atmosphere

Article

Flood Forecasting through Spatiotemporal Rainfall in Hilly Watersheds

Yuanyuan Liu [1,2], Yesen Liu [1,2,*], Yang Liu [3], Zhengfeng Liu [3,4], Weitao Yang [5] and Kuang Li [1,2]

[1] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China; liuyy@iwhr.com (Y.L.); likuang@iwhr.com (K.L.)
[2] Key Laboratory of Water Safety for Beijing-Tianjin-Hebei Region of Ministry of Water Resources, Beijing 100038, China
[3] MWR General Institute of Water Resources and Hydropower Planning and Design, Beijing 100120, China; hehome415@163.com (Y.L.); liuzhengfeng@giwp.org.cn (Z.L.)
[4] Fujian Water Conservancy and Hydropower Survey and Design Institute, Fuzhou 350001, China
[5] Guangxi Water & Power Design Institute Co., Ltd., Nanning 530023, China; ywtcxp2008@163.com
* Correspondence: liuys@iwhr.com

Abstract: Flood prediction in hilly regions, characterized by rapid flow rates and high destructive potential, remains a significant challenge. This study addresses this problem by introducing a novel machine learning-based approach to enhance flood forecast accuracy and lead time in small watersheds within hilly terrain. The study area encompasses small watersheds of approximately 600 km^2. The proposed method analyzes spatiotemporal characteristics in rainfall dynamics to identify historical rainfall–flood events that closely resemble current patterns, effectively "learning from the past to predict the present". The approach demonstrates notable precision, with an average error of 8.33% for peak flow prediction, 14.27% for total volume prediction, and a lead time error of just 1 h for peak occurrence. These results meet the stringent accuracy requirements for flood forecasting, offering a targeted and effective solution for flood forecasting in challenging hilly terrains. This innovative methodology deviates from conventional techniques by adopting a holistic view of rainfall trends, representing a significant advancement in addressing the complexities of flood prediction in these regions.

Keywords: artificial intelligence; manifold learning; spatial and temporal characteristics of rainfall; flood risk management; flood forecasting; LSTM neural network

check for updates

Citation: Liu, Y.; Liu, Y.; Liu, Y.; Liu, Z.; Yang, W.; Li, K. Flood Forecasting through Spatiotemporal Rainfall in Hilly Watersheds. *Atmosphere* **2024**, *15*, 820. https://doi.org/10.3390/atmos15070820

Academic Editor: Tin Lukić

Received: 3 June 2024
Revised: 26 June 2024
Accepted: 4 July 2024
Published: 8 July 2024

1. Introduction

In recent years, with global climate change, frequent occurrences of extreme rainfall have triggered natural disasters such as basin flooding, urban waterlogging, and mountainous flash floods. From 28 July to 1 August 2023, the Haihe River Basin in China experienced a historically rare extreme rainstorm process, with a cumulative surface rainfall of 155.3 mm in the basin and a total precipitation of 49.4 billion m^3 [1]. On 20 July 2021, Zhengzhou City in Henan Province, China, experienced a maximum hourly rainfall of 201.9 mm, which set a new record for the maximum hourly rainfall intensity for a city in China at that time [2]. With the rapid development of socio-economic conditions, the losses caused by flood disasters are also exponentially magnified. Flood disasters have become an undeniable issue, emerging as a major natural disaster currently affecting social development and operations, and they are increasingly prevalent.

Before the arrival of floods, accurately predicting flood occurrences, preemptively assessing flood risks, and timely evacuation to prevent casualties are crucial aspects of flood disaster risk management. Traditional flood forecasting methods primarily involve watershed hydrological models, which, based on the physical characteristics of the underlying surface, approximate the watershed's water cycle process by incorporating infiltration curves, unit hydrographs, evapotranspiration formulas, and other elements. These models include the Sacramento model in the United States [3], the Tank model in Japan [4], China's Xin'anjiang model [5], and so on. In recent years, with technological advancements and algorithm improvements, the application of distributed hydrological models in flood forecasting has become increasingly widespread. The work of Abdelmounim [6] concerns the distributed hydrological modeling of the Azzaba catchment area in Haut-Sebou, "Morocco", considering the chronological sequence of phenomena and the influence of the climatic and physical–hydrogeological parameters of the basin. Dawen Yang [7] applied a GBHM (geomorphology-based distributed hydrological model) to the Chao Phraya Basin in Thailand. Khan [8] developed a distributed hydrological model of the Teesta River Basin using SWAT (Soil Water Assessment Tool) to assess the potential changes to the water balance. Saavedra [9] used a distributed hydrological model to simulate hydrological processes in the Agatsuma River Basin at hourly time steps. Guo [10] proposed a DEM-based distributed hydrological model to simulate runoff processes throughout a watershed. Reed [11] used simulation results for 10 basins in Oklahoma and Arkansas, USA, to improve distributed hydrologic model accuracy in small, interior basins, when forced by operational quality radar-based precipitation data. Bashirgonbad [12] used the daily precipitation data of 144 climatology stations in Iran to evaluate the seasonal and monthly pattern of flood-causing precipitation considering seasonal and monthly distribution.

These methods have achieved significant success and application in engineering planning, design, and flood forecasting and scheduling. However, traditional hydrological models generalize complex surface water processes, resulting in high prediction accuracy but with numerous mathematical model parameters and inherent uncertainties. Additionally, flash floods in hilly areas demand timely forecasts and warnings, imposing higher requirements on the accuracy and timeliness of flood predictions.

In recent years, with the flourishing development of the water conservancy industry and the dense network of monitoring stations, a significant amount of historical monitoring data has been accumulated. People have gained a deeper understanding of the patterns of rainfall and floods. In response to the demands of this new situation, effectively utilizing and mining massive historical hydrological data to further enhance the accuracy and timeliness of flood forecasting is a pressing issue that needs to be addressed.

With the continuous enrichment of historical hydrological data and the flourishing development of data mining and artificial intelligence technologies, research on data-driven machine learning techniques for flood forecasting has become a hot topic [13]. Hitokoto [14] used ANN models for the Abashiri River catchment, and river stage prediction up to 6 h showed very good accuracy. Luppichini [15] employed a Long Short-Term Memory (LSTM) model in the Arno River in Italy, for flood warning forecasts, and explored the reliability of this model. Do Hoai [16] presented an empirical–statistical downscaling method for precipitation prediction which used a feed-forward multilayer perceptron neural network for the Thu Bon River Basin, located in Central Vietnam. Akbari [17] proposed LSTM and the precipitation estimation from remotely sensed information using artificial neural networks for short-term quantitative precipitation forecasting. Amrul [18] proposed a support vector machine regression model to forecast flood water levels in the downstream area for different lead times in Kelantan River in Malaysia. Yuxuan Luo [19] proposed a Spatiotemporal Hetero Graph-based Long Short-Term Memory (SHG-LSTM) model for multi-step-ahead flood forecasting, and the SHG-LSTM model outperformed the LSTM and S-GCN models, with an average reduction in the volume error (VE) of 6.5% and 11.1%.

These models are based on AI algorithms that utilize rainfall and flood data to forecast hydrologic processes at the basin outlet. However, they rely on single-station or surface-averaged rainfall processes, as well as flooding processes in the pretemporal sequence at the outlet, as inputs. They also have a short flood prediction period.

Due to the complexity of spatial and temporal changes in rainfall in hilly areas, flood forecasting in hilly areas is also the focus and difficulty of flood forecasting. Complex mountainous regions remain a challenging task even for modern raingauge networks [20].

However, for watershed floods, especially in hilly areas with small basins, they are directly related to factors such as the precipitation and the spatial–temporal variations in rainfall centers in the watershed. Without considering human factors, under similar conditions of precipitation and spatial–temporal changes, the flooding processes at the surface also exhibit a certain degree of similarity. Based on this principle, this paper introduces the manifold learning algorithm from machine learning into the recognition of storm features and flood forecasting, proposing a flood prediction method based on the manifold learning algorithm. Manifold learning is a practical algorithm in the field of machine learning, successfully applied in feature classification, extraction, and identification. This method extracts spatiotemporal features from historical storm and flood processes to form a spatiotemporal feature sample library. When forecasting rainfall occurs, by identifying the spatiotemporal features of the current forecasted rainfall trend, the method rapidly matches the most similar historical storm processes from the spatiotemporal feature sample library, drawing analogies from the past to predict the entire flood corresponding to the current forecasted rainfall trend. Using the Zhongping small basin in the Guangxi Zhuang Autonomous Region as an example, the results indicate that this method can effectively predict peak flow, flood volume, peak occurrence time, and flood shape, extending the flood lead time. The forecasting results achieve first-class accuracy and meet the requirements of precision and timeliness for flood forecasting.

For watershed flooding, especially in small watersheds in hilly areas, there is a direct correlation with the amount of rainfall in the region and the spatial and temporal variability in storm centers in the watersheds. Excluding anthropogenic factors, there is some similarity in the flooding process in the subsurface under similar conditions of rainfall and spatial and temporal variability. Based on this principle, this paper introduces the manifold learning algorithm in machine learning into storm feature recognition and flood forecasting, proposing a flood forecasting method. The manifold learning algorithm is a highly practical tool in the field of machine learning, successfully applied in feature classification, extraction, and recognition [21].

