

A rapid forecast method for the process of flash flood based on hydrodynamic model and KNN algorithm

Nie Zhou (✉ NieZhou@whu.edu.cn)

Wuhan University

Jingming Hou

Xi'an University of Technology <https://orcid.org/0000-0002-9097-3804>

Hua Chen

Wuhan University School of Water Resources and Hydropower Engineering

Guangzhao Chen

Xi'an University of Technology

Bingyi Liu

Wuhan University School of Water Resources and Hydropower Engineering

Research Article

Keywords: Rapid forecast, flash flood, KNN algorithm, hydrodynamic model

Posted Date: October 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-2118609/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A rapid forecast method for the process of flash flood based on hydrodynamic model and KNN algorithm

Nie Zhou ¹, Jingming Hou ^{2*}, Chen Hua ¹, Guangzhao Chen ², Bingyi Liu ¹

¹ State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China

² State Key Laboratory Base of Eco-Hydraulic Engineering in Arid Area, Xi'an University of Technology, Xi' an, Shaanxi, China

* Corresponding author: Prof. Jingming Hou, State Key Laboratory Base of Eco-Hydraulic Engineering in Arid Area, Xi'an University of Technology, NO.5 Jinhua Road, Xi'an, China, 710048, E-mail: jingming.hou@xaut.edu.cn

Abstract

Using hydrodynamic models to carry out early warning and forecasting of flash floods is an essential non-engineering measure to reduce the loss. However, most current hydrodynamic models have insufficient forecasting timeliness. To address this issue, a method based on the hydrodynamic model and the K nearest neighbours (KNN) algorithm is proposed to realize a rapid forecast for the process of flash floods. With the rainfall sequence as the input data and the simulation results of the hydrodynamic model as the target data, the rapid forecast of water depth, water velocity and discharge are achieved. Taking the Baogai Temple Basin as the study area, the rapid forecast model (RFM) is established, and its reliability and timeliness are verified. The results suggest that the established model has high accuracy, with 99% of the test data able to control the error of accumulated inundation extent within 20% and the NSE value of the cross-sectional discharge reaching 0.98. For 75% of rainfall scenarios, the maximum average

water depth and velocity error of the cross-sections can control within 7.5% and 10%, respectively. It also has dramatically improved calculation time, and can complete the prediction of the flooding process for the next 10 hours within 25s. It can gain valuable lead time for emergency decision-making and has broad application prospects in dealing with flash floods.

Keywords

Rapid forecast, flash flood, KNN algorithm, hydrodynamic model

1 Introduction

Flash flood is one of the most destructive natural disasters, with the characteristics of rapid water flow, strong scouring force, and incredible destructive power (Ji et al. 2018). Meanwhile, flash floods are highly catastrophic, and there is no effective engineering measure to intercept flash floods currently. It is prone to pose a severe threat to transportation, agriculture and downstream residents. According to the statistics of the Ministry of Water Resources of China, the number of deaths and disappearances caused by flash floods in China accounts for approximately 68.5% of those deaths and disappearances caused by flood disasters (He et al. 2016). Therefore, how to cope with flash floods and reduce losses has become the focus and difficulty of natural disaster prevention research.

Establishing the flash flood forecasting system is an essential non-engineering measure to deal with flash floods (Montanari et al. 2013). The two-dimensional hydrodynamic model has a clear physical meaning. Based on high-precision remote sensing information, it can simulate the entire process of flood evolution and has been widely used in flood simulation (Berardi et

al. 2013; Zhang et al., 2018). However, the hydrodynamic model requires high hardware equipment and computing capabilities, and the simulation process is time-consuming. How realize the real-time simulation and forecast of the flash flood process is still a challenging problem.

Many scholars have conducted extensive research on the accelerated computation of hydrodynamic models. For example, Hou et al. (2021), based on the idea of a non-uniform grid, use fine grids to represent areas of high relief and coarse grids to represent areas of low relief to improve the calculation rate of the model by reducing the number of grids. Hu et al. (2019) took the conservation characteristics of the algorithm as the starting point, used the concept of the local time step, and gave different time steps to the local grid under the premise of ensuring the stability of the calculation process, thus significantly reducing the overall calculation time. Xu (2016) and Hou et al. (2020) used the Graphic Processing Unit (GPU) acceleration to reduce computation time through multi-threaded computation by shifting the running computational process from the central processing unit (CPU) to the GPU. Kao et al. (2018) proposed multiple graphics cards parallel computing, which allocates computational work to each graphics card and invokes multiple cards simultaneously for accelerated computation to improve the overall computational performance of the model. These studies have achieved specific results, but due to the complexity of the hydrodynamic model solving process, there is still a large gap in achieving real-time simulation of the flooding process.

AI (artificial intelligence) technology has continuously developed for nearly two decades. Machine learning, a core component of AI technology, has also made significant progress,

showing the advantages of being extraordinarily applicable and efficient. By directly establishing the relationship between the input data and the target outcomes, rapid and efficient prediction of the target results can be achieved at a low computational cost (Mekanik et al. 2013). It has been widely used in image recognition, unmanned driving and medical health fields, showing outstanding advantages of high efficiency and reliability (Shin et al. 2016; Li et al. 2021; Briguglio et al. 2021).

KNN algorithm is a machine learning algorithm developed on the basis of a vector space model, which has the advantages of mature theory, stable calculation and good robustness. It has been widely used in image classification, medicine, and unmanned vehicles.(MA et al. 2010; Uddin et al. 2022; FOUAD et al. 2022) Meanwhile, it gained some application in flood forecasting in recent years, showing accurate and efficient characteristics, and has achieved a series of results. For example, Han et al. (2015) used the KNN algorithm to correct the real-time flood forecast in mountain areas. Their research showed that the KNN algorithm performed well and could effectively reduce the flood peak prediction error. Liu et al. (2017) applied KNN and Kalman filtering to flood forecasting in the Huaihe river basin. Their results showed that the KNN algorithm could predict the flood situation effectively in the next 8 hours. Naghibi et al. (2020) applied the KNN and RF algorithms to the rationality analysis of flood control engineering construction. Using topographic, hydrogeological, hydrological and stratigraphic lithological parameters as input conditions, they could quickly judge whether the area was suitable for flood control engineering construction. Their research showed that the areas under the ROC curves of the RF and KNN models are 97.1% and 94.6%, respectively,

indicating that RF and KNN can be effectively applied in flood control projects. Shahabi et al. (2020) combined the KNN algorithm with cosine, cubic, and weighted to forecast flooding in northern Iran spatially and showed that their model enhanced the predictive accuracy of the Cubic-KNN model (AUC=0.660), which they recommended for broader application to the sustainable management of flood-prone areas. Hou et al. (2021) applied the KNN algorithm to the rapid prediction of urban inundation. The model can complete the prediction of urban inundation in the 20s, which can gain a lot of lead time for dealing with short-duration heavy rainfall scenarios. Khatri et al. (2022), combining the KNN algorithm, random forests, and GIS tools, constructed a model to forecast flood locations in Mumbai in August 2020, and the results showed that their model was able to identify some of the flood locations correctly and could be used to predict locations with a high probability of flooding.

These studies show that the KNN algorithm provides the possibility for real-time flash flood forecasting and has broad application prospects in flood forecasting. The application of machine learning algorithms requires sufficient training samples, and the current rapid flood forecasting process is usually based on several years of historical data from hydrological stations. However, extreme flood events are limited and difficult to measure effectively, so machine learning algorithms still have limitations in practical applications.

This study aims to combine the hydrodynamic model and KNN algorithm to propose an RFM that is less dependent on historical measured data and has a broader range of applications. Based on the digital elevation data, the hydrodynamic model is used to simulate the flood routing process under various rainfall scenarios, providing sufficient training data. Directly

establishes the correspondence between the rainfall scenarios and the flooded results through the KNN algorithm and realizes the real-time forecast of the process of the flood. This study took the Baogai Temple Basin in Hunan Province, China, as an example to construct a forecast model and took the inundation extent, water depth, velocity and discharge as evaluation indexes to verify the reliability of the established rapid prediction model.

The main composition of this paper is as follows. After a brief introduction, the framework for constructing the RFM, the KNN algorithm, and the hydrodynamic model are presented. Next, the RFM for flash floods is constructed using the Baogai temple basin as the study area. The model's reliability is analyzed in Section 4, and the conclusions are summarized in Section 5.

2 Methodology

This section mainly introduces the overall structural framework of the realization of flash flood forecasting and the hydrodynamic model and KNN algorithm used in the realization process.

2.1 Framework of the RFM

The framework of the RFM is shown in Figure 1.

Step 1: generate sufficient rainfall data to cover rainfall under different conditions;

Step 2: use the hydrodynamic model to simulate the rainstorm-induced flash flood information corresponding to all grids in the area with various rainfall conditions based on DEM (Digital Elevation Matrix) and the infiltration capacity of the underlying surface in the study area;

Step 3: Divide the obtained rainstorm data into a training set and test set;

Step 4: Use rainfall data as the input parameters of the RFM and the results of the hydrodynamic model as the corresponding output data, and use the training set data for model training and use the test set to debug, optimize and improve the forecast model until the forecast performance of the model meets the requirements;

Step 5: Verify the forecasting ability of the rapid forecasting model by driving it with the verified rainfall data.

