020), LSTM network for the probabilistic daily streamflow forecasting (Zhu et al. 2020), hybrid decision tree-based ML models for short-term water quality prediction (Lu and Ma 2020), etc. The corresponding studies show that ML has great prospects in flood prediction.

However, these studies almost focus on large-scale river basins and carry out large-scale low-precision simulations. ML methods require a large amount of training data for learning. Since urban flood inundation is usually caused by short-duration rainstorms, it is difficult to effectively obtain enough historical inundation data, so ML technology is rarely applied to urban flood inundation research. With the continuous development of hydrodynamic models in recent years, the precise simulation of urban flooding process can be realized based on high-precision terrain and rainfall information, which provides the possibility for the application of machine learning technology in urban flooding prediction as explained in the following.

The purpose of this study is to realize the real-time prediction of urban flood inundation to meet the reference needs of emergency decision-making. We propose a rapid forecasting model combining a hydrodynamic model with ML algorithms, the hydrodynamic model calibrated by measured data is used to simulate sufficient urban flood inundation data under various rainfall conditions, and the inundation data are used as training data to generate a rapid forecasting model using machine learning algorithms. The established model can generate the corresponding inundation map within 20 s, thus helping decision-makers to take necessary measures to reduce losses.

The organizational structure of this paper is as follows. In Sect. 2, we first introduce the main workflow, as well as the hydrodynamic model and ML algorithms we used. In Sect. 3, we present the study area and rainfall data. We establish the rapid prediction model and analyze the forecasting ability in Sect. 4 and the conclusion is summarized in Sect. 5.

2 Methodology of rapid forecasting model for urban flood inundation

Based on the hydrodynamic urban flood model and ML algorithm, a rapid prediction model of urban flood inundation is established in this study. The coupling model mainly includes the following processes: (1) obtain the digital elevation model (DEM) of the study area and collect enough rainfall data of each type to enable it to represent the various conditions of urban heavy rain; (2) use the hydrodynamic model to obtain the inundation area and volume data caused by rainstorms; (3) extract the characteristic parameters of rainfall and reduce the number of unnecessary parameters through correlation analysis to improve the model performance and reduce the amount of time wasted in model training; (4) use ML algorithms to fit the input parameters and output data; and (5) save the trained model and use the test sets to verify the reliability of the model (Fig. 1).

2.1 Hydrodynamic-based urban flood model

The hydrodynamic-based urban flood model takes the 2D SWE (shallow water equation) as the governing equation, its conservation scheme can be expressed in vector form as Eqs. (1–3):

$$\frac{\partial q}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} = S \tag{1}$$

$$\mathbf{q} = \begin{bmatrix} h \\ uh \\ vh \end{bmatrix}, \mathbf{F} = \begin{bmatrix} uh \\ u^2h + gh^2/2 \\ uvh \end{bmatrix}, \mathbf{G} = \begin{bmatrix} vh \\ vuh \\ v^2h + gh^2/2 \end{bmatrix} \tag{2}$$

$$\mathbf{S} = S_{\mathrm{f}} + \mathrm{S}_{\mathrm{b}} = \begin{bmatrix} i \\ -gh\partial z_b/\partial x \\ -gh\partial z_b/\partial y \end{bmatrix} + \begin{bmatrix} 0 \\ -C_f u\sqrt{u^2 + v^2} \\ -C_f v\sqrt{u^2 + v^2} \end{bmatrix} \tag{3}$$

where t represents the time; x and y are the Cartesian coordinates; q denotes the vector of conserved flow variables consisting of h and uh and vh, which are the water depth and unit-width discharges in the x and y-directions; F and G are the flux vectors in the x- and

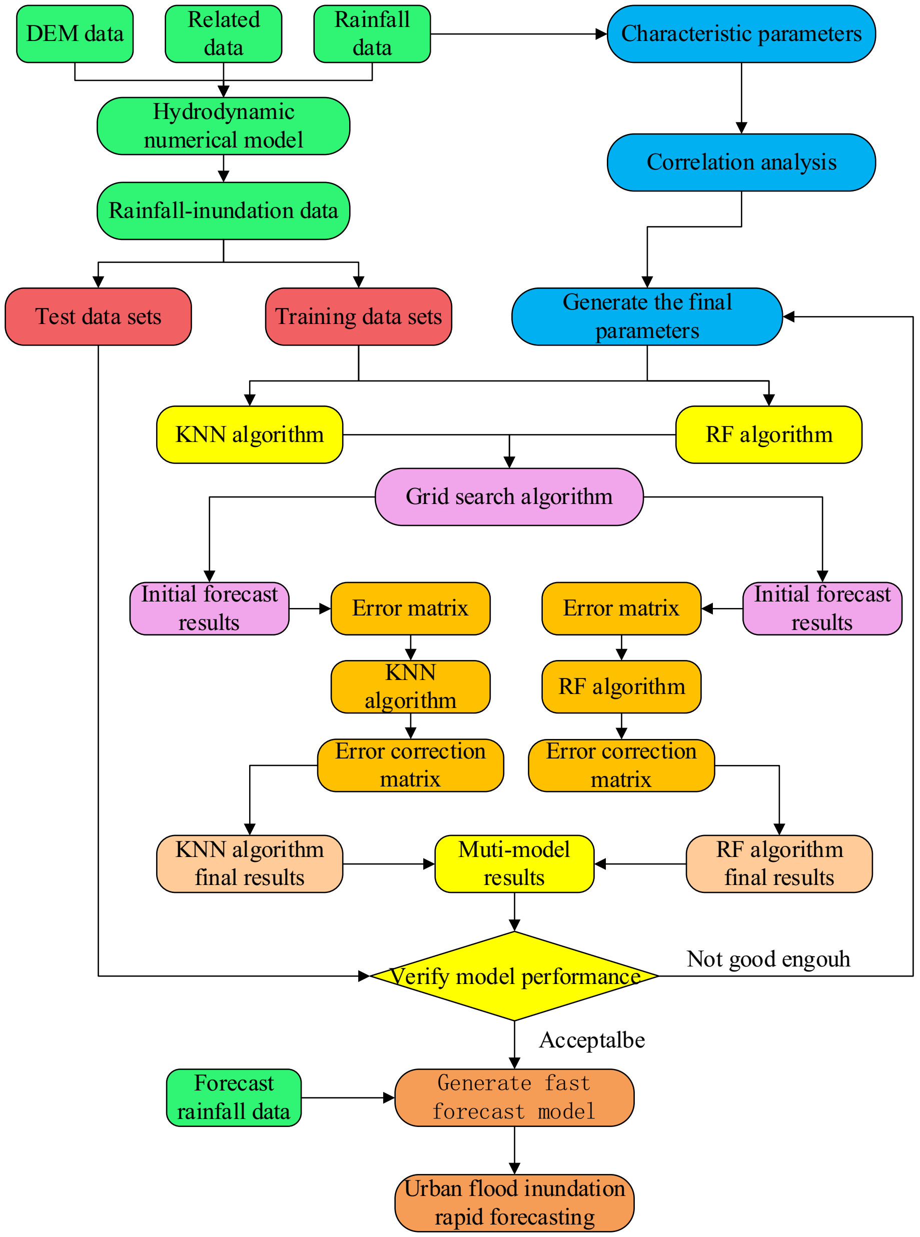


Fig. 1 Framework for the multiple machine learning model

y-directions, respectively; *g* is gravity; *S* is the source vector that may be further subdivided into net rain source terms *i*, slope source terms S_b and friction source terms S_f; *Z* represents the bed elevation; C_f depends on the Manning coefficient and can be expressed as $C_f = gn^2/h^{1/3}$, where *n* is the Manning coefficient.