The method performs a spatiotemporal feature extraction of historical storm flooding processes to form a library of spatiotemporal feature samples. When rainfall is forecasted to occur, the spatial and temporal features of the current forecast rainfall trend are recognized. This allows for the quick matching of the most similar historical rainstorm processes from a sample library of spatial and temporal features, with historical flood processes identified as the result of flood forecasting under the current forecast rainfall trend.

In this paper, the Zhongping sub-basin of the Guangxi Zhuang Autonomous Region of China is taken as an example. The results show that the method can effectively predict the flood peak flow, flood volume, peak present time, and the shape of flood change, extending the flood forecast period. The prediction result reaches the accuracy of a Class A forecast, meeting the requirements of flood forecast accuracy and timeliness. Unlike traditional flood forecasting methods, this method can predict not only the total rainfall, maximum rainfall intensity, and other rainfall characteristics but also the entire flood process, including flood volume, peak flow, peak time, and flood shape changes. This can meet the accuracy and timeliness requirements for flood forecasting in small mountainous basins. This study effectively extends the flood forecast period, filling the gap in the timeliness of AI-based flood forecasting technology.

2. Data and Methods

2.1. Data

The Zhongping River is located in the northeast of the Guangxi Zhuang Autonomous Region, flowing from the south to the north, with a fan-shaped watershed. The area of the basin is 596 km^2, and the average slope drop is 5.04‰, with a river length of 63 km. For mountainous rivers, the river’s slope is steeper, causing flooding to rise and fall quickly. As shown in Figure 1, there are three rainfall stations: Wangtian, Dachang, and Liuxiang, along with the Zhongping hydrological station. In this paper, data from these three rainfall stations and one hydrological station for the years 2002 to 2023 are used to build a flood forecasting model.

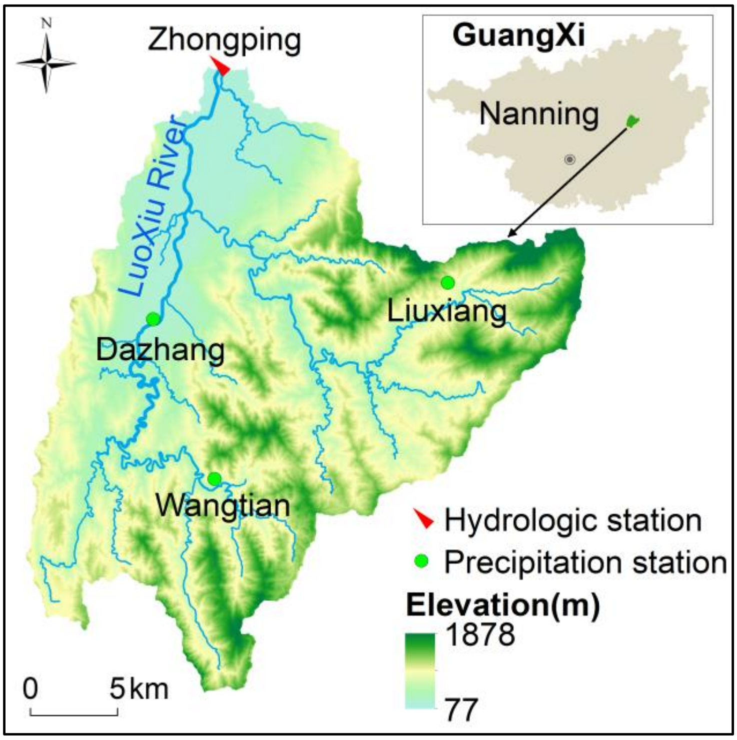


Figure 1. Research scope.

First, data cleaning and filtering are conducted. This study considers periods of continuous rainfall less than 0.1 mm for more than 2 h as indicating no effective rainfall. The specific process is as follows: If the 5 min rainfall at a single station exceeds 10 mm and is an isolated data point, the record is considered unreasonable. If no rainfall is detected within a 5 km × 5 km range from a given rainfall station, but the 5 min rainfall at that station exceeds 10 mm, the recorded rainfall at that station is deemed unreasonable. In such cases, the unreasonable records are replaced with interpolated values using rainfall data collected from other stations within the 5 km × 5 km range. The interpolation method used in this study is the Inverse Distance Squared Weighting (IDW) method. After cleaning the rainfall data, 115 heavy rainfall events were identified from 2002 to 2023, forming a rainfall–flood event sample library. Subsequently, the heavy rainfall–flood processes were digitized and structured, constructing a high-dimensional array for these events in both temporal and spatial dimensions. The algorithm was then applied to analyze the sample library of high-dimensional arrays, with details as follows.

2.2. Methods

Using the manifold learning algorithm, specifically the Locally Linear Embedding (LLE) algorithm, and the dynamic clustering algorithm, K-Means (KM), the dynamic spatiotemporal distribution characteristics of various rainfall types are obtained [22]. Comparative learning and feature recognition are then applied to the corresponding flood process characteristics. This allows for the identification of historical heavy rain–flood processes that are the most similar to the current rainfall dynamic trends. Consequently, it becomes possible to predict future flood processes under the current rainfall trends. The specific technical workflow is illustrated in Figure 2.

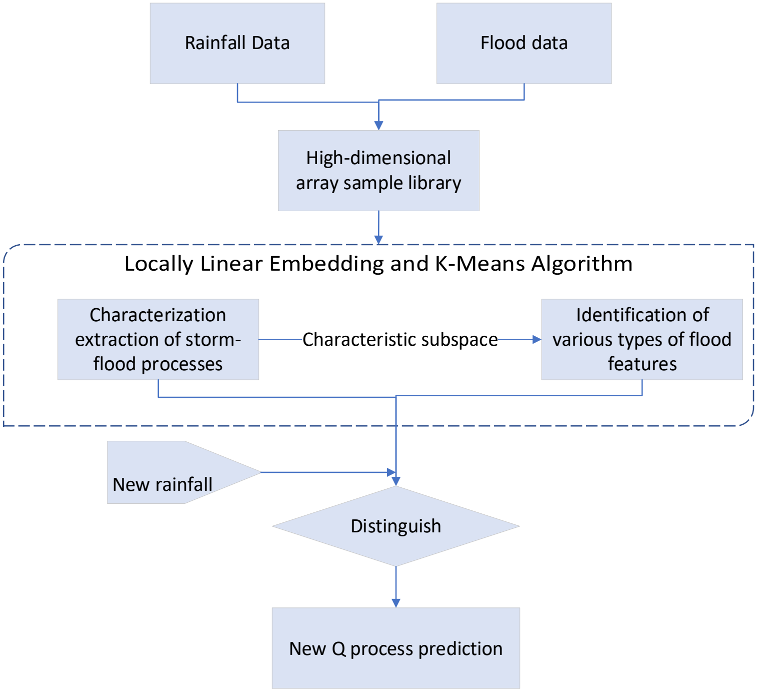

Rainfall Data
Flood data
High-dimensional array sample library
Locally Linear Embedding and K-Means Algorithm
Characterization extraction of storm-flood processes
Characteristic subspace
Identification of various types of flood features
New rainfall
Distinguish
New Q process prediction

Figure 2. Method and flow chart.

2.2.1. Construction of Spatiotemporal Distribution Dynamic Feature Matrix for Heavy Rainfall and Floods

Due to significant differences in rainfall and flood volume among different events, this study aims to facilitate the comparison of the dynamic features of rainfall and floods for various rainfall processes. For this purpose, matrices representing the temporal and spatial proportions of rainfall and floods are constructed for different durations of rainfall events. These matrices, describing the spatiotemporal distribution characteristics of rainfall and the morphological features of flood processes using the proportions of rainfall and flood volumes, enable a mathematical description of the spatiotemporal dynamic development features of multiple rainfall–flood events. A sample set Ω for heavy rain–flood processes is established, as shown in Equations (1) and (2).

$$\Omega = \{P_1, P_j, \ldots P_N\} \tag{1}$$

where Ω represents the historical heavy rainfall sample set, including N instances of heavy rain, where $j = 1, 2, 3\ldots N$, and N is the number of rainfall events.

$$P_j = \begin{bmatrix} r^j_{11} & r^j_{1t}\cdots & r^j_{1m} \\ r^j_{21} & r^j_{2t}\cdots & r^j_{2m} \\ \vdots & r^j_{it} & \vdots \\ r^j_{s1} & r^j_{s2}\cdots & r^j_{sm} \\ q^j_1 & q^j_t\cdots & q^j_m \end{bmatrix} \tag{2}$$

P_j represents the j-th storm–flood percentage matrix, where r^j_{it} signifies the percentage of rainfall at time t at the i-th rainfall station in the j-th storm–flood process concerning the total rainfall at all stations at that specific time. q^j_t denotes the flow at time t of the watershed outlet cross-section during the j-th storm–flood process as a percentage of the total for the entire flood process. Here, $i = 1, 2, 3\ldots s$, where s is the number of rainfall stations, and $t = 1, 2, 3\ldots m$, where m is the number of time slots.

$$r^j_{it} = R^j_{it}/\sum_{t=1}^{m} R^j_{it} \tag{3}$$

$$q_t^j = Q_t^j / \sum_{t=1}^{m} Q_t^j \tag{4}$$

In Equation (3), R_{it}^j represents the rainfall amount at the i-th rainfall station at time t during the j-th rainfall event. In Equation (4), Q_t^j represents the flow at the watershed outlet cross-section at time t during the j-th rainfall event.