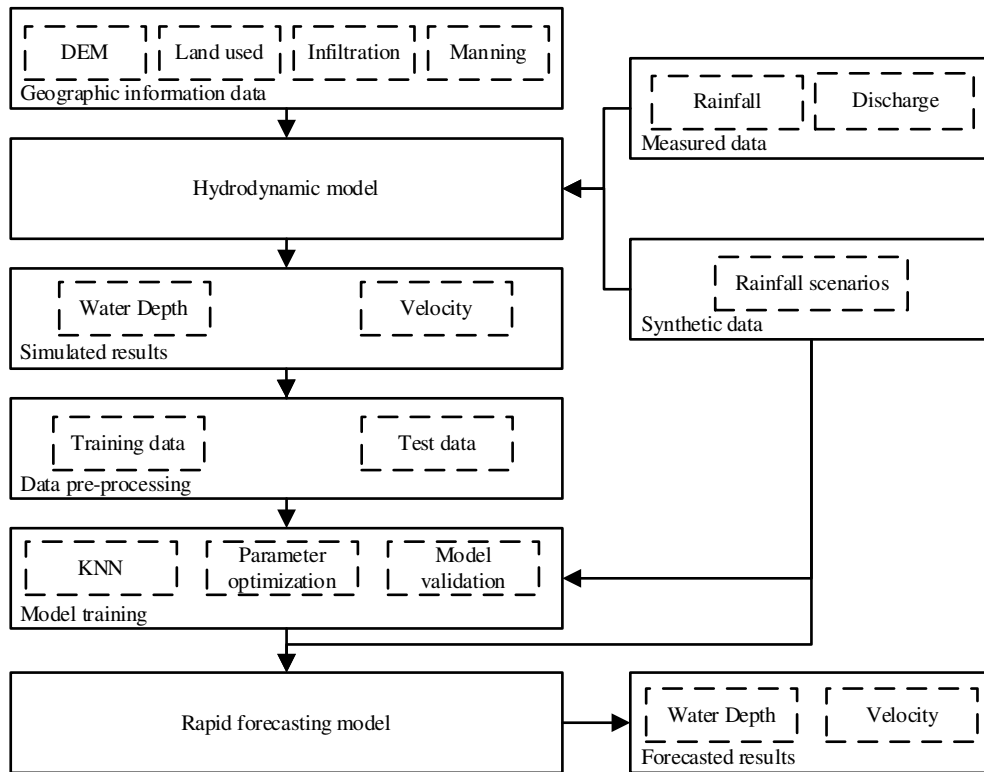


Figure 1. Framework of the RFM

2.2 Hydrodynamic model

The hydrodynamic-based urban flood model takes the 2D SWE (two-dimensional shallow water equation) coupled with hydrological processes, ignoring kinematic viscosity terms, Coriolis force, wind stress and turbulence viscosity terms. Its conservation scheme can be expressed in vector form as Eqs (1-3):

$$\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = \mathbf{S} \quad (1)$$

$$\mathbf{q} = \begin{bmatrix} h \\ q_x \\ q_y \end{bmatrix}, \mathbf{F} = \begin{bmatrix} uh \\ uq_x + gh^2/2 \\ uq_y \end{bmatrix}, \mathbf{G} = \begin{bmatrix} vh \\ vq_x \\ vq_y + gh^2/2 \end{bmatrix} \quad (2)$$

$$\mathbf{S} = \mathbf{S}_f + \mathbf{S}_b = \begin{bmatrix} i \\ -gh\partial z_b / \partial x \\ -gh\partial z_b / \partial y \end{bmatrix} + \begin{bmatrix} 0 \\ -C_f u \sqrt{u^2 + v^2} \\ -C_f v \sqrt{u^2 + v^2} \end{bmatrix} \quad (3)$$

142 where t represents the time; x and y are the Cartesian coordinates; $\mathbf{q}$ denotes the vector of
 143 conserved flow variables consisting of h and q_x and q_y , which are the water depth and unit-
 144 width discharges in the x and y directions; $q_x = uh$ and $q_y = vh$, where u and v are depth-averaged
 145 velocities in the x and y directions, respectively; $\mathbf{f}$ and $\mathbf{G}$ are the flux vectors in the x and y
 146 directions, respectively; g is gravity; $\mathbf{S}$ represents the source vector that may be further
 147 subdivided into net rain source terms i , slope source terms S_b and friction source terms S_f ; Z_b
 148 represents the bed elevation; C_f depends on the Manning coefficient and can be expressed as C_f
 149 $= gn^2/h^{1/3}$, where n is the Manning coefficient.

150 The model adopts the dynamic wave method, which considers that physical problems are
 151 the most comprehensive to simulate the flood routing process, and uses Godunov's finite
 152 volume method to discretely solve the 2D SWE. Through the explicit second-order Runge-
 153 Kutta (Hubbard 1999) method, we constructed a MUSCL (Monotonic Upwind Scheme for
 154 Conservation Laws) with second-order space-time accuracy to ensure the conservation of
 155 matter and effectively solve the discontinuity problem (Hou et al. 2015). To solve complex
 156 problems such as abrupt flow and discontinuity, the model uses the HLLC (Haren-Lax-van Leer
 157 contact) approximate Riemann solver to solve the mass and momentum flux on the unit
 158 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 grids, 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 to maintain the stability of the calculation results (Hou et al. 2013), and GPU parallel technology is used to ensure the model's computational efficiency.

2.3 KNN algorithm

KNN algorithm was proposed by Cover and Hart based on the vector space model (Keller et al. 1985). After continuous improvement, it has become a primary classification and regression algorithm in machine learning. The KNN algorithm's input is the instance's feature vector, which corresponds to the point in the vector space, and the output is the category or regression value of the instance (Violetto & Noro 2019; Liu et al. 2016). The KNN algorithm assumes that given an instance category or real-valued training dataset when performing regression or classification calculations, the KNN algorithm will continuously divide the feature vector space. For a new instance, the distance between it and each existing instance in the training dataset will be calculated using the distance formula, from which the nearest K instances will be selected and the new instance predicted by majority voting and averaging. The

180 choice of K value, distance measurement and decision rules are the three essential elements of
 181 the KNN algorithm.

182 When the input sample and the K value are determined, the KNN algorithm will return the
 183 K samples closest to the input sample, as shown in Eqs (4-5):

$$\forall \beta \in NN_{\alpha}(K), \forall \gamma \in TD - NN_{\alpha}(K) \quad (4)$$

$$d(\alpha, \beta) < d(\alpha, \gamma) \quad (5)$$

184 Where the β represents the K closest samples to input sample α , γ are the other samples
 185 that besides the K closest samples to the sample α , $NN_{\alpha}(K)$ is sample set comprised of the K
 186 closest samples to the input sample α , TD represents the all sample dataset, $d(\alpha, \beta)$ as the
 187 distance between the sample α and β , and $d(\alpha, \gamma)$ as the distance between the sample α and γ .

188 Assuming that the sample is represented by an N-dimensional space vector, that α can be
 189 represented by $X_{\alpha} = (x_{\alpha 1}, x_{\alpha 2}, x_{\alpha 3}, \dots, x_{\alpha n})$, and β can be represented by $X_{\beta} = (x_{\beta 1}, x_{\beta 2}, x_{\beta 3}, \dots, x_{\beta n})$,
 190 the distance between the sample α and β can be represented by Eq (6).

$$d(\alpha, \beta) = \left(\sum_{i=1}^n (x_{\alpha i} - x_{\beta i})^p \right)^{\frac{1}{p}} \quad (6)$$

191 Where $d(\alpha, \beta)$ is the Euclidean distance between sample α and β , $x_{\alpha i}$ and $x_{\beta i}$ are the i
 192 eigenvalue of sample α and β respectively; p is a constant greater than or equal to 1, where ,
 193 $d(\alpha, \beta)$ is the Manhattan distance when $p = 1$, the Euclidean distance when $p = 2$, and the
 194 maximum of the coordinates of each distance when $p = \infty$. In two-dimensional space, with
 195 different values of p , the graphs formed by the points at a distance of 1 from the origin $d(\alpha, \beta)$
 196 are shown in Figure 2.

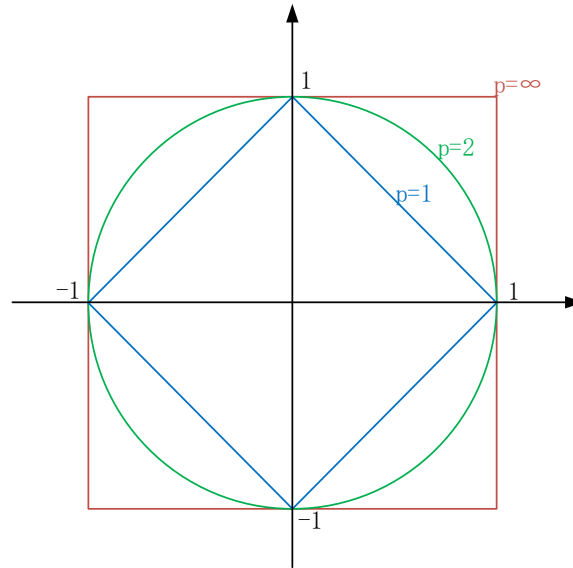


Figure 2. Schematic diagram of distance in two-dimensional space

When dealing with classification problems, the information provided by the KNN is often combined by voting, meaning that the outcome with the highest frequency of occurrence will be used as the predicted classification outcome. In regression problems, the mean value is often used for determination. That is, the mean of the target values of the KNN will be used as the target outcome for the new sample. Additionally, the model can also assign different weights to the nearest neighbours according to the proximity between the input samples and the nearest neighbours and determine the target outcome of the new samples by means of inverse distance weights. In this study, the inverse distance weights were used to determine the final forecast results so as to take into account the similarity of the rainfall scenarios.

2.4 Model performance Indicators

In order to comprehensively verify the performance of the RFM, the Inundation extent, water depth and velocity were selected as the verification data, and various model evaluation indexes were used for analysis, including R^2 (R-square), RMSE (Root mean square error), NSE (Nash-Sutcliffe coefficient) and RME (relative average error).