The model adopts the dynamic wave method, which comprehensively considers the combined effects of inertia, pressure, gravity, and friction terms on water flow, and can most effectively simulate the evolution of water flow under complex boundary conditions. It discretely solves the 2D SWE by the Godunov's finite volume method, which focuses on constructing discrete equations from a physical point of view. Each discrete equation is an expression of the conservation of a certain physical quantity on a finite volume, which can ensure that the discrete equation has conservation characteristics. Through the second-order explicit Runge–Kutta (Hubbard 1999) method, we constructed a monotonic upwind scheme for conservation laws (MUSCL) with second-order space–time accuracy to ensure the conservation of mass and effectively solve the discontinuity problem (Hou et al. 2015). To solve complex problems such as abrupt flow and discontinuity, the model uses the HLLC (Haren-Lax-van Leer contact) approximate Riemann solver to calculate the mass and momentum flux on the unit interface. The static water reconstruction method (Sivakumar et al. 2009) is used to address the problem of negative water depth at the boundary between wet and dry cells, and the flow rate is used to replace the single width flow rate as the calculation variable to effectively convert the second-order formula prone to instability into a stable first-order formula when the water depth is lower than a certain value or the flow rate is higher than a certain value. On the premise of ensuring the calculation accuracy, the slope surface source term in the calculation cell is converted to the flux on the boundary of the cell to ensure it also meets the full stability condition in complex terrain calculations. The friction source term uses the implicit splitting point method optimized by Liang and Marche (Hou et al. 2013) to maintain the stability of the calculation results. At the same time, GPU parallel technology is used to accelerate the simulation process and ensure the model's computational efficiency.

2.2 Machine learning models for urban flood inundation

ML, the core of artificial intelligence, includes many kinds of algorithms and has been widely used in many fields. Based on the random forest (RF) and K-nearest neighbor (KNN) algorithms, this study aims to investigate and build a rapid forecasting model for urban flood inundation.

2.2.1 Random forest model

The decision tree (Quinlan 1986, 2014; Breiman 1984) algorithm is a nonparametric supervised learning method that can summarize decision rules from a series of unordered and irregular features, and present these rules as a tree graph to solve classification and regression problems. It can effectively process a large amount of data, but the single decision mechanism of decision tree will be greatly affected by the characteristic parameters. As it is easy to overfit based on the training sets, the classifier will achieve a perfect performance on the training sets, but show poor performance on the testing sets, so a single decision tree will have difficulty obtaining an acceptable result (Breiman 2001).

The RF algorithm was proposed by Leo Breiman in 2001 by combining bagging integrated learning theory with the random subspace method (Breiman 2001; Kwok et al. 1990; Ho 1998). And RF uses the decision tree as the base classifier, as shown in Fig. 2. It contains multiple decision trees trained by the bagging algorithm and combines the results of multiple decision trees. When new samples need to be predicted, the simulation results of multiple evaluators will be considered to obtain the comprehensive

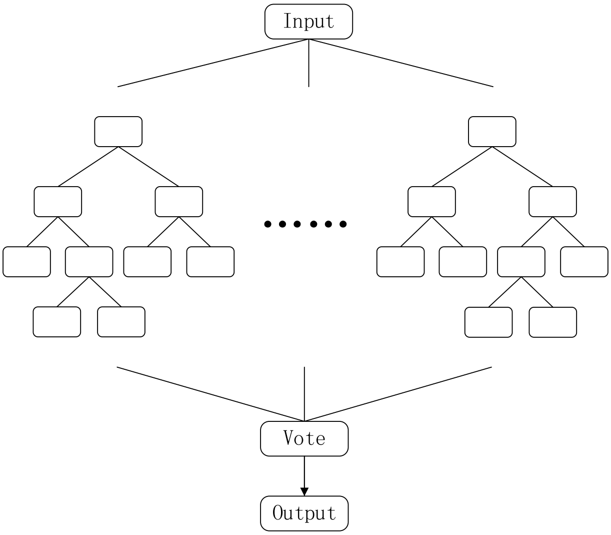


Fig. 2 Schematic diagram of the principle of RF algorithm

results to overcome the instability of a single model and achieve a better regression and classification performance.

The improvement in RF algorithm is mainly manifested in the randomness of the training sets and characteristic parameters. First, the training data are generated by sampling with replacement, and N sub-datasets are constructed randomly by the bagging method. In each sub-dataset, the elements are allowed to repeat, as shown in Fig. 3. Second, like the sub-dataset, characteristic parameters of different decision trees are also randomly chosen from the proposed features, then the optimal feature is selected as a root node to generate decision trees according to the impurity. The third step is establishing a voting mechanism. When predicting the results of a new sample, each decision tree will give its own results, and the final output results will be determined by voting on these trees. This method can effectively prevent the overfitting of the decision tree and make the model generalizable and achieve a better performance on the new data.

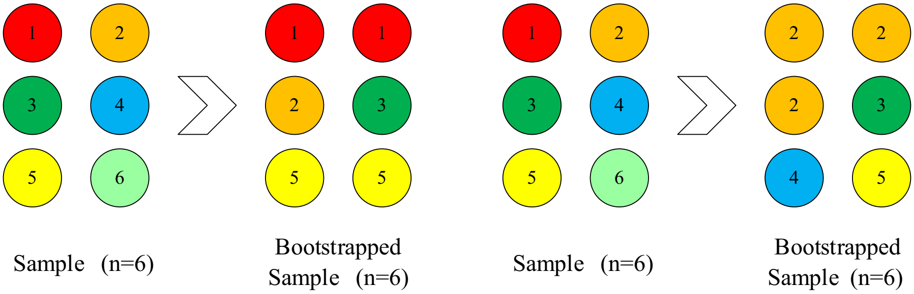


Fig. 3 Schematic diagram of sample selection

2.2.2 K-nearest neighbor model

The KNN model was proposed by Cover and Hart in 1968 based on the vector space model (Keller et al. 1985). In the KNN algorithm, each sample is regarded as a point or vector in R^n space. The basic idea of the KNN regression algorithm is to use the neighborhood algorithm to find K samples that are closest to the target sample in the training sets, and use these *K* samples for estimation. This algorithm has the advantages of maturity, simplicity, and good robustness to noise in training sets and has been widely applied in many fields (Vialetto and Noro 2019; Huang et al. 2017; Liu et al. 2016). Its main steps are as follows:

(1) The existing samples are instantiated, that is, converted into the form (x, f(x)), where x is the characteristic parameter of the sample, x is represented by ($x_1, x_2,\ldots, x_n$), and x_n is the nth property of the sample; that is, the number of feature parameters is equal to the dimension composed of vectors. After instantiation, all the samples constitute the training sets and test sets.