2.2.2. Dimensionality Reduction Analysis Based on LLE Algorithm

The high-dimensional data sample library describing the spatial and temporal distribution characteristics of rainstorms represents a nonlinear high-dimensional data space. Directly analyzing it involves a certain degree of uncertainty, requiring downsizing before analysis. Dimensionality reduction involves representing the original data with a smaller number of "effective" features, extracting the main features without diminishing the information contained in the original data.

Through dimensionality reduction, analysis efficiency can be significantly improved, enhancing the accuracy of the results [23]. In this study, the Locally Linear Embedding (LLE) algorithm is employed for dimensionality reduction analysis on this high-dimensional data. The LLE algorithm, proposed by S.T. Roweis [24] et al., is an unsupervised dimensionality reduction method for nonlinear data. It is a flow learning algorithm where local linearity reflects global nonlinearity, preserving the topology of the original data in the dimensionality-reduced data.

Since the LLE algorithm maintains the local linear characterization of both high- and low-dimensional spaces, the classification results in the low-dimensional space are also reasonable in the high-dimensional space. The LLE algorithm assumes that the data are linear in a small range around them, expressing each sample point linearly in terms of its neighboring data. This local linear relationship remains constant in both high- and low-dimensional spaces. The LLE algorithm consists of three main parts:

1. In the high-dimensional space, find the K nearest samples to sample x_i by using the Euclidean distance measure.
2. For each sample x_i, find the linear relationship of the K nearest neighbors in its neighborhood, and obtain the linear relationship weight coefficient W_i
3. Assuming that the linear relationship weight coefficients W_i remain constant in the K-neighborhood in both high- and low-dimensional spaces, reconstruct the sample data in low dimensions using the weight coefficients W_i, $x_i \in R^D \to y_i \in R^d$, $\mathrm{d} \ll \mathrm{D}$.

First, for N data points in the high-dimensional space $\{x_1, x_2, \cdots x_N\} \in R^D$, calculate the Euclidean distance of each sample point x_i from all other samples, and then select the K samples with the smallest distance $\{x_{i1}, x_{i2}, \cdots x_{ik}\}$.

Each x_i be expressed linearly in terms of the nearest K samples $\{x_{i1}, x_{i2}, \cdots x_{ik}\}$

$$x_i \approx \overline{x}_i = \sum_{j=1}^{k} w_{ij} x_j \tag{5}$$

$$\sum_{j=1}^{k} w_{ij} = 1 \tag{6}$$

Using the mean square deviation as the loss function, the following can be obtained, as shown in Equation (7):

$$f(W) = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} w_{ij} x_j \right\|_2^2 \tag{7}$$

The weighting factor W was solved for the minimum value of Equation (7).

The LLE algorithm assumes that high-dimensional samples are mapped into the low-dimensional space, where the local linearity of the preserved samples in the high-dimensional space is maintained, and the weight coefficients are kept unchanged. This ensures that the points $\{x_1, x_2, \cdots x_N\} \in R^D$ in the high-dimensional space are mapped into the low-dimensional space as $\{y_1, y_2, \cdots y_N\} \in R^d (d \ll D)$.

2.2.3. Dynamic Cluster Analysis

The dimensionally reduced sample set $Y \in R^{d \times N}$ (where d is the low-dimensional spatial dimension of the projection, and N is the number of samples) is categorized into r subsets, with samples approximated within each subset and variations observed between subsets. Features specific to each class are extracted by determining the center of mass for each subset.

In this study, dynamic clustering methods are predominantly employed to classify samples' post-dimensionality reduction [25]. The fundamental concept of dynamic cluster analysis involves iteratively finding a partitioning scheme of r clusters. This minimizes the overall error when using the means of these r clusters to represent corresponding sample categories. The algorithm starts by randomly selecting r sample points as the initial clustering centers for r subsets. The distance between all samples and these initial centers is calculated. Samples are then assigned to the subset with the closest center. This process clusters all samples into subsets automatically based on distance, establishing the initial classification categories and subsets.

The mean of all samples in each subset is calculated to obtain new generation clustering centers. The distances of all samples from the new centers are computed, and the process is repeated iteratively. The p th and p + 1 th generation clustering centers are compared, and convergence is considered if the difference is within a specified range. This yields the final subsets and clustering centers for each subset.

Although this clustering method converges rapidly and provides improved results, its outcomes are significantly influenced by the selection of the initial clustering center. Consequently, after iterative convergence, this paper constantly compares and analyzes to assess the reasonability of the number of subsets and the initial subset center. Adjustments are made accordingly, repeating the iterative clustering process until a reasonable number of spatially distributed feature categories and clustering centers are determined. The calculation steps are as follows:

1. $\Phi = \{Y_1, Y_2, \ldots Y_N\}$ is the sample set analyzed, Y_i represents the mapping points in the low-dimensional space, M is the maximum number of iterations, r is the number of subsets initially divided, and C represents the r subsets $C = \{C_1, C_2, \cdots, C_r\}$. Initially, $C_j = \emptyset$, j = 1, 2, ... r.
2. Randomly select r samples from Φ as the initial r subsets of each center vector $Z_j^0 = \{z_1, z_2, \ldots\ldots z_r\}$ (0 is the initial value of the iteration number).
3. For n = 1, 2, ... N, calculate the distance $d_{ij} = \|Y_i - z_j\|_2^2$ between sample $Y_i (Y_i \in \Phi)$ and each clustering center $Z_j = \{z_1, z_2, \ldots\ldots z_r\}$. If $d_{ij} = min\{d_{ij}\} i = 1, 2, \cdots N$, then $Y_i \in C_j$. Update $C_j = C_j \cup Y_i$.
4. For j = 1, 2, ... r, recalculate the center vector $Z_j^1 = \frac{1}{C_j} \sum_{Y_{i \in C_j}} Y_i$ for all sample points in C_j.
5. Keep repeating the iteration; if $Z_j^{p+1} \neq Z_j^p$, j = 1, 2, ... r, go back to step 3, and repeat the iterative calculation. If $Z_j^{p+1} = Z_j^p$, j = 1, 2, ... r, the operation ends.
6. Output the subsets $C = \{C_1, C_2, \cdots, C_r\}$, the samples $y_1^{C_i}, y_2^{C_i}, \ldots\ldots y_o^{C_i}$ belonging to each subset, and the mean $Z_j^{p+1} = \{z_1, z_2, \ldots\ldots z_r\}$.

2.2.4. Reconstruction of Spatiotemporal Feature Spaces

The subsets $C = \{C_1, C_2, \cdots, C_r\}$ and the means Z_j^{p+1} of each subset obtained from the aforementioned clustering method do not represent the sought-after feature space but rather the feature space of the dataset after dimensionality reduction.

The LLE algorithm utilized in this paper assumes that the local linear relationship between the high-dimensional space and the low-dimensional space remains unchanged. In other words, the linear relationship between a sample x_i in the high-dimensional space and its neighboring samples is the same as the local linear relationship between its mapped point y_i in the low-dimensional space and its corresponding neighboring samples.

The samples that belong to the same subset in that space are also similar in the higher-dimensional space. Similarly, samples belonging to the same subset in the lower-dimensional space are classified into the same subset in the higher-dimensional space.

This implies that the samples in each subset $\mathrm{C} = \{C_1, C_2, \cdots, C_r\}$ in the lower-dimensional space also correspondingly belong to the same subset $\mathrm{B} = \{B_1, B_2, \cdots, B_r\}$ in the higher-dimensional space. The mean $\mathrm{S}_j = \frac{1}{B_j} \sum\limits_{x_i \in B_j} x_i \in R^D$ of each subset in the high-dimensional space serves as the center of clustering for each category in the high-dimensional space. In other words, it characterizes the dynamic spatiotemporal distribution of the samples belonging to that category.

2.2.5. Spatiotemporal Dynamic Feature Recognition and Distinguishing of Storm–Flood Events

For an upcoming rainstorm, the aforementioned algorithm is utilized to project into the spatiotemporal feature space. In adherence to the principle of the smallest distance in the feature space, samples with the smallest distance in the spatiotemporal feature subspace of historical rainstorms are identified. This process aims to identify historical rainstorms with the most similar spatiotemporal features to the current rainfall process. The flood process corresponding to this identified historical storm sample is considered the flood forecast result for the current forecast storm, as expressed in Equation (8):

$$\min(d_{2D}(Y_t; Y_i)) = \min(\|Y_t - Y_i\|_2) \tag{8}$$

Here, Y_t is the feature matrix of d samples to be recognized, and Y_i is the feature matrix of storm samples.