213 R^2 is defined as Eq (7), and the range of R^2 is 0-1, the closer R^2 value is to 1, the better
 214 fitting effect is.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

215 RMSE is shown as Eq (8), and the closer RMSE value is to 0, the better fitting effect is.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \quad (8)$$

216 The Nash-Sutcliffe coefficient was proposed by Nash and Sutcliffe in 1970, as shown in
 217 Eq (9). The value of NSE is negative infinity to 1, and NSE is close to 1, indicating that the
 218 model is of good quality and the model has high reliability; NSE is close to 0, indicating that
 219 the simulation result is close to the average level of the observations, that is, the overall result
 220 is credible, but the process simulation error is large; NSE is far less than 0, the model is not
 221 credible.

$$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

222 RME defined as Eq (10), and the closer the RME is to 0, the better fitting effect is.

$$RME = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad (10)$$

223 where y_i is the measured data, $\hat{y}_i$ means the simulated data, $\bar{y}$ is the mean value of the
 224 measured data, n is the amount of data.

225 **3 Applications**

226 This part takes the Baogai Temple Basin as an example to introduce the basic data required
 227 to establish the RFM and the process of building the model.

3.1 Study area

Baogai Temple Basin, located in Hunan Province, China, belongs to a humid subtropical climate zone with hot and humid summer but cold and dry winter and an average annual rainfall of about 1569mm. The total area of the basin is about 22.1km² (Liu et al. 2019), with a complex topography and steep river banks, making it prone to flash floods and geological hazards. There are several branches in the basin, and the main river is Baogai River, with a total length of about 9.1km. A reservoir is in the downstream, and the prediction of the upstream flow is of great significance for the reservoir operation.

3.2 Datasets for real-time forecasting

3.2.1 Underlying surface data

Terrain data is an important factor that affects the reliability of the model. If the accuracy of terrain is too coarse, it will not accurately reflect the underlying surface information and cause large errors in the simulation results. As the terrain accuracy increases, the simulation effect will gradually improve, but the simulation calculation cost will also increase exponentially. To accurately reflect the river information, DEM data with a resolution of 10m (567,312 grids) was chosen for this study, and the DEM information of the study area is shown in Figure 3.

DEM information can reflect the elevation information of the surface but cannot represent the land use of the underlying surface. For this reason, combined with the DOM (Digital Orthophoto Map) data of the study area, using the maximum likelihood classification method, the study area is divided into two categories according to the land used (Figure 4), including

sloping fields and rivers. Based on the research of Ma et al. (2015), the manning coefficient of the river channel is set as 0.02, the infiltration capacity is set as 20mm/h, the manning coefficient of the sloping field is set as 0.2, and the infiltration capacity is set as 6mm/h.

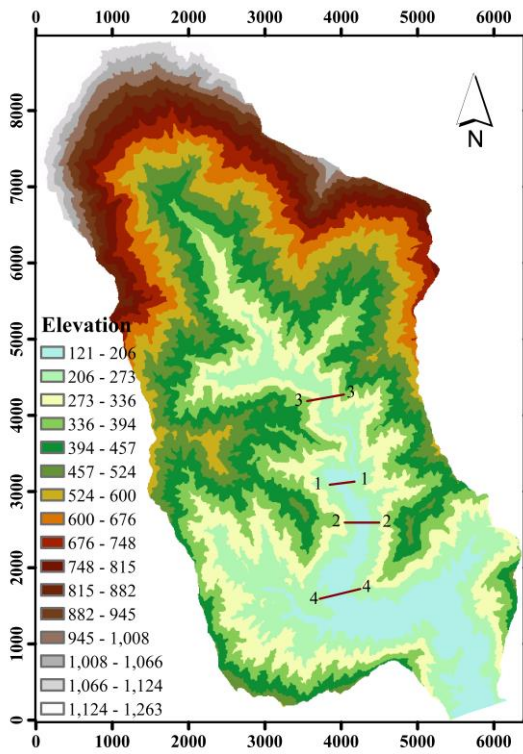


Figure 3 The DEM of study area

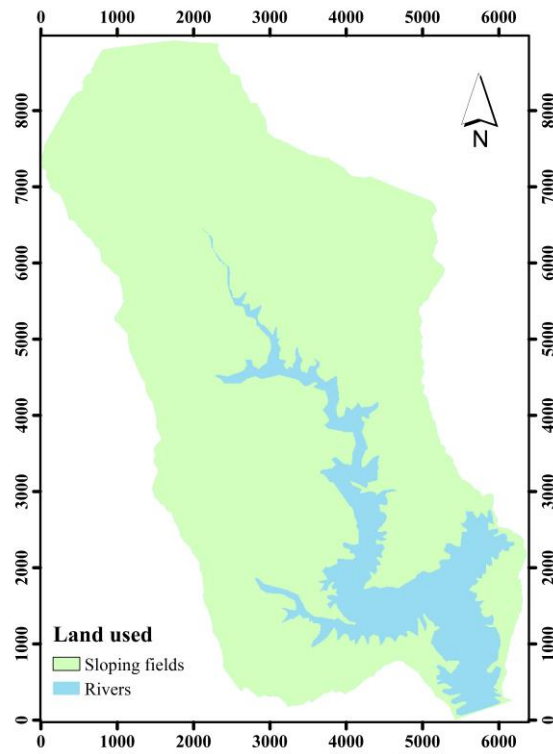


Figure 4 The Land used of study area

3.2.2 Rainfall data

The formation of a flash flood is usually caused by prolonged rainfall or locally short-duration heavy rainfall, and the rapid forecast method of flash flood proposed by the author mainly aims at the flash flood caused by the locally short-duration heavy rainfall. Since insufficient measured extreme heavy rainfall events can't meet the requirements of model training, based on the short rainfall duration and heavy rainfall characteristics of the study area, this study consulted the latest local hydrological manual and constructed 260 design rainfall scenarios based on the study area's rainstorm formula (Eq 11) and the designed rain pattern (Shao 2014). For the accuracy of the short-term numerical forecast models can reach 1 h, and

based on weather radar, can develop 0-6h high-precision nowcasting (Li et al. 2021; Liu et al. 2020), the RFM is constructed the rainfall data with a time resolution of 30min to reflect the process of flood more accurately. The designed rainfall return period covers one-year to 500-year scenarios, and the rainfall duration is mainly focused on 2h, 4h and 6h, which can cover most of the short-duration floods in the study area (Table 1).

$$i = \frac{472.029 + 421.3056 \lg T}{(t + 68.7627)^{1.3699}} \quad (11)$$

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

Table 1 Parameters of design rainfall

Cases	The rain peak coefficient	return period (a)	duration of rainfall (minutes)
1	0.35	1	120
2	0.35	1	240
3	0.35	1	360
...	...	...	...
100	0.45	50	120
101	0.45	50	240
102	0.45	50	360
...	...	...	...
258	0.45	500	120
259	0.45	500	240
260	0.45	500	360

3.3.3 Generate training data of fast forecast model

Sufficient training data is the basis for the construction of rapid forecasting models. The DEM information, land use and rainfall data are used as input conditions for the model to drive the hydrodynamic model to simulate the process of storm-induced flooding. For the heavy rain and humid air when flash floods form, the process of water vapour evaporation is not considered in the simulation process. The downstream outlet is set to free flow out of the open boundary, and the other positions are set as the closed boundary. That is, during the simulation process, only the infiltration and the process of forming a confluence and flowing to the outlet of

rainwater are considered. At the same time, the model calculation grid resolution is set to 10m. To ensure the stability of the calculation, the Courant number is set to 0.5. The model outputs the 10h process of rainfall-induced flooding with a time interval of 30min of each rainfall, and each set of result files is composed of 567,312 grids of water depth and velocity data. After simulations, 260 rainfall-induced flooding results were finally formed, and a total of 5,200 sets of result files were obtained. Based on this, a machine learning sample information database was established.

3.3 RFM construction

The RFM construction is mainly completed through python3.6, Scikit-learn, Numpy and Pandas frameworks. The model takes the rainfall sequence as the input data and the hydrodynamic model simulation result file as the target data for training.

The sample data obtained from the hydrodynamic model simulations were divided into a training set and a test set according to 9:1, with 234 rainfall fields as the training set and 26 rainfall fields as the test set. The validation dataset was not distinguished because the K-fold cross-validation algorithm was used in the model training process. The training set data is used to train and construct the RFM, and the test set is used to test the forecast performance of the model.

The framework of the proposed RFM showed in Figure 5. When training the model, the input data is composed of a 234*12 matrix, where 234 represents 234 training rainfall data, and 12 is the number of input parameters taken for each rainfall. This research is mainly for short-duration heavy rainfall scenarios, the rainfall duration is less than or equal to 6h, and its time

resolution is 30min. Based on this, the RFM sets a total of 12 parameters for each rainfall. The target data is composed of a multi-dimensional matrix of $234 \times 20 \times 3 \times 567,312$, where 234 is the number of accumulated training rainfall sessions; 20 means 20 moments, this research simulates the next 10h flooding process formed by short-duration heavy rain, and output result files at 30min intervals, so there are 20 moments of result files for each rainfall; 3 is the number of data types that constitute the result file, including water depth and the velocity in x and y direction; 567312 is the number of grids that constitute the result file.