(2) Given a new test sample x_i, a distance formula is used to calculate the distance between x_i and each original sample in the training sets, and K samples closest to x_i are screened out. The distance formulas mainly include the Euclidean distance and Manhattan distance. In this paper, the Euclidean distance formula with good performance in both the training sets and the test sets is selected through the comprehensive comparison of the fitting effect. The formula is shown in Eq. (4).

$$L\left(x_{\mathrm{i}}, x_{\mathrm{j}}\right) = \left(\sum_{l=1}^{n}\left(x_{\mathrm{i}}^{(l)} - x_{\mathrm{j}}^{(l)}\right)^{2}\right)^{\frac{1}{2}} \tag{4}$$

where x_{i} and x_{j} are two samples and $x_{\mathrm{i}}^{(l)}$ and $x_{\mathrm{j}}^{(l)}$ are the l eigenvalues of x_{i} and x_{j}, respectively. L(x_{i}, x_{j}) is the Euclidean distance between x_{i} and x_{j}.

(3) According to the proximity between the selected K samples and the unknown sample, the predicted results of the K samples are assigned to the new test samples as the forecast value according to their weight (Fig. 4).

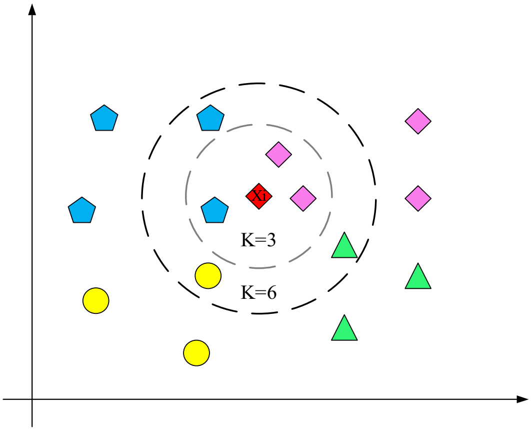


Fig. 4 KNN model schematic diagram

2.3 Parameter correlation analysis

Redundant parameters not only cannot improve the predicted accuracy, but also may cause more errors (Hu et al. 2020) while increasing the complexity of the model. Therefore, a correlation analysis between the characteristic parameters of rainfall and the inundation situation is carried out. Due to the complex terrain, it is unrealistic to independently verify the data in each grid. The rationality of the parameter selection was verified by calculating the correlation parameters between the rainfall characteristic parameters and the accumulated inundation area and volume of the region.

This study mainly uses the Pearson correlation coefficient (Sedgwick 2012) for the correlation analysis, and the correlation calculation formula is shown in Eqs. (5–7). The correlation criterion is shown in Table 1.

$$\rho_{\mathrm{xy}} = \frac{\mathrm{cov}(x,y)}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \tag{5}$$

$$\mathrm{cov}(x,y) = E(xy) - E(x)E(y) \tag{6}$$

$$\sigma_{\mathrm{x}} = \sqrt{E\left(x^2\right) - E^2(x)}, \sigma_{\mathrm{y}} = \sqrt{E\left(y^2\right) - E^2(y)} \tag{7}$$

where ρ_{xy} is the Pearson correlation coefficient and E(x), E(y) and E(xy) are the mathematical expectation of x, y and xy, respectively; cov (x, y) is the covariance between x and y; σ_{x} and σ_{y} are the variances of x and y.

2.4 Parameters optimization of forecasting model

The fitting degree of the ML algorithm to the training data is mainly determined by the algorithm and its parameters, such as the number of decision trees, maximum depth, maximum number of features in the random forest algorithm and the neighbors K, distance formula, and algorithm selection in KNN will directly affect the reliability of the model. In order to improve the performance of the ML algorithm, based on the Scikit-learn framework (Varoquaux et al. 2015), the grid search algorithm is used to optimize the parameters of the learning algorithm. Grid search is an exhaustive algorithm that can automatically simulate the combination of different parameters and compare errors through cross-validation methods to determine the most suitable model parameters for the current training data to improve the accuracy of the model (Liu et al. 2014).

Table 1 Value ranges of the Pearson correlation coefficient and their corresponding correlations

The absolute value of the Pearson correlation coefficient	Correlation
0–0.2	Very weak or no correlation
0.2–0.4	Weak correlation
0.4–0.6	Moderate intensity correlation
0.6–0.8	Strong correlation
0.8–1	Very strong correlation

In this study, the cross-validation coefficient is finally selected sixfold cross-validation, and the optimal parameters of the algorithm are finally determined by the grid search algorithm as shown in Table 2.

2.5 Error correction matrix

In order to reduce the accumulated error caused by the hydrodynamic and the machine learning model, the study generated an error matrix through the simulation results of the hydrodynamic model and the prediction results of the machine learning algorithm. Based on ML algorithm, the rainfall characteristic parameters were used as input values and the error matrix as the target values to construct the error modified model, so the prediction result of the ML model is generated by superimposing the initial prediction result and the error correction matrix (Fig. 5).

2.6 Multi-model for urban flood inundation

Due to the characteristics of single machine learning algorithm, no matter how to optimize the parameters, there may still be large errors in individual rainfall forecasts. In order to improve the overall reliability of the rapid forecasting model in urban flood inundation, the research carried out weighted redistribution of the simulation results of RF and KNN algorithm to obtain the comprehensive results of multiple models. The formula is shown in Eq. (8).

$$R = \frac{1}{\mathrm{S}_{\mathrm{KNN}} + \mathrm{S}_{\mathrm{RF}}} \left(R_{\mathrm{KNN}} \cdot S_{\mathrm{KNN}} + R_{\mathrm{RF}} \cdot S_{\mathrm{RF}} \right) \tag{8}$$

R is the final result of the multi-model, S_{RF} and S_{KNN} represent the R^2 values of the RF and KNN models, respectively, and R_{RF} and R_{KNN} represents the predicted values of the RF and KNN models in the grid.

In this study, R^2, mean absolute error (MAE), root mean square error (RMSE), and mean relative error (MRE) are mainly used to verify the reliability of the model. The calculation method is shown in Eq. (9–12):

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \overline{y}_i\right)^2} \tag{9}$$

$$\mathrm{MAE} = \frac{1}{m} \sum_{i}^{m} \left| \left(y_i - \hat{y}_i\right) \right| \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m} \sum_{1}^{m} \left(y_i - \hat{y}_i\right)^2} \tag{11}$$

$$\mathrm{MRE} = \frac{1}{m} \sum_{i=1}^{m} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{12}$$

Table 2 Selection of main parameters of machine learning algorithm

Model	RF model			KNN model		
Parameters	Number of decision trees	Maximum depth	Maximum number of features	Number of neighbors K	Distance formula	Algorithm
Value	40	6	6	5	Europe	Ball_tree

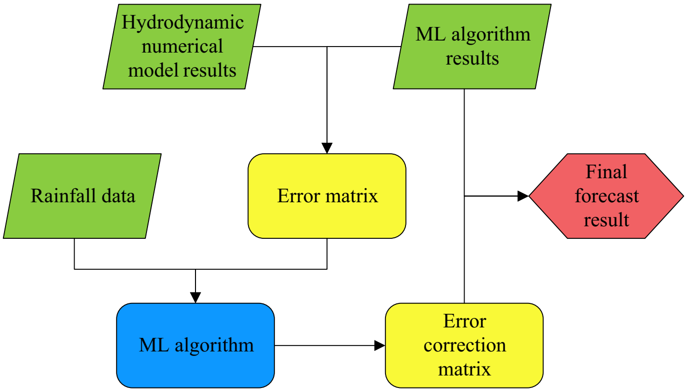


Fig. 5 Schematic diagram of error correction matrix construction

where y_i is the true value on the test set, $\overline{y}_i$ is the average of the true values on the test set, $\hat{y}_i$ is the predicted value, and m is the number of values.