To further illustrate the comparison between the methodology presented in this paper and the traditional neural network approach, a flood forecasting model based on the Long Short-Term Memory (LSTM) neural network model is also constructed.

2.2.6. LSTM Neural Network Model

The Long Short-Term Memory (LSTM) neural network model is an enhanced version of the Recurrent Neural Network (RNN) model, developed by Hochreiter and Schmidhuber [26]. The enhancement introduced by LSTM to the RNN is primarily manifested in the addition of a hidden state C_t to the RNN hidden layer. Additionally, three gates—the forget gate, input gate, and output gate—are incorporated to address the issue of gradient vanishing or gradient explosion in RNN models [27], as illustrated in Figure 3.

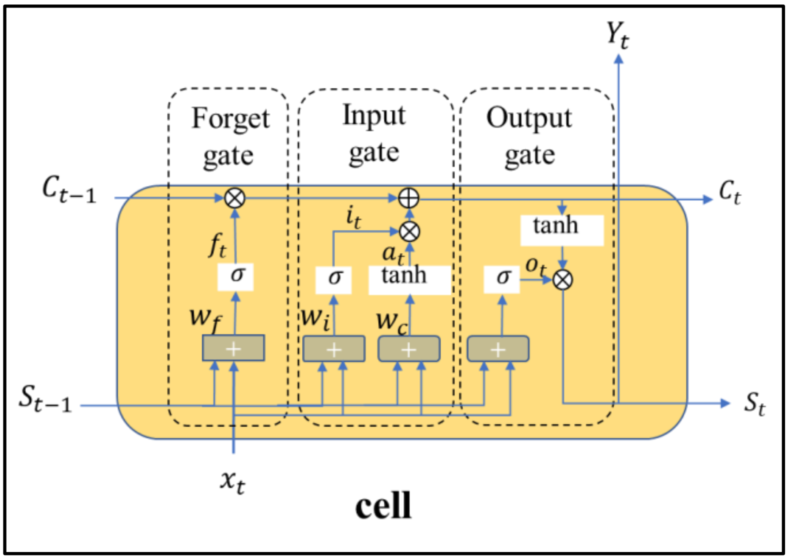


Figure 3. LSTM neural network model structure.

The output of the oblivion gate is f_t, as shown in Equation (9):

$$f_t = \sigma\left(W_f S_{t-1} + U_f x_t + b_f\right) \tag{9}$$

In Equation (9), the output of the forget gate f_t is determined by the sample x_t of this time series and the outputs of the previous time's hidden layer S_{t-1}, where σ represents the activation and sigmoid functions. As the output of the sigmoid function falls within the range [0, 1], the output of the forgetting gate f_t also lies within [0, 1]. This value indicates the probability of forgetting the hidden cell state from the previous layer. W_f and U_f are the matrix parameters of the model, and b_f is the linear bias parameter.

The output of the input gate is depicted in Equations (10) and (11), where the activation function in Equation (10) is a sigmoid function, and the activation function in Equation (11) is a tanh function.

$$i_t = \sigma(W_i S_{t-1} + U_i x_t + b_i) \tag{10}$$

$$a_t = \tanh(W_a S_{t-1} + U_a x_t + b_a) \tag{11}$$

W_i, U_i, W_a, and U_a are linear parameters, and b_i and b_a are linear bias parameters. Update the state of C_t from these two outputs

$$C_t = C_{t-1} \odot f_t + i_t \odot a_t \tag{12}$$

where $\odot$ is the Hadamard product.

The output of the output gate is shown in Equation (13), where the activation function is a sigmoid function:

$$o_t = \sigma(W_o S_{t-1} + U_o x_t + b_o) \tag{13}$$

Then, the output S_t of the implicit layer is obtained by the product of the output o_t of the output gate and C_t

$$S_t = o_t \odot tanh(C_t) \tag{14}$$

Then, the predicted output is shown in Equation (15), and the activation function is the sigmoid function.

$$\hat{y}_t = \sigma(VS_t + b_t) \tag{15}$$

The above $W_f, U_f, b_f, W_a, U_a, b_a, W_i, U_i, b_i, W_o, U_o, b_o, V, b_\mathrm{t}$ are parameters, and similar to the standard RNN algorithm, the LSTM iteratively updates all the parameters by gradient descent.

As depicted in Figure 4, the input factors for the flood forecasting model based on the LSTM algorithm encompass the rainfall data from each rainfall station in the initial three time series and the flow data from the outlet cross-section during the same time span. The model's output corresponds to the flow data for the subsequent time series, thereby categorizing it as an LSTM model featuring multi-factor input and a single output.

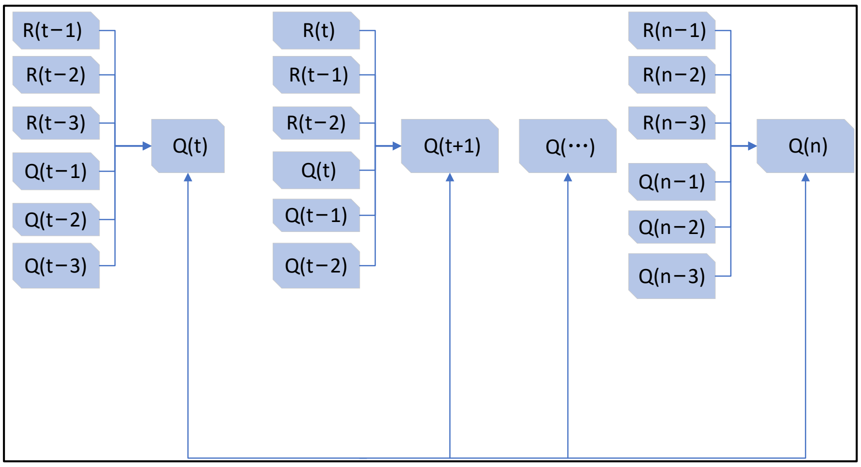


Figure 4. Input–output structure diagram.

By utilizing the rainfall and flow data from each rainfall station in the preceding time series, the model predicts the flow rate. Through continuous cycles of training, the objective is to achieve a comprehensive forecast of the entire rainfall process and, consequently, predict the flow rate at the outlet section.

2.2.7. Identification of Forecasted Floods

When identifying storm characteristics and matching flood processes, multiple instances of storm–flood events can be considered. To achieve optimal forecasting performance, this paper also establishes evaluation metrics for flood forecasting. The forecast results are assessed, and the best flood process is selected as the outcome of the flood forecast. The evaluation primarily focuses on peak flow intensity, forecast errors in flow rates at each time step during the flood process, and the morphological aspects of the flood process, comparing forecasted values with actual measurements. This encompasses the following:

1. Peak flow error ΔQ_m

$$\Delta Q_m = \frac{\left| Q_{mmeasure-} Q_{mforecast} \right|}{Q_{mmeasure}} \times 100\% \tag{16}$$

2. Peak timing error ΔT_m

$$\Delta T_m = \left| T_{mmeasure} - T_{mforecast} \right| \tag{17}$$

3. Root mean square error ($RMSE$) between flow rates at each time step

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} (y_i - f_i)^2} \tag{18}$$

f_i and y_i denote the predicted and measured results of the model, respectively.

4. The coefficient of determination (R^2) indicating the similarity between predicted and simulated data curves.

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y_i})^2} \tag{19}$$

In this study, the coefficient of determination, R^2, is employed to assess the degree of similarity between the predicted and observed curves. A higher R^2 value, approaching 1, indicates a better fit between these two curves, signifying closer alignment between the model's predictions and the actual observations. The formula for calculating R^2 is presented in Equation (19), where f_i and y_i denote the model's predicted results and observed measurements, respectively.

5. Comprehensive indicator

The identified multiple flood events have varying emphases. For instance, some recognition results exhibit good agreement in peak flow rates, while the shape of the flood process does not match well. On the other hand, some results demonstrate a good match in the flood process, but there are significant differences in peak flow rates. To provide a more objective evaluation, this paper defines a comprehensive indicator, Comindex, to assess the final recognition results. The specific formulation is presented in Equation (20).

$$Comindex = \frac{1}{RMSE} + R^2 \tag{20}$$

This indicator is inversely proportional to the RMSE and directly proportional to R^2. A larger value of the indicator indicates smaller errors in predicted flow rates at each time step and a closer resemblance of the predicted flood process to the actual flood process.

3. Results and Discussion

3.1. Result

Using the above algorithm, we illustrate the process with the example of the Zhongping small watershed. According to the criteria mentioned above, a total of 115 intense rainfall events were identified between 2002 and 2023, forming a sample library of rainfall–flood events. Out of these, 110 events were used as training samples, and an additional 5 rainfall events were randomly selected as samples for identification.