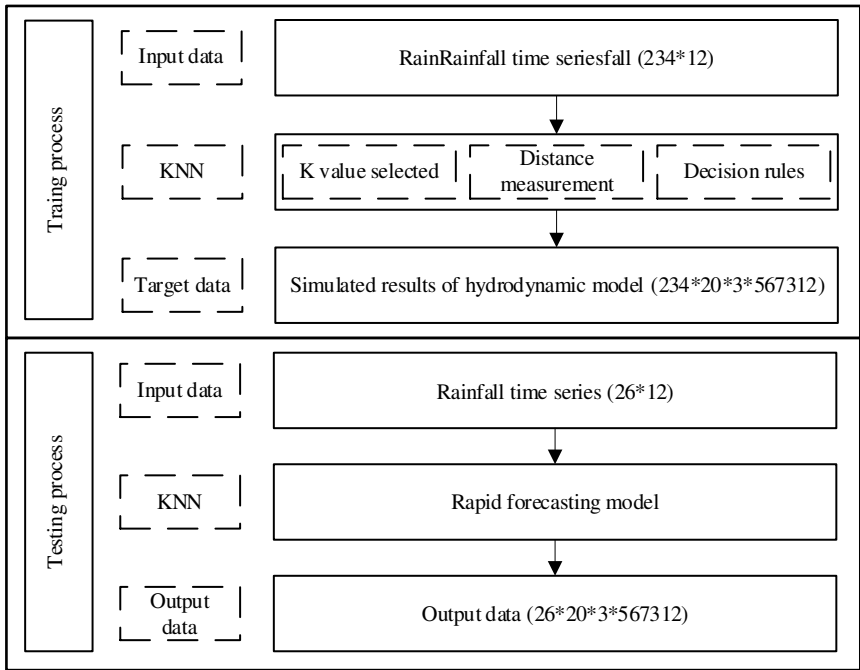


Figure 5 The framework of the proposed RFM

Then, model training is performed according to the input parameters and target data, where each row of input parameters corresponds to the output result one-to-one. In the process of model training, considering the internal parameters of the algorithm, such as the number of neighbours K in the KNN algorithm, the selection of the distance formula, and the way of determining the prediction result will all have a greater impact on the accuracy parameters of the model's prediction performance. Accordingly, in the process of model training, a grid search

algorithm is used, which is an exhaustive algorithm. That is, the possible values of all parameters of the model algorithm are assigned to the matrix, and it will automatically select the parameters one by one, combine and perform simulation training, and select the best parameter combination for the simulation of training data through the cross-validation method to improve the reliability of the RFM. The main parameter selected for the KNN model is shown in Table 2.

Table 2 Selection of main parameters of KNN algorithm

Parameters	K value	Distance measurement	Decision rules	Algorithm
Value	3	Manhattan distance formula	Uniform weights	KD tree

Finally, 26 field test data were used as input data, and the corresponding forecast results were obtained through the established RFM. The reliability of the prediction results was verified by comparing them with the simulation results of the hydrodynamic model.

4 Results and discussion

4.1 Analysis of hydrodynamic model

Since insufficient historical extreme heavy rainfall events to meet the model training requirements, this study constructs a hydrodynamic model to simulate the flood inundation data under various extreme short-duration heavy rainfall scenarios. The RFM is constructed by using the KNN algorithm to establish a direct link between the rainfall data and the hydrodynamic model simulation results. The reliability of the hydrodynamic model will directly affect the reliability of the constructed RFM. Therefore, the study firstly verifies the reliability of the constructed hydrodynamic model.

Using the Baogai Temple basin as the study area, the hydrodynamic model was calibrated using a heavy rainfall event on 9 May 2012, monitored by the Qing Shui hydrological station

in the basin. The rainfall lasted from 05:00 to 21:00, with a cumulative rainfall time of 16h and a cumulative rainfall of 130mm, with a maximum single hourly rainfall of 47.5mm. The main rainfall was concentrated between 07:00 and 13:00, during which the cumulative rainfall exceeded 125mm, accounting for 96% of the total rainfall, and the specific rainfall process is shown in Figure 6.

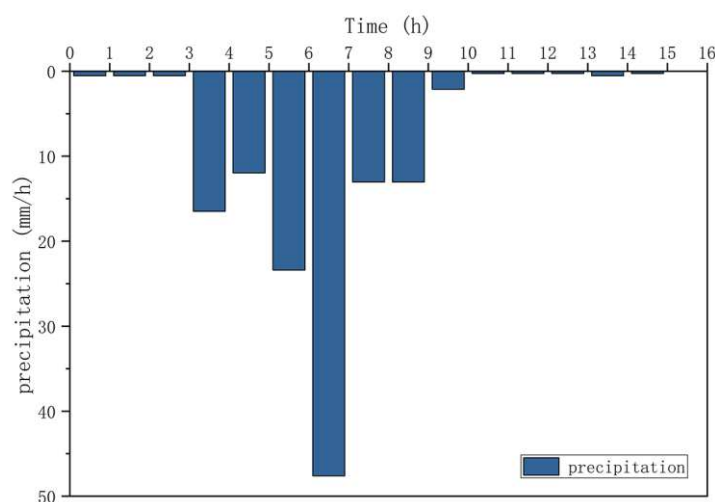


Figure 6 Measured rainfall process on May 9,2012

The downstream boundary is set as the open boundary of the free flow, and the remaining boundary is defined as the closed boundary. That is, the process of rainfall forming a flood in the river basin and flowing towards the outlet of the river basin was simulated without considering the inflow situation outside the river basin. The reliability of the model was verified by comparing the measured water level and flow process in section 1 (Figure 3).

Due to the heavy rainfall, the rainwater quickly converges in the basin, forming a flash flood and instantaneous flows in excess of 200 m³/s. The comparison between the simulation result and the measured discharge process in section 1 is shown in Figure 7. The figure shows that the simulated discharge process is in good agreement with the measured data, the NSE is 0.70, and the measured peak flow approximately occurs at the 8th hour, with a peak flow of

235m³/s. The peak flow obtained from the simulation of the hydrodynamic model based on physical processes also occurs at the 8th hour, with a simulated peak flow of 226.9m³/s, which is slightly less than the measured value, but the difference is less than 10m³/s, with a relative error of less than 5%, indicating that the simulation results are accurate and effective, and the simulation results of the built simulation in this basin have high accuracy.

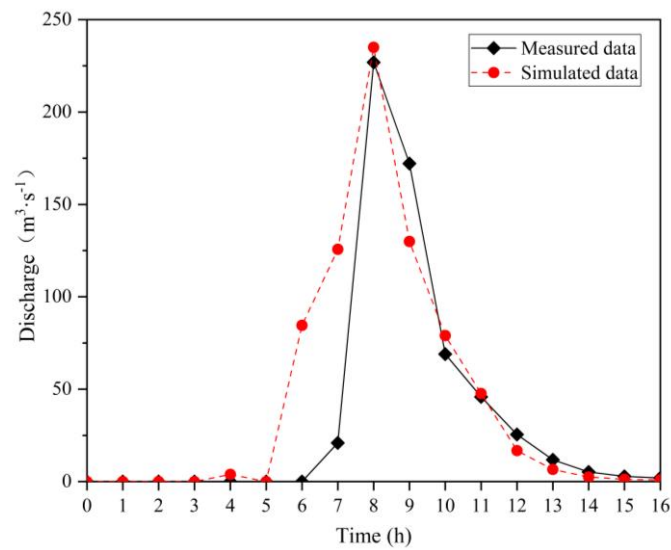


Figure 7 The comparison between the simulation result and measured discharge at section 1

However, despite using GPU acceleration in the model runs and calculations with a GeForce RTX 2070 super graphics card, it still took 2760 seconds to simulate the 16-hour flooding process. As a result, its timeliness is still difficult to guarantee in the face of short-term storms.

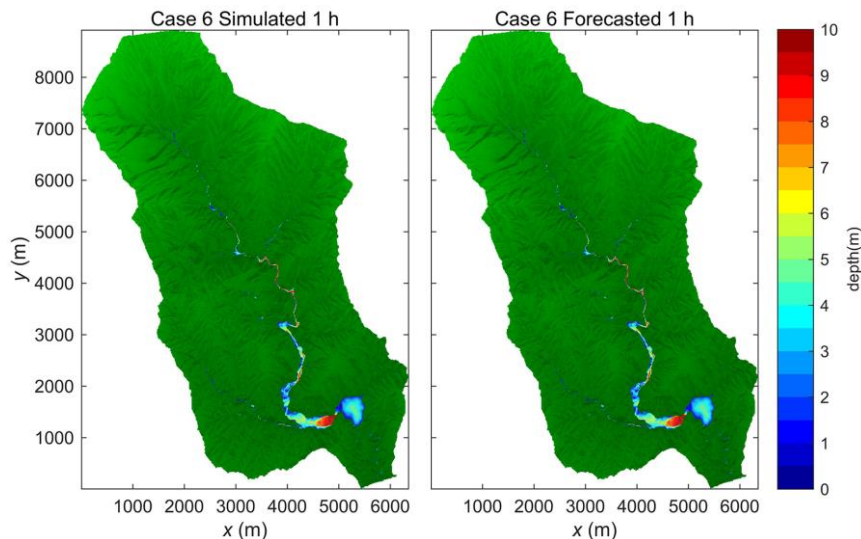
4.2 Analysis of real-time forecasting model

The 10-hour flood evolution of 260 rainfall events was simulated by the hydrodynamic model with a temporal resolution of 30 min. With 20 moments of data included in each rainfall scenario, a total of 5200 result files were available. At a ratio of 9:1, 234 of the rainfall scenarios were used as model training data, and 26 rainfall scenarios were used as test data. Inundation

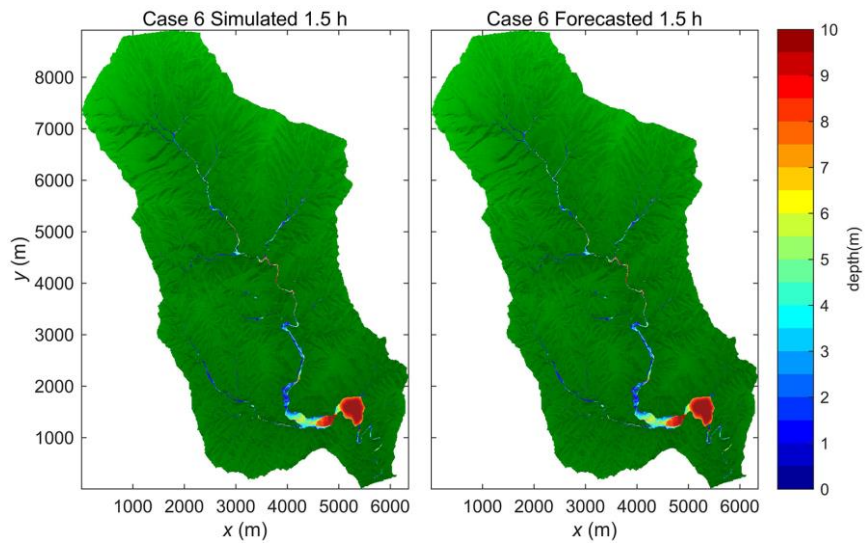
extent, water depth and flow velocity were used as the main analysis objects, and RMSE and NSE were used as the main indicators to test the prediction performance of the RFM. Specific analyses are presented below.