3 Study area and rainfall data

3.1 Study area

The Xixian New Area, located in Shaanxi Province, China, has a temperate continental climate. The average annual rainfall is approximately 500 mm, but more than 50% of the annual rainfall events are concentrated from July to September, and the rainfall is often heavy rainstorms with short durations. The terrain in this area is complex, and there are many low-lying sections, which are prone to landslides, urban flood inundation, and other meteorological disasters during the rainy season. Therefore, we select a part of the Xixian New Area with an area of 2.432 km^2 as the study area.

Since the width of a road is generally approximately 15 m, a coarse grid will not be able to represent the characteristics of the roads, and over precise terrain can improve the accuracy to some extent, but the calculation time will increase considerably. Therefore, the horizontal resolution of the terrain data used in the study is 2 m, composed of 640 × 950 cells, and the digital elevation map is shown in Fig. 6. Based on the maximum likelihood classification method, the study area was divided into five classes: roads, houses, bare land, grassland, and forest. The area of each type of land is shown in Table 3. According to the Xi’an rainstorm intensity formula adopted in the design scheme, the design return period of the drainage pipe of the study area is once a year, which can cope with the rainstorm intensity of 10.74 mm/h. Therefore, the designed drainage capacity of the pipe network in the study area was equivalent to the infiltration rate of 10.74 mm/h, according to the equivalent drainage method. (Hou et al. 2017; Li et al. 2020). The Manning coefficient and infiltration of different land use types are determined with reference to urban drainage standards and literature (Gao 2014; Li 2017) (Fig. 7).

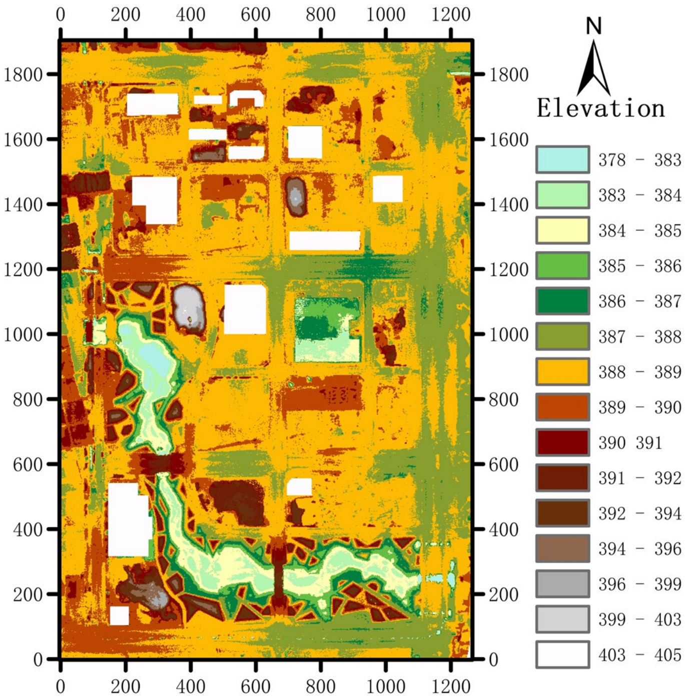


Fig. 6 DEM of the study area

Table 3 Information related to land use

Land used	Area (m^2)	Infiltration (mm/h)	Manning coefficient
Grass	234,689	28.39	0.060
Bare land	29,956	19.43	0.030
Forest	105,442	37.55	0.200
Road	184,942	0	0.014
House	49,893	0	0.015

3.2 Rainfall data

At present, because most of the high-precision public meteorological rainfall forecast data have a temporal resolution of 1 h, the rainfall temporal resolution adopted in this work is 1 h. Because the measured rainfall data in the study area are limited, historical rainfall data cannot cover all possible types of rainstorms. Therefore, in the process of model training and verification, both the design rainfall and measured rainfall comprising 180 fields are added in this study. The historical rainfall data were obtained from the Fengxi sponge city control platform. Bi Xu et al. (2015) pointed out that the rainfall in Xi'an urban area presents mostly short-duration torrential rains, and most of them are single-peak rainfall, its rainfall characteristics are highly similar to the Chicago rainfall pattern. Therefore, the design rainfall data were generated by the Chicago rain pattern generator according to the formula of rainstorms in this area.

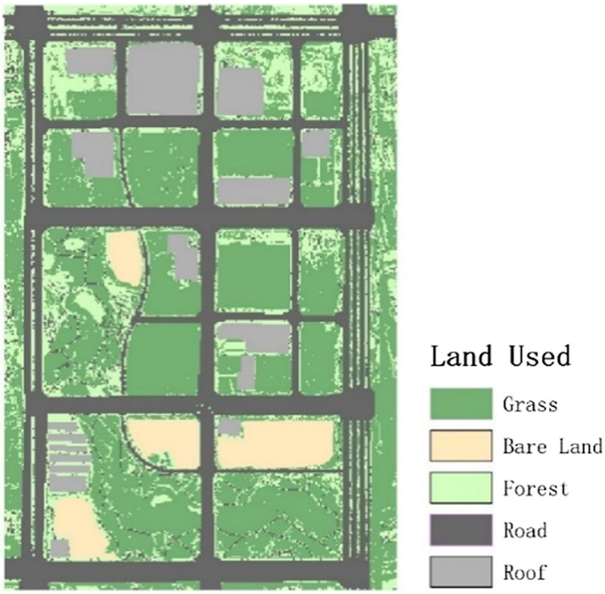


Fig. 7 Land used distribution in the study area

The formula of rainstorms in the study area (Hou et al. 2019) is as follows:

$$i = \frac{446.3676 \times (1 + 1.971\,\lg p)}{(t + 7.4246)^{0.8124}} \tag{13}$$

where i is the rainstorm intensity (mm/h); p is for the rainstorm recurrence period (a); and t is the rainfall duration (minutes).