Firstly, the spatiotemporal characteristics of historical intense rainfall events were extracted, and cluster analysis was conducted based on these features. Building upon this, the current dynamic rainfall process was identified, recognizing the most similar historical intense rainfall event. The flood process corresponding to this identified intense rainfall event was extracted as the forecast result for the ongoing rainfall. Evaluation was carried out based on the peak flow error ΔQ_m, peak timing error ΔT_m, root mean square error between flow rates at each time step (*RMSE*), coefficient of determination for curve similarity (R^2), and a comprehensive indicator to assess the similarity between the identified historical flood process and the sample for identification.

When identifying the sample for identification, the first one-quarter duration, first one-third duration, and first one-half duration of the rainfall process were initially considered. These segments were projected onto the historical intense rainfall sample library for identification, finding corresponding historical intense rainfall samples. Subsequently, the identification was performed with the entire rainfall process to find the historical intense rainfall sample corresponding to the complete identification sample. The experiment showed that the results based on the first one-quarter duration differed from those based on the complete rainfall process, while the results based on the first one-third and first one-half durations were consistent with the results based on the complete process. In other words, using the proposed method in this paper, when the ongoing rainfall completes one-third of its total duration, it is possible to quickly predict the subsequent rainfall process, identify the most similar historical intense rainfall event, and achieve the early recognition of flood risk based on the identified rainfall and its corresponding flood process.

A map depicting the dynamic distribution of each storm and a comparison of the flooding process is illustrated in Figures 5–9. In the five figures, panel A displays the samples earmarked for identification, panel B showcases the historical samples identified from the sample library, and panel C exhibits the predicted flooding results from both this paper's algorithm and the LSTM model. The total duration of rainfall was T hours, and the specific storm identification results along with flood forecasting results are detailed in Tables 1 and 2.

Table 1. Comparison table for identifying storm results.

Serial Number	Group Name	Rainfall	Average Surface Rainfall (mm)	Errors (%)	Maximum Rainfall at Single Station (mm)	Errors (%)	Maximum Rainfall Intensity (mm)	Errors (%)
1	A1	23/05/2022	37	6.22	46	6.52	13	15.38
	B1	10/05/2022	39.3		43		11	
2	A2	23/05/2015	45	9.62	55	16.36	19.5	23.08
	B2	02/04/2014	40.67		46		15	
3	A3	11/06/2022	116	3.71	158	21.52	29	22.41
	B3	26/08/2019	120.3		124		35.5	
4	A4	05/06/2014	32.33	3.19	42	14.29	27.5	21.82
	B4	10/08/2011	31.3		36		21.5	
5	A5	20/06/2022	61.67	7.04	81	7.41	30	8.33
	B5	07/09/2015	57.33		75		27.5	
		Average error		5.95		13.22		18.21

Samples in A group are to be identified, samples in B group are identification results.

Table 2. Comparison table of flooding processes by feature recognition.

Serial Number	Group Name	Rainfall	Q (m^3/s)	Errors (%)	W ($\times 10^5$ m^3)	Errors (%)	Time of Flood Peaks (h)	DT (h)
1	A1	23/05/2022	40.2	23.63	458.85	12.22	18	1
	B1	10/05/2022	49.7		514.92		17	
2	A2	23/05/2015	72.7	9.22	653.94	18.11	24	1
	B2	02/04/2014	66		535.53		23	
3	A3	11/06/2022	1140	2.63	1998.12	25.07	19	2
	B3	26/08/2019	1170		2498.98		21	
4	A4	05/06/2014	60.9	1.64	414.32	3.36	14	0
	B4	10/08/2011	61.9		400.41		14	
5	A5	20/06/2022	110	4.55	928.64	12.59	8	1
	B5	07/09/2015	105		811.76		9	
		Average error		8.33		14.27		1

Samples in A group are to be identified, samples in B group are identification results.

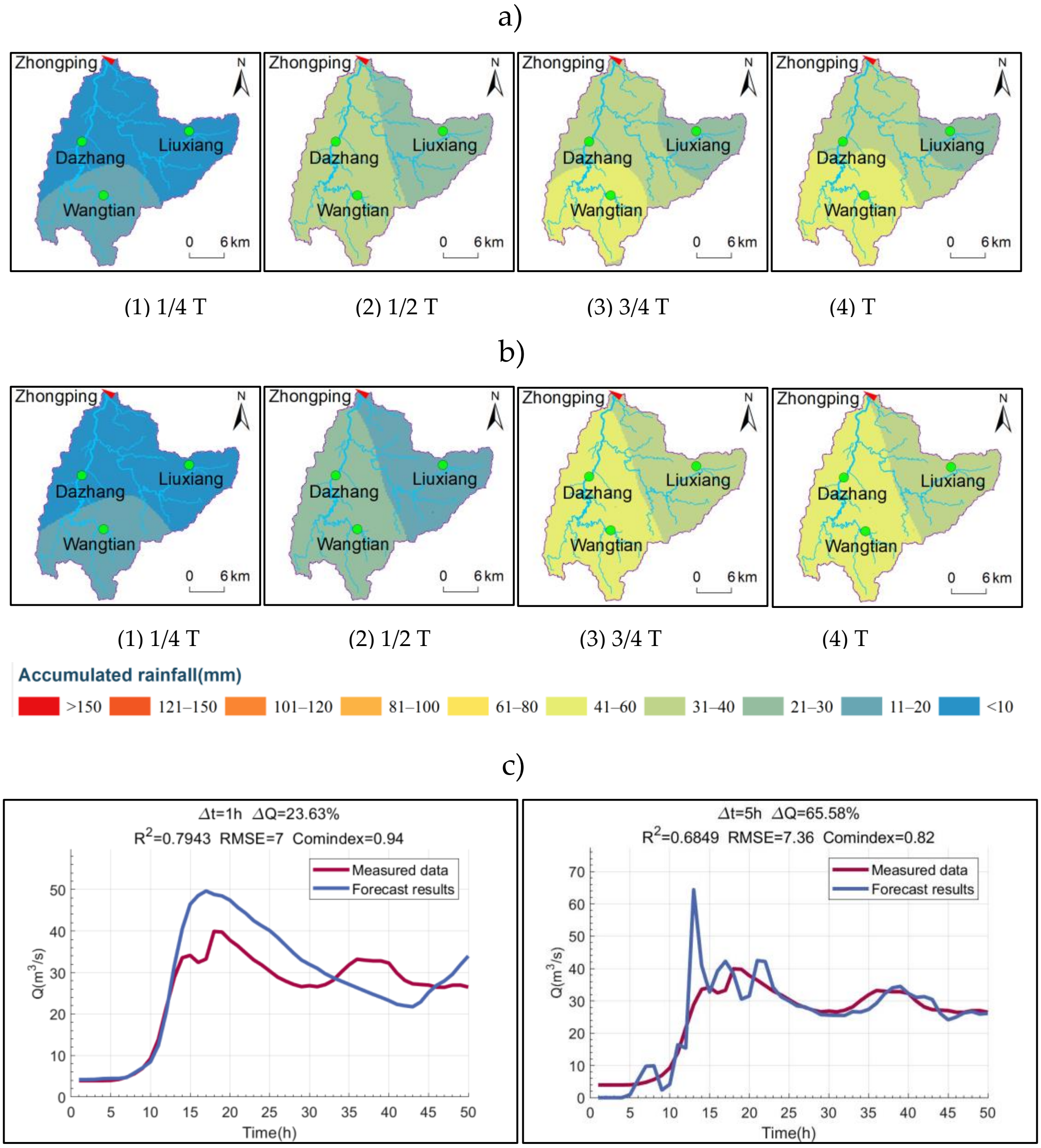
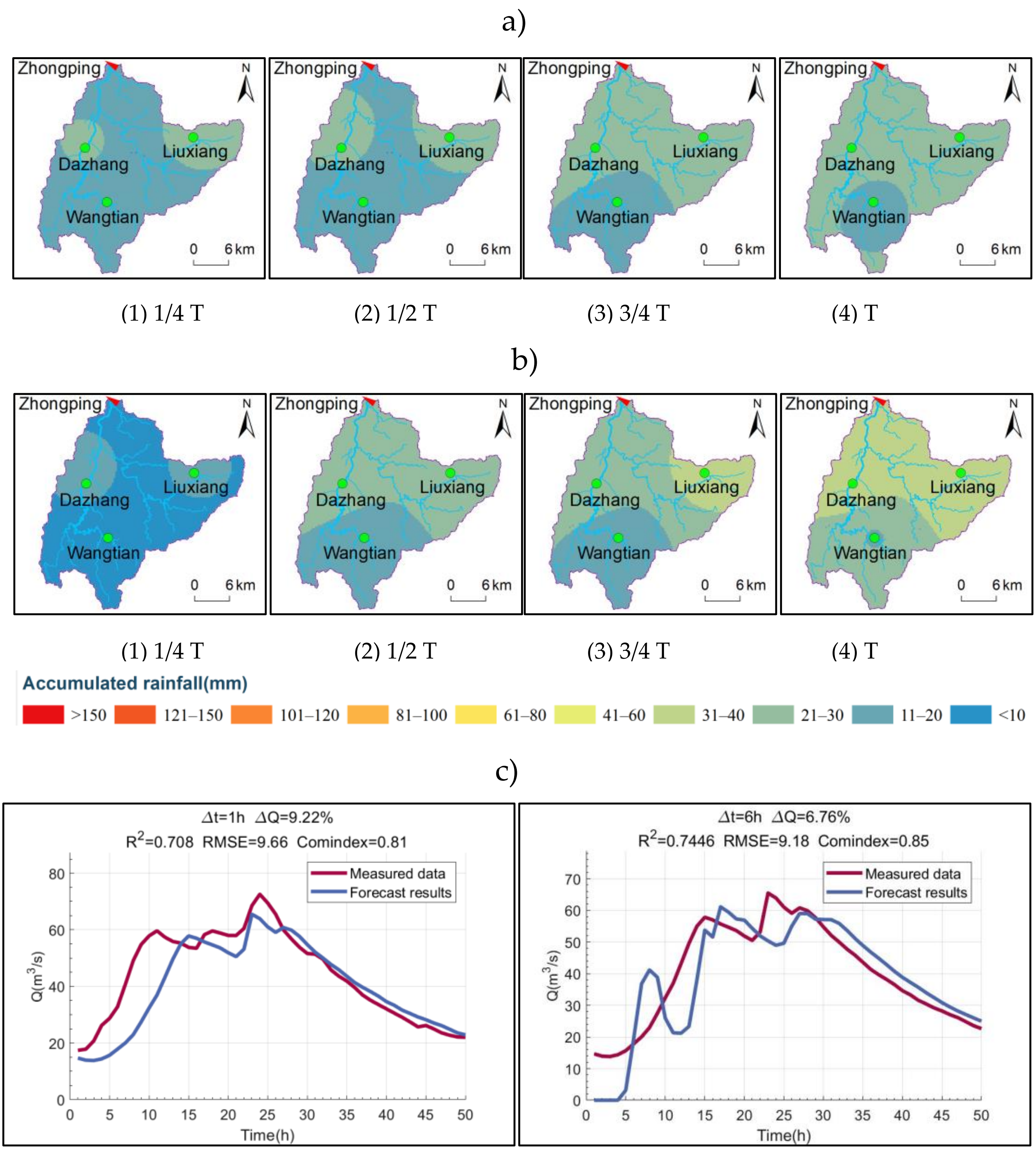