4.2.1 Analysis of the forecast results of whole basin

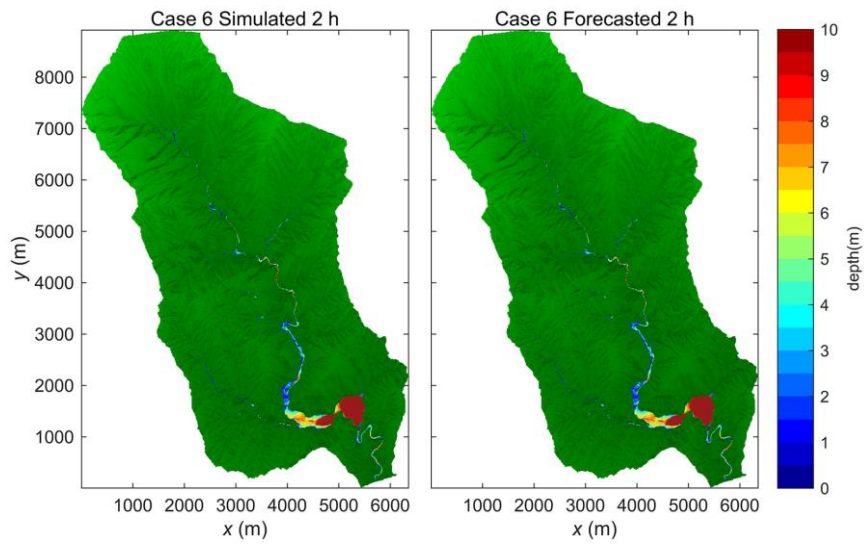
First of all, the inundation extent was analyzed by counting the water depth in each cell. In order to make the inundation boundary more obvious and avoid over-estimation of the inundation extent, the inundation area below with water depth below 0.3m was ignored in the process of drawing the inundation map and data analysis. That is, the water depth of fewer than 0.3m was set to 0. Part of rainfall scenarios flood routing map as shown in Figure 8. The inundation process obtained from the RFM varies smoothly without any abnormal water depth that violates the overall trend, which is consistent with the objective rule. Meanwhile, the prediction results are basically consistent with the overall situation of the hydrodynamic model simulation results, indicating that the constructed RFM is generally stable and reliable.



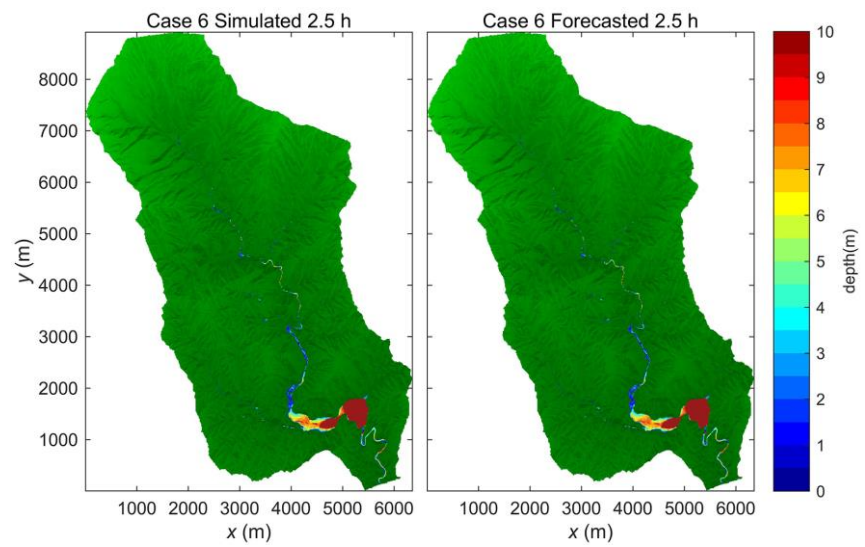
(a) $t = 1h$



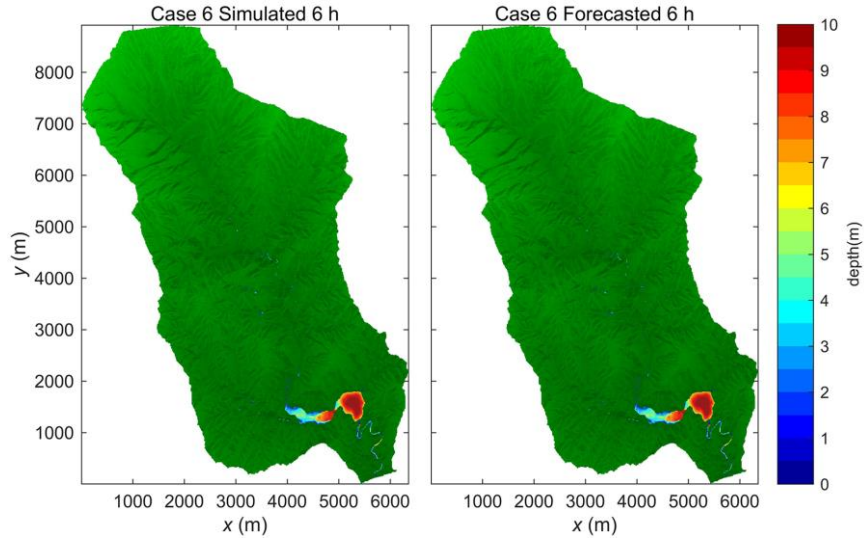
(b) $t = 1.5h$



(c) $t = 2h$



(d) $t = 2.5h$



(e) $t = 6h$

Figure 8 Comparison of inundation extent between hydrodynamic model and RFM

Analysis of the error of inundation extent at each moment showed in Figure 9. Of the cumulative 26 rainfall scenarios, 520 moments of forecasts, 99% of the data forecast errors can be controlled to within 20%. The results with larger forecast errors are mainly concentrated in the 1-2h time period at the beginning of the rainfall, and the forecast effects tend to be stable after 2h with smaller errors. This is mainly due to the fact that at the early stage of rainfall, the factors affecting the inundation extent are mainly the 1-2h rainfall sequence. Considering the usual temporal resolution of forecast rainfall, the time interval of rainfall sequence was set to 30min, resulting in fewer effective parameters and relatively larger flood forecast errors at the early stage of rainfall. With the increase in the duration of rainfall, the effective forecast parameters increase, the forecasting effect tends to be stable and accurate, and the inundation extent error can be controlled within 10%.

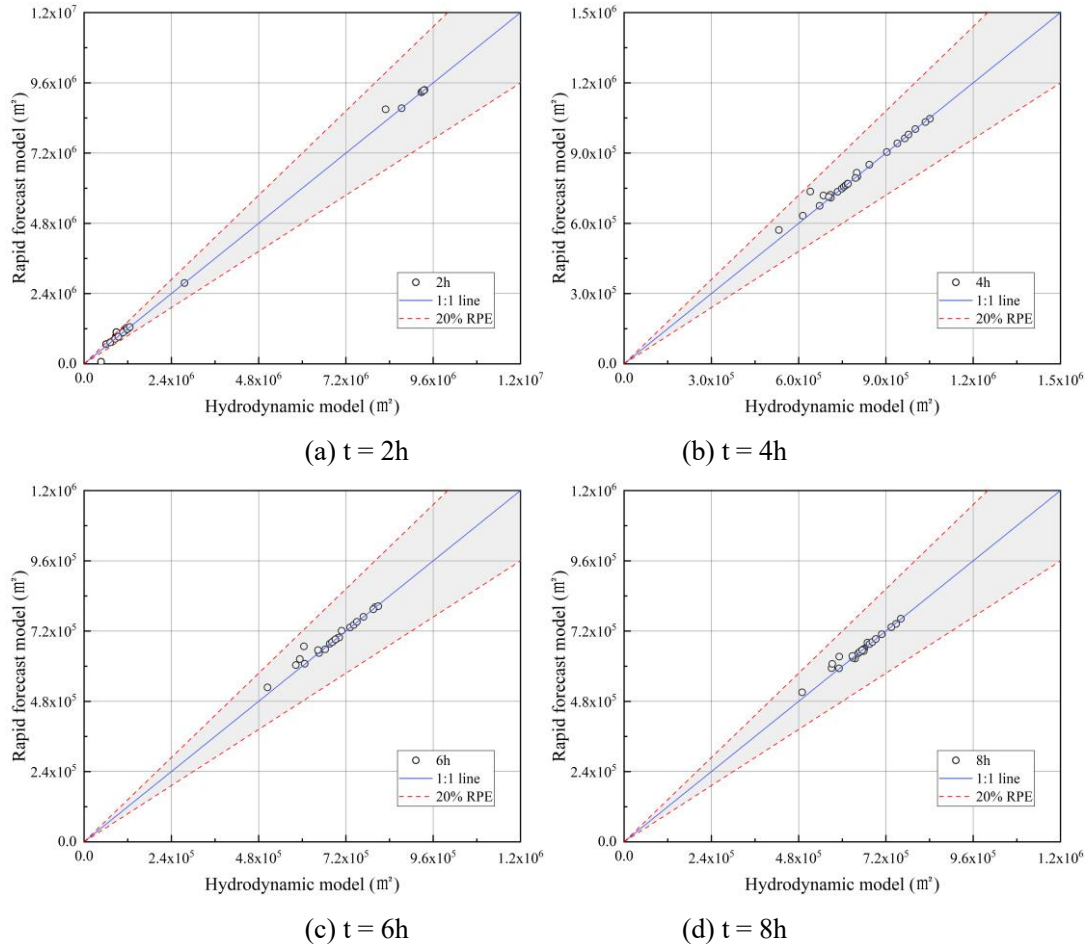


Figure 9 Error analysis of the inundation extent at each moment

To further quantify the forecast performance of the model, this study used RMSE and NSE as evaluation indicators to quantify the error at each moment. The average value of RMSE and R^2 of water depth is 0.990 and 0.044 (Table 3), which shows that the model forecast results are accurate and reliable, and the forecasted process is stable. The maximum error of RMSE (0.131) and the minimum value of R^2 (0.948) both occurred at the 2nd hour, which still had high accuracy.