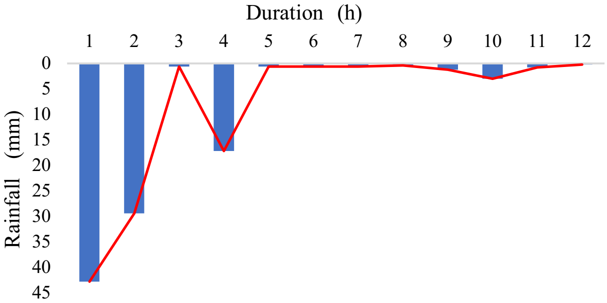
4 Results and discussion

In this section, the reliability of the hydrodynamic model is verified, and since decision-makers tend to be most concerned about the maximum loss caused by rainfall, we mainly constructed the relationship between rainfall and the most serious inundation situation. We chose the R^2, MAE and RMSE to judge the overall stability and accuracy of the constructed ML model. Then, to avoid and overestimation of the reliability of the learning model caused by the areas with few inundation, four regions of A-D in Fig. 9 with severe inundation were selected for further analysis. The learning performance of the model was verified by comparing the relative error of rainstorm-inundation situations between the hydrodynamic model and the ML methods in the main inundation area. Finally, to further improve the forecasting, the results of the two algorithms are merged to obtain the final forecast results.

4.1 Verification of hydrodynamic bases urban flood model

The rainstorm-induced inundation data were generated by the hydrodynamic model simulation. Therefore, the hydrodynamic model simulation performance was first verified in this paper. The rainfall data used for model verification are the measured rainfall data

Fig. 8 Measured rainfall process on August 25, 2016



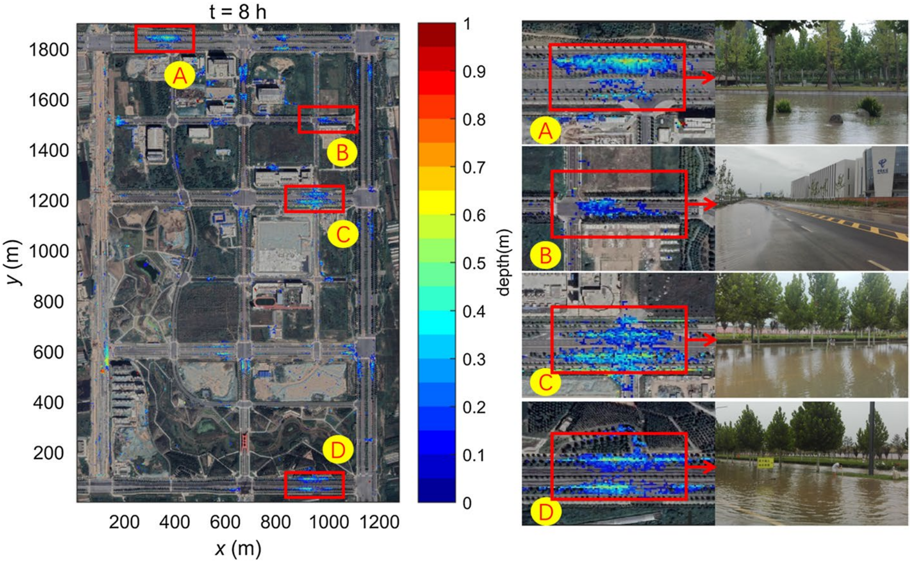


Fig. 9 Simulation results and comparison with measured data

collected by the automatic network weather station in the Xixian New Area on August 25, 2016. The rainfall event lasted from 0:30 to 13:30 on August 25, the cumulative amount of rainfall was 97.40 mm, and the maximum rainfall of 3 h was 72.4 mm. The specific rainfall process is shown in Fig. 8.

The rainstorm caused multiple inundations in the study area. The measured data showed that the depth of inundation on many roads reached 30 cm, and in some areas reached 50 cm, resulting in traffic paralysis on many roads. The simulation results of the hydrodynamic model and the partial enlargement comparison with the actual measurement results are shown in Fig. 9. It can be seen from the figure that the model can accurately reflect the location of water, and the inundation area and depth are basically consistent with the measured data. The results show that the model can accurately simulate the inundation caused by rainstorms. However, with a GeForce RTX 2070 Super graphics card and GPU acceleration technology, the calculation still takes approximately 1 h.

Table 4 Correlation analysis of characteristic parameters

	Person correlation coefficient						
	Cumulative rainfall	Rainfall recurrence period	Rainfall duration	Rainfall peak	Maximum rainfall of 2 h	Maximum rainfall of 3 h	Cumulative rainfall before peak
Accumulated inundation area	0.986	0.888	0.451	0.984	0.971	0.979	0.449
Accumulated inundation volume	0.981	0.848	0.427	0.976	0.971	0.978	0.434

Table 5 Scores of ML model

Score	Random forest		KNN	
	Training data	Testing data	Training data	Testing data
R^2	0.991	0.985	0.986	0.987
MAE(m)	0.002	0.002	0.002	0.002
RMSE(m)	0.008	0.010	0.010	0.010

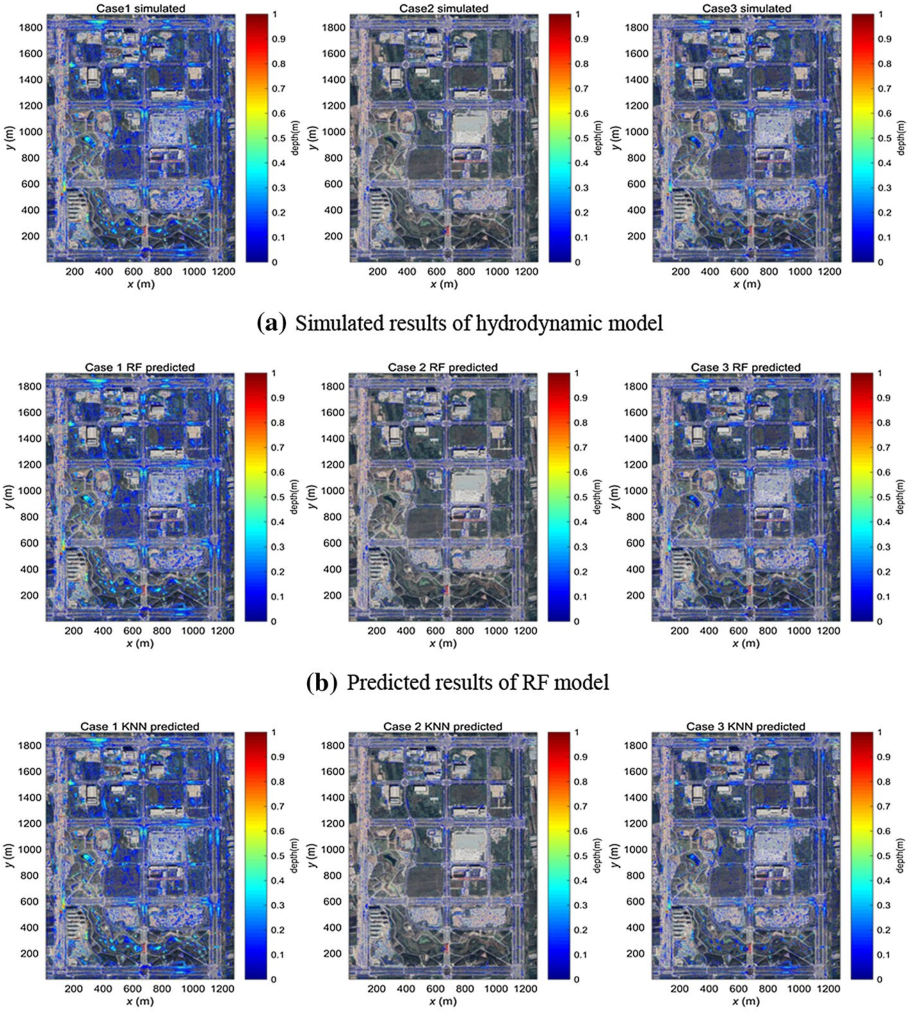
4.2 Effective parameters selected for rapid forecasting model