Figure 5. The first identification case. (**a**) Rainfall processes to be identified (23 May 2022); (**b**) identification results (10 May 2022); (**c**) flood forecast results.



(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

Figure 6. The second identification case. (**a**) Rainfall processes to be identified (23 May 2015); (**b**) identification results (2 April 2014); (**c**) flood forecast results.

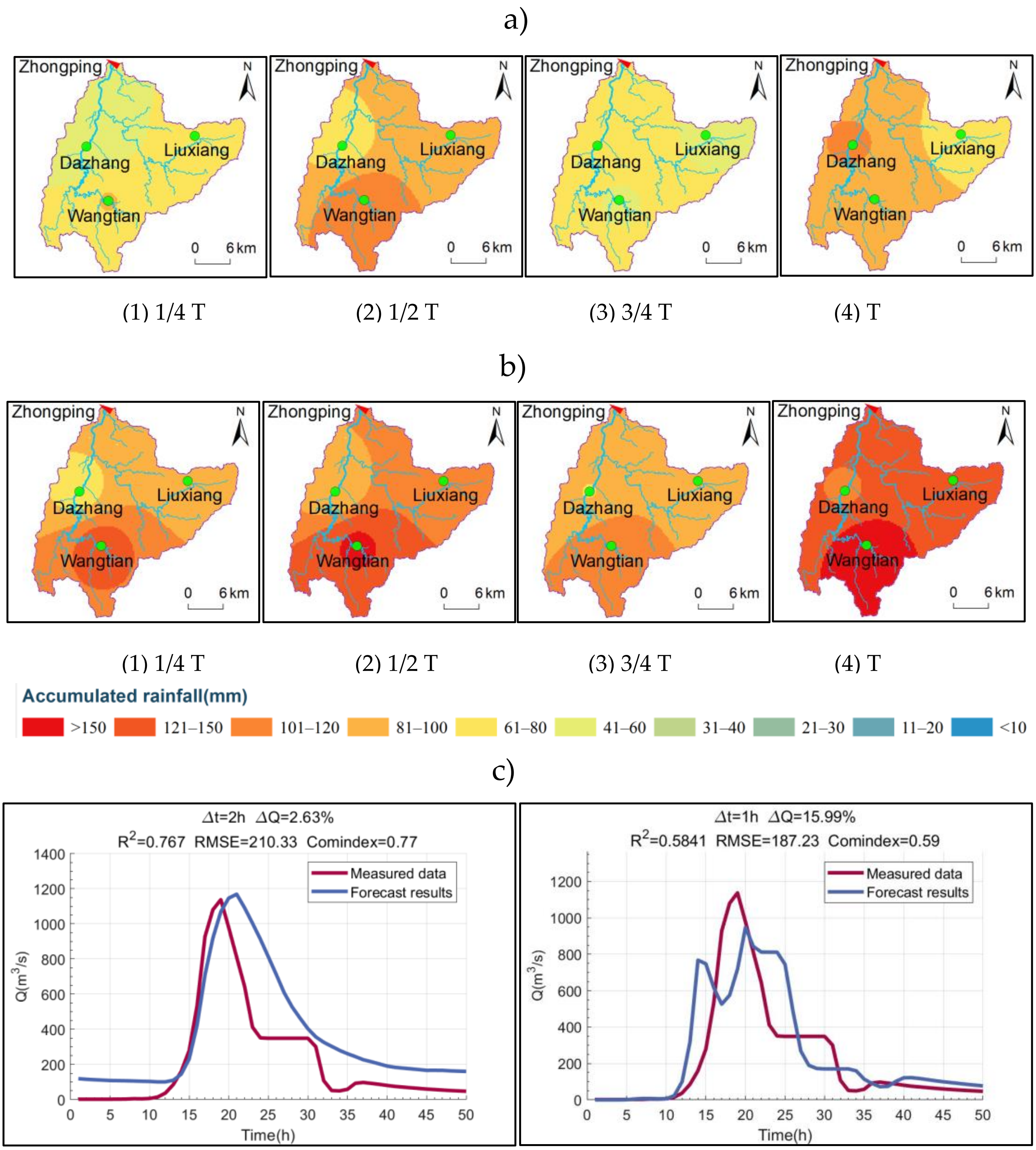
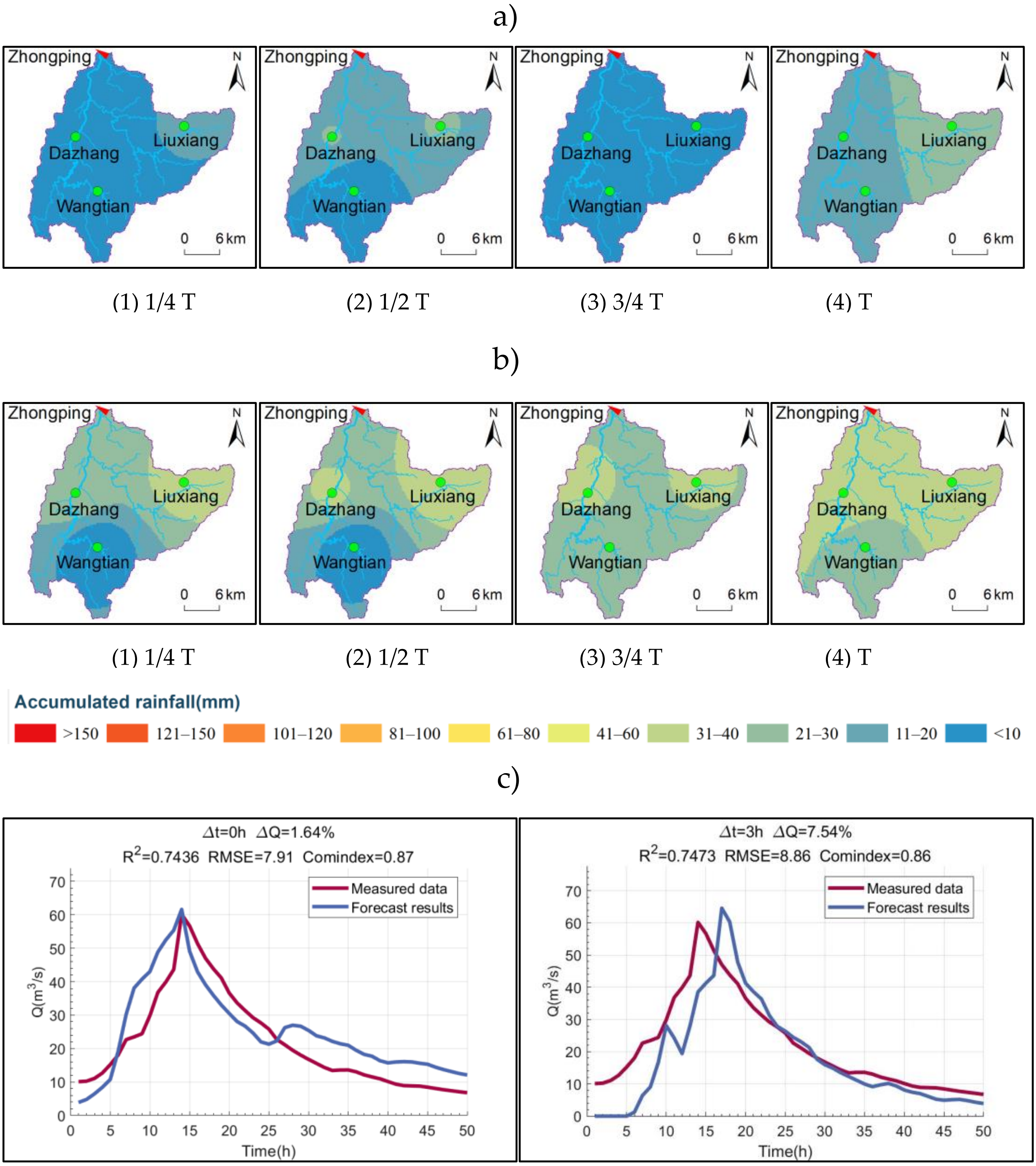