Table 3 Error analysis of the water depth and velocity at each moment

Times (h)	Water depth		Water velocity	
	R ²	RMSE	R ²	R ²
0.5	1.000	0.004	0.892	0.006
1.0	0.968	0.029	0.832	0.049
1.5	0.984	0.037	0.794	0.112
2.0	0.948	0.131	0.816	0.105
2.5	0.985	0.092	0.856	0.088
3.0	0.989	0.054	0.831	0.066
3.5	0.992	0.047	0.773	0.048
4.0	0.993	0.045	0.787	0.036
4.5	0.993	0.043	0.787	0.029
5.0	0.994	0.040	0.808	0.016
5.5	0.994	0.038	0.812	0.014
6.0	0.994	0.037	0.844	0.012
6.5	0.995	0.036	0.855	0.011
7.0	0.995	0.036	0.853	0.010
7.5	0.995	0.035	0.837	0.009
8.0	0.995	0.035	0.877	0.008
8.5	0.995	0.034	0.866	0.008
9.0	0.995	0.034	0.844	0.007
9.5	0.995	0.034	0.843	0.007
10.0	0.995	0.034	0.816	0.006

In addition, the flash flood has the characteristics of high velocity, and high-speed water
 may cause debris flow, landslides and other secondary disasters, so the model also predicts the
 flow velocity. The prediction errors of the RFM for the flow velocity are shown in Table 3. By
 comparing the R² of water depth and flow velocity, it can be seen that the overall prediction
 performance of the RFM for flow velocity is slightly worse. This may be due to the magnitude
 and direction of water velocity being determined by the velocity components in the x and y
 directions, so there may be some accumulation of error. Nevertheless, the R² values for water
 velocity are all greater than 0.77, and the RMSE values are close to 0, indicating a high degree
 of agreement between the water velocity and the overall simulation results of the hydrodynamic
 model. The RFM is able to accurately reflect the velocity distribution within the basin.

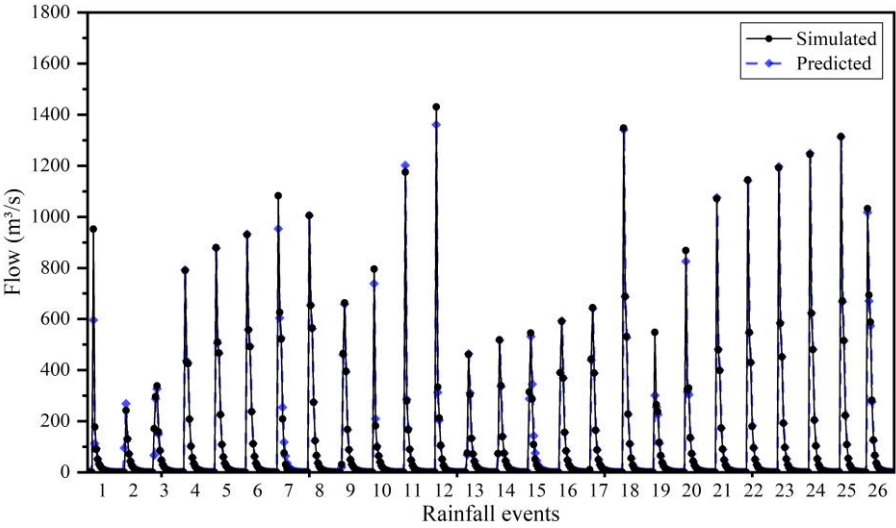
4.2.2 Analysis of the forecast results of local sections

As the study area is mountainous, the terrain is relatively undulating and the river channels

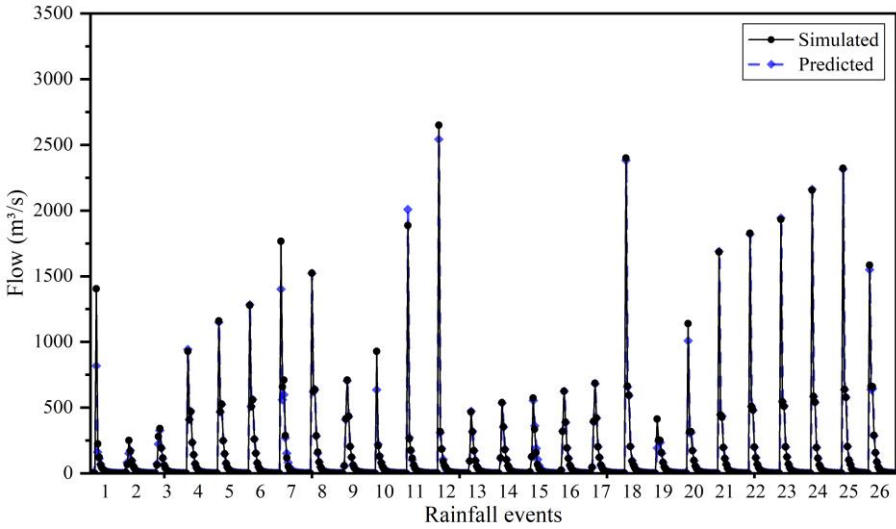
are narrow and long. Rainwater from high lying areas will rapidly collect along the slope towards the river channel, so that the water depth tends to be shallow in areas with high topography. In addition, when analyzing the overall inundation, the threshold value for water depth was set to 0.3 m in order to better distinguish the inundation extent, which may allow the model to obtain near-perfect forecasts in areas with high terrain, ultimately leading to an overestimation of the model's forecast performance. On this basis, four river sections were selected and the water depth, flow velocity and discharge in these sections were analyzed in more detail, and the locations of the sections are shown in Figure 3.

The section discharge for different rainfall conditions at each time period is shown in Figure 10. By comparing the simulation results of the hydrodynamic model and the RFM, it was revealed that the RFM has good forecasting effectiveness on the trend of the discharge and can reflect the time of peak flow occurrence at each section. Similar to the overall basin forecast results, the large forecast errors appear within 1-2h at the beginning of each rainfall, and the forecasting effect is to be stabilized after 2h. This may also be due to the small number of effective control parameters at the onset of rainfall, and by improving the temporal resolution of the forecast rainfall, the accuracy of the model in the pre-forecast can be further improved. Nevertheless, the NSE values between the simulation results of the hydrodynamic model and the RFM forecasts were calculated to be in excess of 0.98 for all sections, with a slightly smaller NSE value of 0.982 for section 1 at the upstream location of the river and the best prediction of 0.99 for section 4 at the downstream location of the river. In addition, the average error of peak flow for sections 1, 2, 3, and 4 were 7.16%, 7.87%, 6.17%, and 5.28%, respectively. The model's

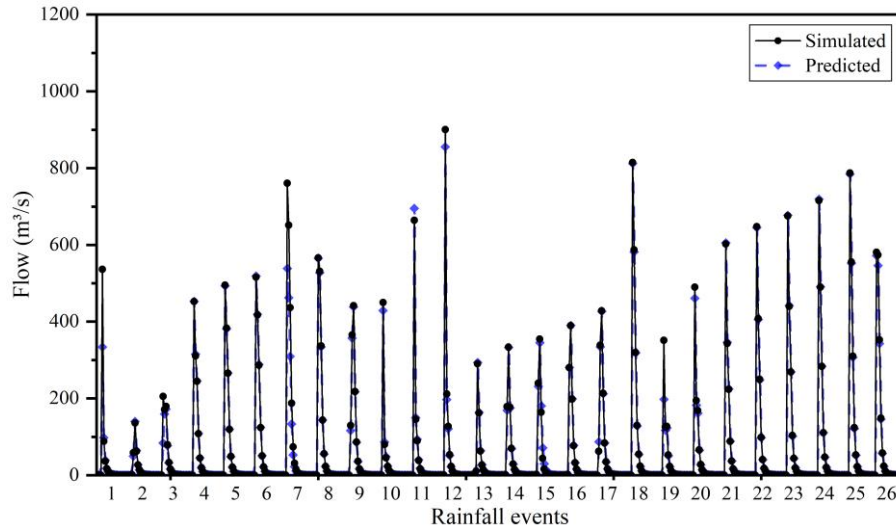
prediction of section discharge can meet the accuracy requirements of emergency decision-making.