The results of the correlation analysis of the characteristic parameters are shown in Table 4, in which the Pearson correlation coefficient between the rainfall duration and the accumulated inundation area is 0.451, and the Pearson correlation coefficient between the cumulative rainfall before the peak and the accumulated inundation area is 0.449, within the range of 0.4–0.6, indicating a moderate correlation, and the correlation with the inundation volume is similar. The other parameters are correlated more with the accumulated inundation area and volume, with correlation coefficients exceeding 0.8, showing that accumulated inundation area and volume have a very strong correlation between them. In general, the selected parameters are reasonable.

4.3 Forecasted results for study area

In total, 180 rainfall events were selected for fitting: 150 events for the training sets and 30 events for the test sets. Model training was conducted at the time when the inundation was the most serious. Each rain simulation result was composed of 608,000 grid simulation data points. Through model parameter optimization, the final scores of the RF and KNN models are shown in Table 5.

In the training sets, the RF model reveals a better fitting performance; the R^2 value is 0.991, while in the test sets it is 0.985, which means that the RF model can effectively fit the data, and the overall performance is good. Compared with the RF model, the KNN model also gets an acceptable score, with an R^2 value of 0.986 in the test sets, and with a R^2 score of 0.987 in the test sets, indicating that the KNN model is also suitable for this type of data learning and can reflect the overall situation. We also found that the RMSE and MAE of both the RF and KNN models are very small, which may be because there are some high-lying areas such as buildings in this region. In this type of area, rainwater will quickly be drained to low-lying areas, so during the simulation, the depth and volume of water are very small under all types of rainfall conditions. Therefore, when the rainfall and inundation results are fitted by ML algorithms, the results will achieve a nearly perfect performance, which in turn causes an overestimation of the effectiveness of the overall regional forecast.



(a) Simulated results of hydrodynamic model

(b) Predicted results of RF model

(c) Predicted results of KNN model

Fig. 10 Comparison of inundation results between hydrodynamic and ML model

Figure 10 shows a comparison diagram of the inundation results obtained by the hydrodynamic model and the ML algorithms. Figure 10a shows the simulation result of the hydrodynamic model. Due to the large infiltration capacity of woodland and grassland areas, the water depth and volume in these areas are relatively shallow, and the topography of the housing area is high, rainwater is quickly discharged to the surrounding roads and toward low-lying sections forming a large area of inundation including low-lying sections. Figure 10b, c shows the ML results. Through comparison, it can be seen that the learning results of the two ML algorithms are very similar to the hydrodynamic model simulation results, and the overall difference is very small. The location and depth of the inundation are basically the same, which is consistent with the above conclusion that the overall performance of the model is reliable.

4.4 Forecasted results for inundation spots

In this part, to prevent the overestimation of the model caused by the areas with few inundation, we selected four main inundation spots (A-D in Fig. 9) as validation areas. Because the terrain may have a small number of noisy points in the region, which will seriously affect the maximum water depth, inundation area and volume are chosen to verify the reliability of the model. The water mean depth, obtained by dividing the water volume by the inundation area, was used to approximate auxiliary verification. The specific relative error of the inundation area, inundation volume, and average water depth of the selected inundation regions are shown in Fig. 11.

From Fig. 11, it can be seen that in the 30 test rainfall events, more than 75% of the forecast relative error of the rainfall events is less than 10%, and the MRE can also be controlled within 10%. The MREs of inundation area and depth are 3.93% and 3.18%, respectively. The RF model has the best prediction performance for water depth, and more than 75% of the relative error of rainfall events can be controlled within 5%, and the prediction

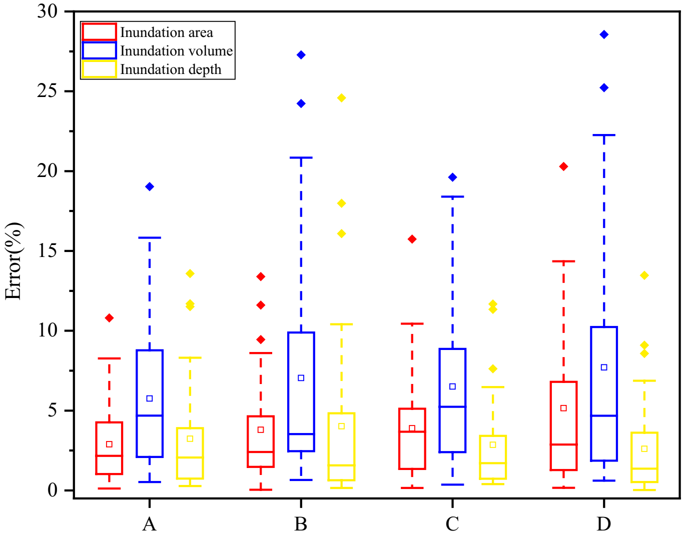


Fig. 11 Prediction errors of RF model

performance of inundation volume is slightly worse, with an MRE of 6.76%. This is because the inundation volume is calculated by the inundation depth and the size of each grid, and there is an error accumulation phenomenon in this process. Although the overall performance of the RF model is good, there may still be a large error in a single rainfall. For example, the maximum error of the inundation volume in the spot D is 28.56%.

The forecasted result of the KNN model is shown in Fig. 12. The KNN model obtains similar scores as the RF model in general, but the performance in selected regions is better, the MRE of inundation volume is 6.21%, and the maximum error is 20.16% appearing in spot D, 8.40% less than that of the RF model. Similar to the RF model, the KNN model performs better in the prediction of the inundation area and average water depth, with relative mean errors of 3.37% and 3.15%, respectively. And the KNN algorithm did not show a large abnormal deviation in the prediction of the inundation area and depth, with maximum errors of 13.40% and 17.94%, respectively, which were more stable than the RF model.

4.5 Forecasted results by applying multi-model

Section 4.4 shows that although the same training sets are used, there are still some differences between the RF model and KNN model in prediction performance, at the same time, both types of models can control the MRE within 10%. Therefore, to further improve forecasting performance, we combine the RF and KNN models, redistributing the predicted results according to their different weights and the results are shown in Fig. 13.