Figure 7. The second identification case. (**a**) Rainfall processes to be identified (11 June 2022); (**b**) identification results (26 August 2019); (**c**) flood forecast results.



a)

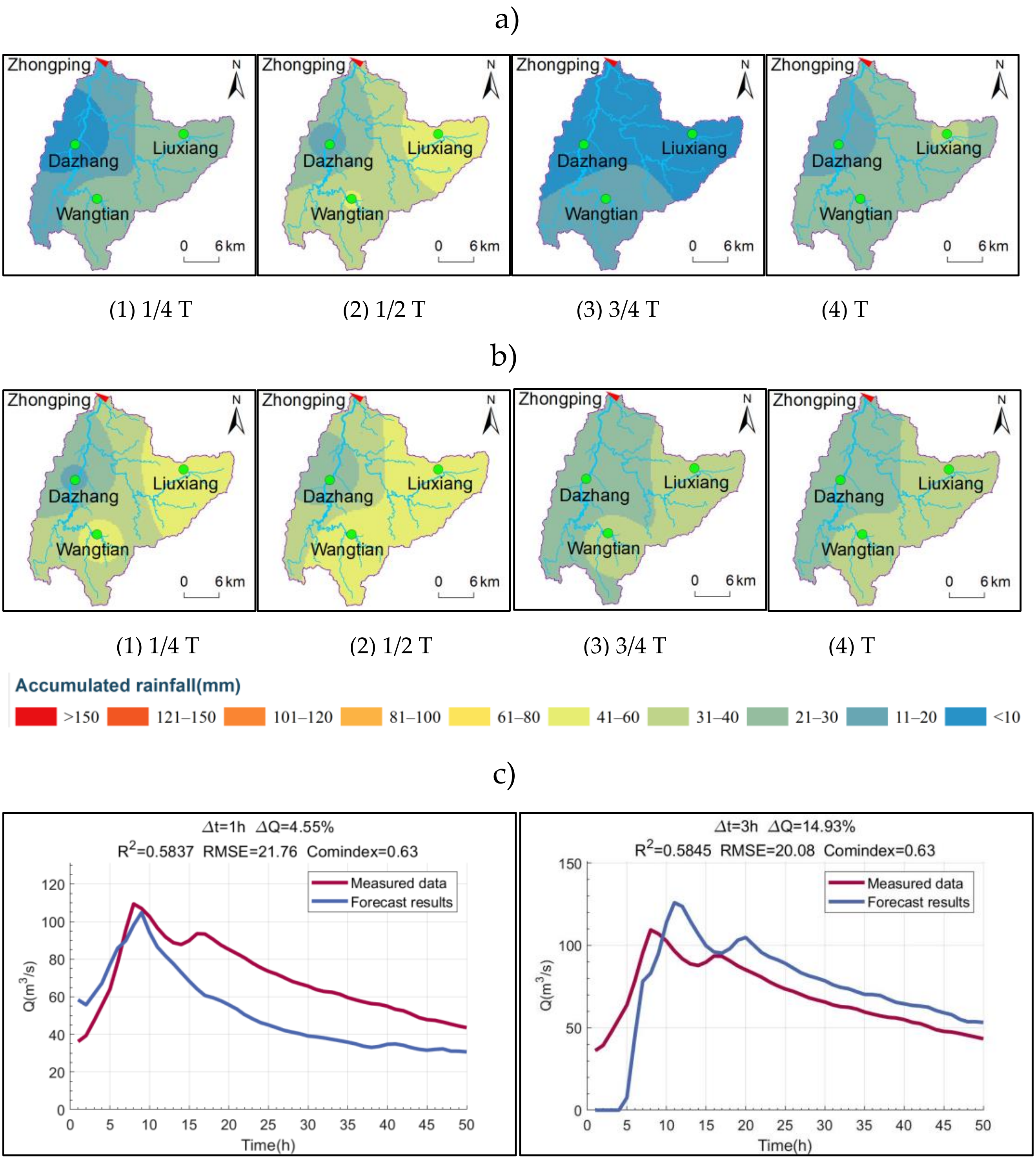
(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

b)

(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

c)

Figure 8. The second identification case. (**a**) Rainfall processes to be identified (5 June 2014); (**b**) identification results (10 August 2011); (**c**) flood forecast results.



(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

(1) 1/4 T (2) 1/2 T (3) 3/4 T (4) T

Figure 9. The second identification case. (**a**) Rainfall processes to be identified (20 June 2022); (**b**) identification results (7 September 2015); (**c**) flood forecast results.

3.2. Discussion

Based on the results of the third group, on 11 June 2022, about 2–3 h after the onset of rainfall, a judgment can be made to issue a rain warning. This rainfall may reach 120 mm, with a maximum rainfall at a single station possibly reaching 124 mm, a maximum 1 h rainfall possibly reaching 35.5 mm, a maximum peak flow possibly reaching 1170 m^3/s, and a total flood volume possibly reaching 24.9898 million m^3. The maximum peak flow may be reached 21 h after the onset of rainfall. This rainfall and flood risk information, which differs from traditional flood forecasting methods, is crucial for flood risk management.

From the results in Figures 5–9 and Table 1, it can be observed that the spatial distribution pattern of the identified historical rainfalls is not exactly the same as that of the rainfalls to be identified. After all, it is almost impossible for two rainfalls to be "exactly the same", and it is not possible to have two rainfalls with the same spatial distribution pattern. However, the identified historical rainfall is quite consistent with the rainfall to be identified in terms of indicators such as average rainfall, maximum rainfall at a single station, maximum hourly rainfall, and the spatial and temporal characteristics of the storm center during the rainfall process.

Using the algorithm proposed in this paper, the average error in identifying surface rainfall is 19.04%, the average error in identifying the maximum rainfall at a single station is 18.09%, and the average error of the maximum 1 h rainfall is 31.56% for the five identified rainfall flooding events. As shown in Tables 2 and 3, a series of indicators proposed in this paper can effectively discern the similarity of floods in terms of the shape of the flood process, flood peak flow, etc., which can enhance the accuracy of the algorithm's flood forecasting.

Table 3. Comparison table of flooding processes by LSTM.

Serial Number	Group Name	Rainfall	Q (m^3/s)	Errors (%)	W ($\times 10^5$ m^3)	Errors (%)	Time of Flood Peaks (h)	⊗T (h)
1	A1	23/05/2022	40.2		458.85		18	
	B12	-	66.56	65.57	452.95	1.29	13	5
2	A2	23/05/2015	72.7		653.94		24	
	B22	-	61.54	15.35	738.67	12.96	17	6
3	A3	11/06/2022	1140		1998.12		19	
	B32	-	957.75	15.99	1554.47	22.20	20	1
4	A4	05/06/2014	60.9		414.32		14	
	B42	-	65.49	7.54	356.53	13.95	17	3
5	A5	20/06/2022	110		928.64		8	
	B52	-	126.43	14.94	1344.51	44.78	11	3
		Average error		23.88		19.04		3.6

Samples in A group are to be identified, samples in B group are flood forecast results.

The flood forecasting results using both the algorithm proposed in this paper and the LSTM neural network model can provide a better prediction of floods. The algorithm proposed in this paper, compared to the LSTM neural network model, has the following advantages:

1. Insufficient information required and long foresight period

Since the mechanisms of the algorithms are different, the algorithm proposed in this paper is based on the identification of the spatiotemporal features of rainfall for forecasting the entire flooding process, while the LSTM neural network model relies on the pretemporal sequence of rainfall and flow data for rolling hour-by-hour forecasting. The two methods require different datasets and forecast periods. The algorithm in this paper can forecast the complete future flood process based only on the first one-quarter time of the rainfall process, effectively extending the foresight period for floods. In contrast, LSTM models require not only the rainfall process but also the flow process in the pretemporal sequence to obtain the future 1 h flow, resulting in a shorter foresight period. The algorithm proposed in this paper is more advantageous in small watersheds in hilly areas where information acquisition is more challenging.