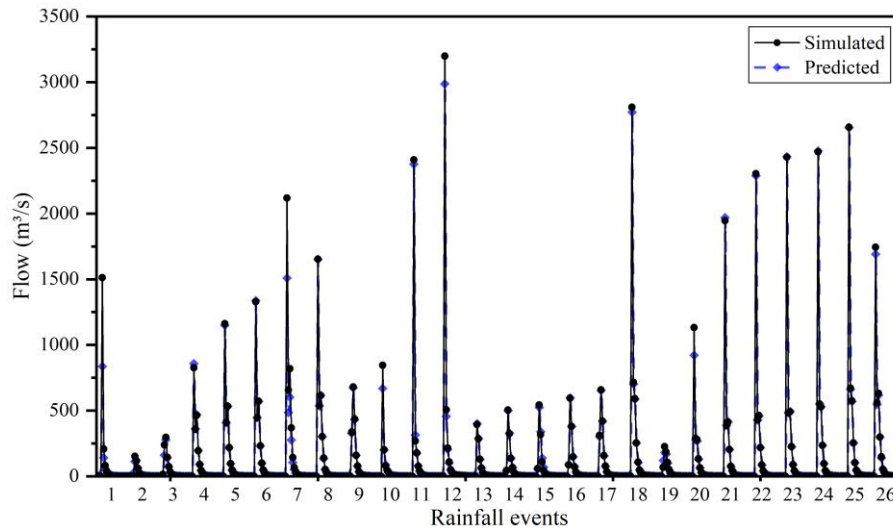
(a) The discharge of section 1



(b) The discharge of section 2



(c) The discharge of section 3



(d) The discharge of section 4

Figure 10 Error analysis of the discharge process of sections

The error analysis of the maximum average velocity of the four sections is shown in Figure 11. The RFM has the best forecasting effect on section 4 ($NSE=0.989$), and the forecasting effect of section 2 is slightly worse ($NSE=0.980$), but the NSE values of the four sections all reach 0.98, which is close to 1, indicating that the built rapid forecast results are highly consistent with the hydrodynamic model simulation results. The average errors of sections 1, 2, 3, 4 were 4.96%, 1.38%, 4.28%, and 3.94%, respectively, and the maximum velocity error of 75% can be controlled within 10% (Figure 11 (a)), demonstrating that the RFM can accurately

predict the maximum velocity of the section under most rainfall-induced flood conditions.

However, Figure 11

b) shows that the maximum mean velocity errors of the RFM are slightly larger in rainfall scenarios 1, 7 and 19. The forecasting still needs to be improved. This may be mainly due to the limited training data of the forecast model, which cannot fully cover enough types of rainfall scenarios. In practical application, it is necessary to further increase the training dataset.

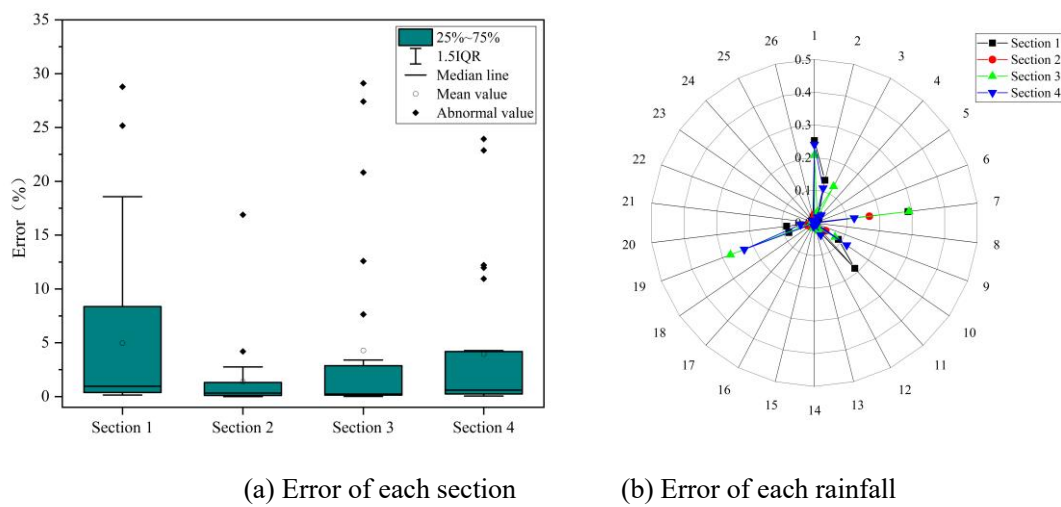


Figure 11 Error analysis of maximum average velocity of sections

The error analysis of the RFM for the maximum average water depth of the section is shown in Figure 12. Under 26 rainfall scenarios, the NSE values between the simulation results of the average water depth of the four sections and the simulation results of the RFM were 0.992, 0.987, 0.992 and 0.983, respectively. As shown in Figure 12 (a), the average error of the maximum average water depth of the sections for more than 75% of the rainfall scenarios can be controlled within 7.5%, and the relative average errors of sections 1, 2, 3, and 4 are 4.34%, 4.49%, 3.42% and 2.96%, respectively, indicating that the RFM achieves forecast results similar to those of the hydrodynamic model. Similar to the velocity forecast results, the RFM forecasts for rainfall scenarios 1, 10 and 19 still need to be improved (Figure 12(b)).

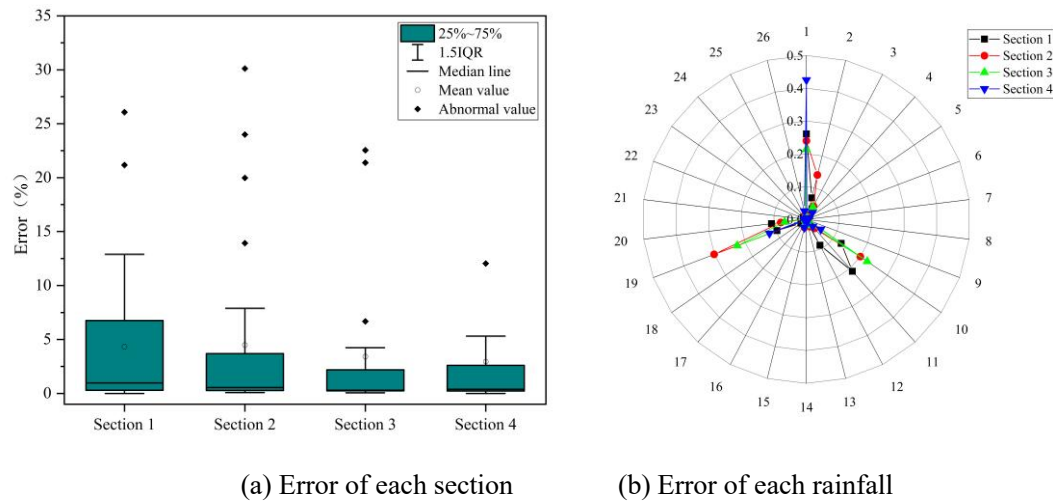


Figure 12 Error analysis of maximum average water depth of sections

4.3 Computational time

The hardware device of this study is based on NIVIDA Geforce GTX 2070 Super Graphics Card, and the CPU is Intel Core I7-8700. By using GPU acceleration techniques, the hydrodynamic model simulates a cumulative time of 40,092s for 26 rainfall scenarios and 1000-1800s for a 10h flooding process for a single rainfall scenario, and the greater the rainfall scenario, the longer the simulation time. The total time for the RFM to simulate 26 rainfall processes is 532s, and the average time for predicting a single rainfall scenario is about 20s, which can be shortened by 50-90 times compared with the hydrodynamic model. And the forecast time will not increase significantly with the increase of rainfall amount and rainfall duration, which is more advantageous in dealing with flash flood events caused by sudden short-duration heavy rainfall.

5 Conclusion

Based on the numerical hydrodynamic model and the KNN algorithm, this research proposes a method to realize the rapid prediction of the flash flood. As insufficient measured data for extreme rainfall events were available, a hydrodynamic model was used to generate the

training data covering flood processes under various types of return period synthetic short-duration heavy rainfall scenarios. And the Baogai Temple Basin was used as the study area to construct and test the reliability and timeliness of the RFM.

The results show that the built model realizes the rapid forecast of the process of flash floods and has high reliability. 99% of the test data can control the accumulated inundation extent error within 20%, and the NSE value of the cross-section discharge reached 0.98. For 75% of rainfall scenarios, the errors of maximum average water depth and velocity at the sections can be controlled within 7.5% and 10%, respectively. Meanwhile, the model can complete the prediction of the flood routing process in the next 10h within 25s. Compared with the traditional hydrodynamic numerical model, the simulation time can be shortened by 50-90 times, and forecasting efficiency has been greatly improved.

In conclusion, the RFM established in this research has the advantages of high efficiency and high precision, which can quickly respond to the flooding process caused by short-duration heavy rainfall. It can gain valuable lead time for emergency decision-making when responding to flash flood events caused by short-duration heavy rainfall and has broad application prospects in the rapid forecast of the flash flood.

Ethical Approval

Informed consent.

Consent to participate

Not applicable.

525 **Consent to Publish**

526 The authors are indeed informed and agree to publish.

527 **Authors contributions**

528 Conceptualization and Methodology: N. Zhou, JM. Hou; Writing-original draft
529 preparation: N. Zhou; Material preparation, collection and analysis: N. Zhou, H. Chen, GZ.
530 Chen; Supervision: BY. Liu; Funding acquisition: JM. Hou.

531 **Funding**

532 This work is partly supported by the Numerical Simulation of Flood Process in Urban
533 Areas with Fine Terrain and Lack of Pipe Network Data (No. 52079106); High Resolution
534 Numerical Simulation of Sediment Carrying Capacity Mechanism and Erosion Process of Over
535 surface Flow on Whole Sand Slope (No. 52009104); Chinesisch-Deutsches Mobilit
536 atprogramm: High – Resolution Numerical Simulating and Predicting Methods for Urban
537 Floods (No. M-0427).

538 **Competing Interests**

539 The authors declare that there is no known competing financial interests or personal
540 relationships that could have appeared to influence the manuscript entitled “A rapid forecast
541 method for the process of flash flood based on hydrodynamic model and KNN algorithm”.