Figure 13 shows that the combination of the two algorithms can further reduce the forecast error to a certain extent. The MRE of inundation area and inundation depth are reduced to 3.27% and 3.16%, respectively, and the MRE of the inundation volume is also reduced to 5.72%. It can be noticed that the maximum errors of the inundation depth and inundation volume at spot B of the multi-model results have both increased compared with the KNN model. This is because KNN model is significantly better than the RF model in

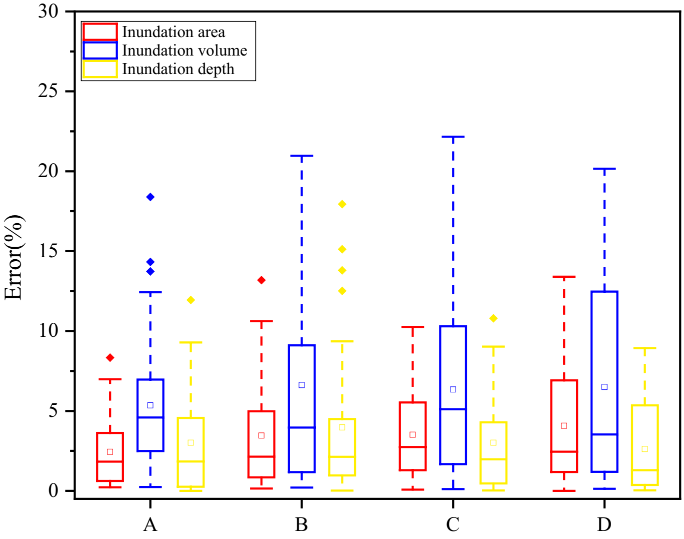


Fig. 12 Prediction errors of KNN model

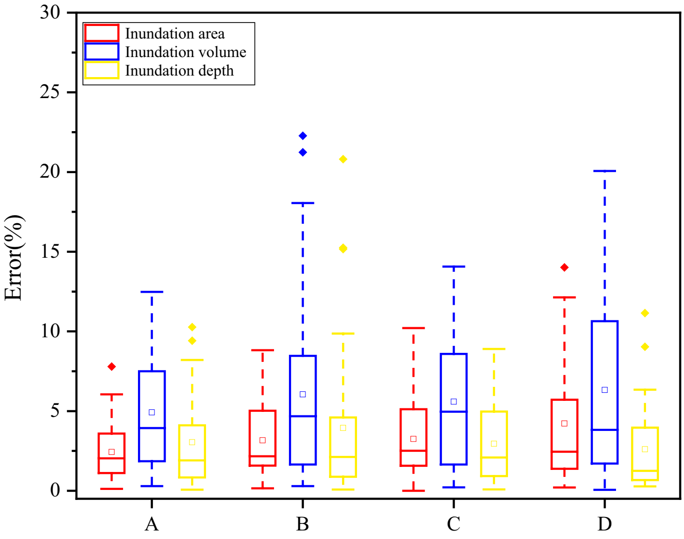


Fig. 13 Prediction errors of multi-model

the maximum error control of spot B, but the maximum errors in spots A, C, and D have been greatly improved. The maximum error of the inundation volume in the spot A has been reduced to 12.48%, the spot C has been reduced to 14.07%, and the inundation volume error in the spot D has also been reduced to 20.06%. From a point of view, it makes the forecasting performance more stable.

From the overall point of view, the multi-model can effectively integrate the prediction results of the two types of models, reduce the abnormal error caused by the uncertainty of a single model, and control the MRE of the inundation area and average depth within 5%. The MRE of inundation volume is also controlled within 10%, and the reliability of the forecast can meet the demand of emergency decision-making on the accuracy of prediction.

4.6 Simulation time

The hardware configuration of this research is based on NIVIDA Geforce GTX 2070 super graphics card, CPU is Core I7-8700, and the comparison of model simulation time is shown in Table 6, the hydrodynamic model needs to be iteratively calculated from time 0, and it takes about 3435.13 s to simulate the inundation evolution process of a single rainfall for 10 h. Based on the RF model, the cumulative simulation time of 30 rainfall data at the most severe time of inundation is 310.25 s, and the average simulation time of a single rainfall is 10.34 s. The KNN algorithm takes a little shorter, and the average rainfall

Table 6 Comparison of simulation time

Model	Hydrodynamic model	KNN model	RF model	Multi-model
Cumulative time (s)	103,053.90	305.90	310.25	329.43
Average time (s)	3435.13	10.20	10.34	10.98

simulation time is 10.20 s. Since multi-model needs to integrate the simulation results of the RF and KNN models, so it takes a bit longer, and the average simulation time of a single rainfall is 10.98 s. The results show that the ML model can provide certain decision support for emergency decision-making and meet the requirements of rapid forecasting of urban flood inundation.

5 Conclusion

This research aims to explore the construction of a rapid forecasting model of urban flood inundation based on high-precision hydrodynamic model and ML algorithms. We obtain the urban flood inundation conditions under various types of local rainfall through a hydrodynamic model. Then, the RF and KNN algorithms are combined to establish the relationship between the rainfall characteristic parameters and the results of inundation, avoiding the iterative calculation of complex equations, and realizing the rapid prediction of urban flood inundation. Taking Fengxi New City, China, as the research area, the simulation effect of the rapid forecasting model is comprehensively verified. The main results are summarized as follows:

To reduce the number of redundant parameters and optimize the learning and forecasting speeds, the Pearson correlation coefficient was used to analyze the correlation of the selected parameters. The results showed that the correlation coefficients of the selected rainfall characteristic parameters, inundation area, and inundation volume were all greater than 0.4, indicating moderate correlation at a minimum, and the selected parameters were reasonable. Only using the KNN or RF model can get a rough inundation situation in a short time, but there may still be large errors in certain rainfall events. For example, the RF model has a maximum water volume error of 28.56% in spot B. In order to further reduce the error, the study constructs multiple models by combining the two algorithms in the form of weight distribution. The error analysis results show that the combination of the two models makes the forecast effect more stable. The MRE of inundation area and depth are less than 5%, and the MRE of inundation volume is 5.72%, which can also be controlled within 10%. It shows that the model can accurately predict the urban flood inundation caused by rainstorm. In terms of efficiency, the built model can output the forecast results within 1 min and generate a distribution map of urban flood inundation, which can provide sufficient lead time for emergency decision-making, helping decision-makers to take more appropriate measures against inundation.

In conclusion, the accuracy and efficiency of the proposed method could meet the requirements of rapid forecasting for urban flood inundation. Future work is planned to reflect the inundation process instead of the maximum inundation time by improving the model.

Acknowledgements This research was supported by the National Key Research Program of China (2016YFC0402704), National Natural Science Foundation of China (52079106; 52009104), and the Shaanxi International Science and Technology Cooperation and Exchange Program (2017KW-014).