2. Higher forecasting accuracy for flood flow and peak time

The algorithm proposed in this paper has been demonstrated to accurately forecast flooding for all five rainfall events. The average error of the predicted flood peak flow is 8.33%, significantly better than the 23.53% prediction error of the LSTM model. Regarding the peak present time, the average error of the algorithm proposed in this paper is only 1 h, which is significantly better than the LSTM model's prediction error of 3.6 h.

4. Conclusions

Based on machine learning algorithms, this paper proposes a novel flood forecasting method that recognizes the dynamic spatiotemporal features of heavy rainfall in small watersheds in hilly areas. A series of indicators are introduced to discriminate the similarity of flooding processes, improving the accuracy of flood forecasting. The conclusions are as follows:

1. The algorithm presented in this paper identifies historical rainstorms similar to current rainstorms in terms of surface rainfall, hourly rainfall, and the spatial and temporal dynamics of the rainstorms. Regarding flood forecasting, the average error in forecasting flood peak flow and peak present time meets the requirements of flood forecasting accuracy. The average error in flood peak flow forecast is 8.33%, and the peak present time is 1 h, satisfying the needs of flood control and emergency response.
2. In comparison to the LSTM neural network model, the algorithm proposed in this paper requires less information and has a longer foresight period to forecast the entire flood process. Additionally, it provides significantly more accurate forecasts for important indicators such as flood flow and peak present time.
3. Due to the limitations of available data, this study only uses the rainfall and flood data of the past 20 years from the Zhongping small watershed as samples. As time progresses, the quantity and quality of rainfall and flood samples will increasingly improve, and with the gradual development and refinement of the technology, more objective, reasonable, and accurate forecasting results can be achieved in the future.
4. The results indicate that the model in this article can provide a general framework for modeling the spatial heterogeneity and correlation of hydro-meteorological variables and achieve accurate and reliable flood forecasts, thereby enhancing the model's applicability in flood prevention platforms and systems.

This method only considers the spatiotemporal characteristics of rainfall; thus, the Zhongping small watershed with minimal human impact was selected for demonstration. In the future, more historical hydrological data from various types of small watersheds in hilly areas can be collected, such as different watershed areas, underlying surface topography, vegetation conditions, river channel morphology, and engineering scheduling conditions. This will enrich the types of learning samples, enhance the intelligence and applicability of the algorithm, and promote its use on a larger scale.

Author Contributions: Conceptualization, Y.L. (Yuanyuan Liu); methodology, Y.L. (Yuanyuan Liu) and Y.L. (Yesen Liu); validation, K.L.; formal analysis, Y.L. (Yang Liu); resources, Z.L. and W.Y.; writing—original draft preparation, Y.L. (Yuanyuan Liu); writing—review and editing, Y.L. (Yuanyuan Liu) and Y.L. (Yesen Liu); visualization, Y.L. (Yesen Liu); funding acquisition, Y.L. (Yesen Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (52394235).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: The authors declare no conflict of interest. Weitao Yang is employee of Guangxi Water & Power Design Institute Co., Ltd. But this paper was not funded by Guangxi Water & Power Design Institute Co., Ltd. The company had no roles in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, or in the decision to publish the articles. The paper reflects the views of the scientists and not the company.

References

1. Qiao, J.H. Experience and enlightenment of prevention 'July 23' flood in the Haihe River. *China Water Resour.* **2023**, *18*, 9–12.
2. Liu, Y.Y.; Zheng, J.W.; Liu, H.W.; Liu, Y.H.; Liu, Y.S.; Liu, X.P.; Feng, M.M.; Liu, S. Warning and thoughts of extreme rainstorm on urban flood prevention. *China Flood Drought Manag.* **2021**, *31*, 21–24.
3. Finnerty, B.D.; Smith, M.B.; Seo, D.J. Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs. *J. Hydrol.* **1997**, *203*, 21–38. [CrossRef]
4. Sugawara, M. Tank model with snow component. In *Study Report of National Research Center for Disaster Prevention*; Japan Science and Technology Agency: Kawaguchi, Japan, 1984; Volume 293.
5. Zhao, R.J.; Zhuang, Y.L.; Fang, L.R. The Xinanjiang Model. In Proceedings of the IASH 129, Hydrological Forecasting Proceeding, Oxfort Symposium, Oxford, UK, 15–18 April 1980; pp. 351–356.
6. Abdelmounim, B.; Benaabidate, L.; Bouizrou, I.; Aqnouy, M. Implementation of Distributed Hydrological Modeling in a Semi-Arid Mediterranean Catchment "Azzaba, Morocco". *J. Ecol. Eng.* **2019**, *20*, 236–254. [CrossRef] [PubMed]
7. Yang, D.; Herath, S.; Oki, T.; Musiake, K. Application of Distributed Hydrological Model in the Asian Monsoon Tropic Region with a Perspective of Coupling with Atmospheric Models. *J. Meteorol. Soc. Jpn. Ser. II* **2001**, *79*, 373–385. [CrossRef]
8. Khan, I.; Ali, M. Potential Changes to the Water Balance of the Teesta River Basin Due to Climate Change. *Am. J. Water Resour.* **2019**, *7*, 95–105. [CrossRef]
9. Saavedra, O.; Koike, T.; Yang, D. Application of a distributed hydrological model coupled with dam operation for flood control purposes. *Annu. J. Hydraul. Eng. JSCE* **2006**, *50*, 61–66. [CrossRef]
10. Guo, S.L.; Xiong, L.H.; Yang, J.; Peng, H.; Wang, J.X. A DEM and physically based distributed hydrological model. *J. Wuhan Univ. Hydraul. Electr. Eng.* **2000**, *6*, 1–5.
11. Reed, S.; Schaake, J.; Zhang, Z. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.* **2007**, *337*, 402–420. [CrossRef]
12. Bashirgonbad, M.; Nia, A.M.; Khalighi-Sigaroodi, S. A hydro-climatic approach for extreme flood estimation in mountainous catchments. *Appl. Water Sci.* **2024**, *14*, 98. [CrossRef]
13. Prodhan, F.A.; Zhang, J.H.; Hasan, S.S. A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. *Environ. Model. Softw.* **2022**, *149*, 105327.1–105327.14. [CrossRef]
14. Hitokoto, M.; Sakuraba, M. Applicability of the deep learning flood forecast model against the flood exceeding the training events. In Proceedings of the Annual Conference of JSAI, Shizuoka, Japan, 28–31 May 2018; Available online: https://confit.atlas.jp/guide/event/jsai2018/subject/1D1-03/detail (accessed on 3 July 2024).
15. Luppichini, M.; Barsanti, M.; Giannecchini, R.; Bini, M. Deep learning models to predict flood events in fast-flowing watersheds. *Sci. Total Environ.* **2022**, *813*, 151885. [CrossRef] [PubMed]
16. Do Hoai, N.; Udo, K.; Mano, A. Downscaling Global Weather Forecast Outputs Using ANN for Flood Prediction. *J. Appl. Math.* **2011**, *2011*, 246286.1–246286.14. [CrossRef]
17. Akbari Asanjan, A.; Yang, T.; Hsu, K.; Sorooshian, S.; Lin, J.; Peng, Q. Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. *J. Geophys. Res. Atmos.* **2018**, *123*, 12.543–12.563. [CrossRef]
18. Amrul, F.; Aminaton, M.; Shahrum, A. Forecasting of Malaysia Kelantan River using Support Vector Regression Technique. *Int. J. Comput. Syst. Sci. Eng.* **2021**, *39*, 297–306.
19. Luo, Y.X.; Zhou, Y.L.; Chen, H.; Xiong, L.H.; Guo, S.L.; Chang, F. Exploring a spatiotemporal hetero graph-based long short-term memory model for multi-step-ahead flood forecasting. *J. Hydrol.* **2024**, *633*, 130937. [CrossRef]
20. Lee, J.; Choi, J.; Jang, S.; Kim, S. Effect of mountainous rainfall on uncertainty in flood model parameter estimation. *Hydrol. Res.* **2024**, *55*, 221–236. [CrossRef]
21. Hochkirchen, T. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2010**, *173*, 467. [CrossRef]
22. Liu, Y.Y.; Li, L.; Liu, Y.S.; Chand, P.W.; Zhang, W.-H. Dynamic spatial-temporal precipitation distribution models for short-duration rainstorms in Shenzhen, China based on machine learning. *Atmos. Res.* **2020**, *237*, 104861. [CrossRef]
23. Belkin, M.; Niyogi, P. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*; MIT Press: Cambridge, MA, USA, 2003.
24. Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef]
25. Hartigan, J.A.; Wong, M.A. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100–108. [CrossRef]
26. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Schmidhuber, J.; Gers, F.; Eck, D. Learning Nonregular Languages: A Comparison of Simple Recurrent Networks and LSTM. *Neural Comput.* **2014**, *14*, 2039–2041. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.