542 **Availability of data and materials**

543 The data and code that support the study are available from the corresponding author upon
544 reasonable request.

545 **References**

- 546 Ji ZH, Wu XH (2018) A Review about the risk assessment of torrential flood disaster. Disaster
547 Science, 33(01):162-167+174.480. <https://doi.org/10.3969/j.issn.1000-811X.2018.01.029>
- 548 He BS, Ma MH, Li Q et al (2017) Summary and analysis of my country's mountain flood
549 disaster prevention in 2016. China Flood Control and Drought Relief (01): 95-98.
550 <https://doi.org/10.16867/j.cnki.cfdm.2017.01.016>
- 551 Montanari A, Young G, Savenije HHG et al (2013) “Panta Rhei—Everything Flows”: Change
552 in hydrology and society—The IAHS Scientific Decade 2013–2022. Hydrological Sciences
553 Journal 58(6): 1256-1275. <https://doi.org/10.1080/02626667.2013.809088>
- 554 Berardi L, Laucelli D, Simeone V et al (2013) Simulating floods in ephemeral streams in
555 Southern Italy by full-2D hydraulic models. International Journal of River Basin Management
556 11(1):1-17. <https://doi.org/10.1080/15715124.2012.746975>
- 557 Zhang DW, Quan J, Ma JM et al (2018) Two-dimensional numerical simulation of watershed
558 surface runoff based on Godunov format. Journal of Hydraulic Engineering 49(07): 787-
559 794+802. <https://doi.org/10.13243/j.cnki.slxb.20180388>
- 560 Hou JM, Wang JH, Tong Y et al (2021) High-efficiency and high-precision flood process
561 simulation based on non-uniform grid. Engineering Science and Technology 53(04): 53-62.
562 <https://doi.org/10.15961/j.jsuese.202000760>
- 563 Hu P, Han JJ, Lei YL (2019) Coupled modeling of sediment-laden flows based on local-time-
564 step approach. Journal of Zhejiang University (Engineering Science Edition) 53(04): 743-752.
565 <https://doi.org/10.3785/j.issn.1008-973X.2019.04.015>

566 Hu P, Lei YL, Han JJ et al (2019) Computationally efficient modeling of hydro-sediment-
 567 morphodynamic processes using a hybrid local time step/global maximum time step. *Advances*
 568 *in Water Resources* 127: 26-38. <https://doi.org/10.1016/j.advwatres.2019.03.006>.

569 Xu D, Xu B, Payet D et al (2016) Numerical simulation of shallow water motion based on
 570 parallel computation using GPU. *Chinese Journal of Computational Mechanics* 33(01): 114-
 571 121. <https://doi.org/10.7511/jslx201601018>

572 Hou JM, Kang YD, Hu CH et al (2020) A GPU-based numerical model coupling
 573 hydrodynamical and morphological processes. *International Journal of Sediment Research*
 574 35(4): 386-394. <https://doi.org/10.1016/j.ijsrc.2020.02.005>

575 Kao NSC, Sheu TWH (2018) Development of a finite element flow solver for solving three-
 576 dimensional incompressible Navier-stokes solutions on multiple GPU cards. *Computers &*
 577 *Fluids* 167:285–291. <https://doi.org/10.1016/j.compfluid.2018.03.033>

578 Mekanik F, Imteaz MA, Shirley GT et al (2013) Multiple regression and Artificial Neural
 579 Network for long-term rainfall forecasting using large scale climate modes. *Journal of*
 580 *Hydrology* 503(503):11-21. <https://doi.org/10.1016/j.jhydrol.2013.08.035>

581 Shin HC, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided
 582 detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions*
 583 *on Medical Imaging* 35(5): 1285-1298. <https://doi.org/10.1109/TMI.2016.2528162>

584 Li Y, Tian S, Huang YJ et al (2021) Driverless artificial intelligence framework for the
 585 identification of malignant pleural effusion. *Translational Oncology* 14(1).
 586 <https://doi.org/10.1016/J.TRANON.2020.100896>.

587 Briguglio W, Moghaddam P, Yousef WA et al (2021) Machine learning in precision medicine
 588 to preserve privacy via encryption. Pattern Recognition Letters 15: 148-154.
 589 <https://doi.org/10.1016/J.PATREC.2021.07.004>.

590 Ma L, Crawford MM, Tian JW (2010) Local Manifold Learning-Based k-Nearest-Neighbor for
 591 Hyperspectral Image Classification. Ieee Transactions on Geoscience and Remote Sensing
 592 48(11): 4099-109. <https://doi.org/10.1109/tgrs.2010.2055876>

593 Uddin S, Haque I, Lu HH et al (2022) Comparative performance analysis of K-nearest
 594 neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports
 595 12(1). <https://doi.org/10.1038/s41598-022-10358-x>

596 Fouad IA, Labib FEM (2022) Role of Deep Learning in Improving the Performance of Driver
 597 Fatigue Alert System. Traitement Du Signal 39(2): 577-88. <https://doi.org/10.18280/ts.390219>

598 Han T, Li Z, Liu K et al (2015) Research on real-time correction method of flood forecasting in
 599 small mountain watershed. Journal of Hohai University Natural Sciences 43(3): 208-14.
 600 <https://doi.org/10.3876/j.issn.1000-1980.2015.03.004>

601 Liu K, Yao C, Chen J et al (2017) Comparison of three updating models for real time forecasting:
 602 a case study of flood forecasting at the middle reaches of the Huai River in East China.
 603 Stochastic environmental research and risk assessment 31(6): 1471-1484.
 604 <https://doi.org/10.1007/s00477-016-1267-x>

605 Naghibi SA, Vafakhah M, Hashemi H et al (2020) Water Resources Management Through
 606 Flood Spreading Project Suitability Mapping Using Frequency Ratio, k-nearest Neighbours,
 607 and Random Forest Algorithms. Natural resources research 29(3): 1915-1933.

608 <https://doi.org/10.1007/s11053-019-09530-4>

609 Shahabi H, Shirzadi A, Ghaderi K et al (2020) Flood Detection and Susceptibility Mapping Using
610 Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of
611 Bagging Ensemble Based on K-Nearest Neighbor Classifier. Remote Sensing 12(2).
612 <https://doi.org/10.3390/rs12020266>

613 Hou J M, Zhou N, Chen G Z et al (2021) Rapid forecasting of urban flood inundation using
614 multiple machine learning models. Natural Hazards 108(2): 2335-2356.
615 <https://doi.org/10.1007/s11069-021-04782-x>

616 Khatri S, Kokane P, Kumar V et al (2022) Prediction of waterlogged zones under heavy rainfall
617 conditions using machine learning and GIS tools: a case study of Mumbai. Geojournal 1-15.
618 <https://doi.org/10.1007/s10708-022-10731-3>

619 Hubbard ME (1999) Multidimensional slope limiters for MUSCL-type finite volume schemes
620 on unstructured grids. Journal of Computational Physics 155(1): 54-74.
621 <https://doi.org/10.1006/jcph.1999.6329>

622 Hou JM, Liang QH, Zhang HB et al (2015) An efficient unstructured MUSCL scheme for
623 solving the 2D shallow water equations. Environmental Modelling & Software 66: 131-152.
624 <https://doi.org/10.1016/j.envsoft.2014.12.007>

625 Sivakumar P, Hyams DG, Taylor LK et al (2009) A primitive-variable Riemann method for
626 solution of the shallow water equations with wetting and drying. Journal of Computational
627 Physics 228(19): 7452-7472. <https://doi.org/10.1016/j.jcp.2009.07.002>

628 Hou JM, Liang QH, Simons F et al (2013) A stable 2D unstructured shallow flow model for

629 simulations of wetting and drying over rough terrains. Computers & Fluids 82:132-147
 630 <https://doi.org/10.1016/j.compfluid.2013.04.015>

631 Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE
 632 transactions on systems, man, and cybernetics (4): 580-585. 549
 633 <https://doi.org/10.1109/TSMC.1985.6313426> 550

634 Vialletto G, Noro M (2019) Enhancement of a Short-Term Forecasting Method Based on
 635 Clustering and KNN: Application to an Industrial Facility Powered by a Cogenerator. Energies
 636 12(23): 4407 <https://doi.org/10.3390/en12234407>

637 Liu K, Li Z, Yao C et al (2016) Coupling the k-nearest neighbor procedure with the Kalman
 638 filter for real-time updating of the hydraulic model in flood forecasting. International Journal
 639 of Sediment Research 31(2): 149-158 <https://doi.org/10.1016/j.ijsrc.2016.02.002>

640 Liu CJ, Wen L, Zhou J et al (2019) Comparison and analysis of flood model and hydrodynamic
 641 method calculation of rainstorm mountain in small watershed. Journal of China Institute of
 642 Water Resources and Hydropower Research 17(04): 262-270+278.
 643 <https://doi.org/10.13244/j.cnki.jiwhr.2019.04.003>

644 Ma XH, Wang WC, Zhu CS et al (2015) Experimental study for Manning coefficient in different
 645 densities of submerged flexible vegetation. Hydrodynamics Research and Progress Series A
 646 (01): 106-111. <https://doi.org/10.16076/j.cnki.cjhd.2015.01.015>.

647 Li SH, Sifton V, Lundgren J et al (2021) Extreme wind climate of the arabian peninsula
 648 characterized by using WRF simulation. Weather and Climate Extremes 33
 649 <https://doi.org/10.1016/j.wace.2021.100330>

650 Liu J, Qiu QT, Li CZ et al (2020) Advances of precipitation nowcasting and its application in
651 hydrological forecasting. Advances in Water Science 31(01): 129-142
652 <https://doi.org/10.14042/j.cnki.32.1309.2020.01.014>
653 Shao YM (2014) The new generation of rainstorm intensity formula in Chinese cities. China
654 Construction Industry Press.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Highlights.docx](#)