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

Anupam S, Pani P (2020) Flood forecasting using a hybrid extreme learning machine-particle swarm optimization algorithm (ELM-PSO) model. Model Earth Syst Environ 6(1):341–347. https://doi.org/10.1007/s40808-019-00682-z

Bi X, Cheng L, Yao DS, Wang BP, Wang L, Jin LN, Yang XC (2015) Analysis on urban rainstorm pattern of Xi'an. J Anhui Agric Sci 43(35):295–297. https://doi.org/10.13989/j.cnki.0517-6611.2015.35.107

Breiman L (1984) Classification and regression trees. Wadsworth Int Group. https://doi.org/10.1002/widm.8

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

Chang LC, Amin MZM, Yang SN, Chang FJ (2018) Building ANN-based regional multi-step-ahead flood inundation forecast models. Water 10(9):1283. https://doi.org/10.3390/w10091283

Chew AWZ, Law WK, Vu TT (2020) Optimizing speedup performance of computational hydrodynamic simulations with UPC programming model. J Comput Civ Eng 34(2):060200011–060200015. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000876

Gao EP (2014) Research on manning coefficient of different vegetated slope. Dissertation, Beijing Forestry University

Guha-Sapir D, Hoyois P, Below R, Vanderveken A (2016) Annual Disaster Statistical Review 2015: The Numbers and Trends. CRED, Université Catholique de Louvain: Brussels, Belgium. https://doi.org/10.13140/RG.2.2.10378.88001

Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844. https://doi.org/10.1109/34.709601

Hou P (2012) Pearson's correlation coefficient. Bmj 345:e4483. https://doi.org/10.1136/bmj.e4483

Hou J, Liang Q, Simons F, Hinkelmann R (2013) A stable 2D unstructured shallow flow model for simulations of wetting and drying over rough terrains. Comput Fluids 82:132–147. https://doi.org/10.1016/j.compfluid.2013.04.015

Hou JM, Liang QH, Zhang HB, Hinkelmann R (2015a) An efficient unstructured MUSCL scheme for solving the 2D shallow water equations. Environ Model Softw 66:131–152. https://doi.org/10.1016/j.envsoft.2014.12.007

Hou JM, Guo KH, Wang Z, Jing HX, Li DL (2017) Numerical simulation of design storm pattern effects on urban flood inundation. Adv Water Sci 28(6):820–828. https://doi.org/10.14042/j.cnki.32.1309.2017.06.003

Hou JM, Li DL, Wang X, Guo KH, Tong Y, Ma Y (2019) Effects of initial conditions of LID measures on runoff control at residential community scale. Adv Water Sci 30(1):45–55. https://doi.org/10.14042/j.cnki.32.1309.2019.01.005

Hu X, Song L (2018) Hydrodynamic modeling of flash flood in mountain watersheds based on high-performance GPU computing. Nat Hazards 91(2):567–586. https://doi.org/10.1007/s11069-017-3141-7

Hu X, Shi L, Lin L, Magliulo V (2020) Improving surface roughness lengths estimation using machine learning algorithms. Agric For Meteorol. https://doi.org/10.1016/j.agrformet.2020.107956

Huang M, Lin R, Huang S, Xing T (2017) A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. Adv Eng Inf 33:89–95. https://doi.org/10.1016/j.aei.2017.05.003

Hubbard ME (1999) Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids. J Comput Phys 155(1):54–74. https://doi.org/10.1006/jcph.1999.6329

Jhong YD, Chen CS, Lin HP, Chen ST (2018) Physical hybrid neural network model to forecast typhoon floods. Water 10(5):632. https://doi.org/10.3390/w10050632

Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 4:580–585. https://doi.org/10.1109/TSMC.1985.6313426

Kwok SW, Carter C (1990) Multiple decision trees. Mach Intell Pattern Recogn 9:327–335. https://doi.org/10.1016/B978-0-444-88650-7.50030-5

Li GY (2017) Comparative study of soil infiltration under different land uses in loess hilly regions. Dissertation, Northwest A&F University

Li DL, Hou JM, Xia JQ, Tong Y, Yang D, Zhang DW, Gao XJ (2020) An efficient method for approximately simulating drainage capability for urban flood. Front Earth Sci. https://doi.org/10.3389/FEART.2020.00159

Liang Q, Xia X, Hou J (2016) Catchment-scale high-resolution flash flood simulation using the GPU-based technology. Proc Eng 154:975–981. https://doi.org/10.1016/j.proeng.2016.07.585

Liu C, Yin SQ, Zhang M, Zeng Y, Liu JY (2014) An improved grid search algorithm for parameters optimization on SVM. Appl Mech Mater 644:2216–2219.

Liu K, Li Z, Yao C, Chen J, Zhang K, Saifullah M (2016) Coupling the k-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting. Int J Sedim Res 31(2):149–158. https://doi.org/10.1016/j.ijsrc.2016.02.002

Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 249:126169. https://doi.org/10.1016/j.chemosphere.2020.126169

Mekanik F, Imteaz MA, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. J Hydrol 503(503):11–21. https://doi.org/10.1016/j.jhydrol.2013.08.035

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. https://doi.org/10.1023/A:1022643204 87

Quinlan JR (2014) C4. 5: Programs for machine learning. Elsevier.

Rocca ML, Miliani S, Prestininzi P (2020) Discrete Boltzmann numerical simulation of simplified urban flooding configurations caused by dam break. Front Earth Sci 8:346

Sivakumar P, Hyams DG, Taylor LK, Briley WR (2009) A primitive-variable riemann method for solution of the shallow water equations with wetting and drying. J Comput Phys 228(19):7452–7472. https://doi.org/10.1016/j.jcp.2009.07.002

Smith MW, Carrivick JL, Hooke J, Kirkby M (2014) Reconstructing flash flood magnitudes using 'structure-from-motion: a rapid assessment tool. J Hydrol 519:1914–1927. https://doi.org/10.1016/j.jhydrol.2014.09.078

Varoquaux G, Buitinck L, Louppe G, Grisel O, Mueller A (2015) Scikit-learn: machine learning without learning the machinery. GetMobile Mobile Comput Commun 19(1):29–33. https://doi.org/10.1145/2786984.2786995

Vialetto G, Noro M (2019) Enhancement of a short-term forecasting method based on clustering and knn: application to an industrial facility powered by a cogenerator. Energies 12(23):4407. https://doi.org/10.3390/en12234407

Wang H, Song L (2020) Water level prediction of rainwater pipe network using an SVM-based machine learning method. Int J Pattern Recognit Artif Intell 34(02):2051002. https://doi.org/10.1142/S0218001420510027

Wu F, Ramis R, Li Z (2018) A conservative MHD scheme on unstructured Lagrangian grids for Z-pinch hydrodynamic simulations. J Comput Phys 357:206–229. https://doi.org/10.1016/j.jcp.2017.12.014

Wu ZN, Shen YX, Wang HL, Wu MM (2020) Urban flood disaster risk evaluation based on ontology and Bayesian Network. J Hydrol 583:15. https://doi.org/10.1016/j.jhydrol.2020.124596

Xie K, Ozbay K, Zhu Y, Yang H (2017) Evacuation zone modeling under climate change: a data-driven method. J Infrastruct Syst 23(4):040170131–040170139. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000369

Yin Z, Feng Q, Wen X, Deo RC, Yang L, Si J, He Z (2018) Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. Stoch Env Res Risk Assess 32(9):2457–2476. https://doi.org/10.1007/s00477-018-1585-2

Yuan X, Chen C, Lei X, Yuan Y, Adnan RM (2018) Monthly runoff forecasting based on LSTM–ALO model. Stoch Env Res Risk Assess 32(8):2199–2212. https://doi.org/10.1007/s00477-018-1560-y

Zhu S, Luo X, Yuan X, Xu Z (2020) An improved long short-term memory network for streamflow forecasting in the upper Yangtze river. Stochastic Environ Res Risk Assess. https://doi.org/10.1007/s00477-020-01766-4

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