IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 7, 2019, accepted September 4, 2019, date of publication September 13, 2019, date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941234

CRML: A Convolution Regression Model With Machine Learning for Hydrology Forecasting

CHEN CHEN[1,2], (Senior Member, IEEE), QIANG HUI[1], QINGQI PEI[1], (Senior Member, IEEE), YANG ZHOU[3], BIN WANG[4], NING LV[1], AND JI LI[3]

[1]State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China
[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China
[3]Goldenwater Information Technology Development Company Ltd., Beijing 101400, China
[4]School of Communication Engineering, Xi'an University of Science and Technology, Xi'an 710054, China

Corresponding author: Chen Chen (cc2000@mail.xidian.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0121400, in part by the National Natural Science Foundation of China under Grant 61571338, Grant U1709218, Grant 61672131, and Grant U1636209, in part by the Key Research and Development Plan of Shaanxi Province under Grant 2017ZDCXL-GY-05-01, Grant 2019ZDLGY13-04, and Grant 2019ZDLGY13-07, in part by the Key Laboratory of Embedded System and Service Computing, Tongji University, under Grant ESSCKF2019-05, and in part by the Xian Key Laboratory of Mobile Edge Computing and Security under Grant 201805052-ZD3CG36.

ABSTRACT Hydrologic disasters often result in substantial property damage and casualties. Therefore, hydrology forecasting, especially the flooding, has become a hot research spot in all countries of the world. Based on the basic principle of flooding formation, this paper proposes a data-driven hydrology forecasting model, i.e., the CRML (Convolution Regression based on Machine Learning). This model could reflect the impact of hourly rainfall on the future flow changes and the flow changes are predicted by superimposing these impacts. First, our work is implemented on historical data onto the Xixian River Basin in Henan Province, China. Through the data filtering, the training set of our model is constructed by using the flood process selection algorithm proposed in this paper. Next, the gradient descent algorithm is used to update the weights of the model, and the optimal weights are verified by ten flooding events generated in the past ten years. Finally, the numerical results show that the qualified rates of our model in predicting flood peak flow and its arrival time is approximately 90% and 100%, respectively. Compared with the latest popular artificial intelligence schemes, our model structure is clear and concise. And combined with the physical meaning of the traditional model and machine learning technology, our model can accurately complete the task of long and short lead time hydrology forecasting.

INDEX TERMS Hydrology forecasting, convolution regression algorithm, flood process selection algorithm, gradient descent, machine learning.

I. INTRODUCTION

The hydrology plays an important role in the sustainable development of human society, which is the lifeblood of human activities, agriculture and industry. Its important to pay more attention on hydrology forecasting, especially the flood forecasting. In fact, accurate prediction of water level has been a hot topic since ancient times [1]. The basis of an accurate prediction is a large amount of measured data onto hydrology, meteorology and geology. Actually, due to the great number of influential factors, the flood forecasting is a complex nonlinear process, and neither a single mathematical nor physical model can accurately describe this process [2]. Experts and scholars in the field of hydrology have proposed many flood forecasting models and systems based on the principles of hydrology in the past few decades.

The Hydrologic Engineering Center's-Hydrologic Modeling System(HEC-HMS) is a basin rainfall-runoff model system developed by the United States Army Corps of Engineers (USCE) Hydrologic Engineering Center (HEC). It is a semi-distributed secondary flood runoff model with physical concepts. The model can calculate the peak amount, runoff and peak time of the total runoff process at the outlet of the basin by calculating the runoff and convergence of each sub-basin [3]. Zhao Renjun et al. proposed the

The associate editor coordinating the review of this manuscript and approving it for publication was Gilberto Pastorello.

XAJ model, considered the relationship between the model and natural conditions, and added meteorological, climatic, geological and geomorphological conditions to the model [4]. Liu Z et al. proposed the TOPKAPI model, which assumed that the movement of water in the soil, the surface and the inner side of the channel can be simulated by motion waves and is widely used in hydrology forecasting all over the world [5].

The traditional hydrological models analyze the process of flood formation from the physical and mathematical aspects by modeling the natural environment to conduct hydrology forecasting. However, these models usually contain many parameters, and each parameter is related to local hydrological and natural data which are often difficult to obtain completely and accurately. Studies have shown that the model parameter settings have a great impact on the prediction results [6], if the required parameters cannot be obtained completely, the results will not be as expected.

In addition, with the fast development of statistical machine learning [7]–[9] and computing technologies [10]–[12], many data-driven artificial intelligence models were also emerging. In 1995, Charles A et al. proposed a hydrology forecasting method based on component analysis. By analyzing the principal factors affecting the flood generation of different regions, a flash flood forecasting model was established [7]. Rudolf Scitovski et al. proposed a short-term and long-term water level prediction method for one river measurement location in 2012. Long-term forecasting is considered as the problem of investigating the periodicity of water level behavior at least one year by using linear-trigonometric regression and short-term forecasting within several days is based on the modification of the nearest neighbor method [8]. However, this method is not generalized. In 2016, Wahid Palash et al. proposed a ReqSim (Requisite Simplicity) model based on linear regression, which adds historical water level data, historical rainfall data and weather-model-generated forecasted rainfall data into the model [9]. This model works fairly well for 1-10-day forecasts. But it is impossible to accurately predict the peak of the flood and its arrival time which are most important in flood forecasting.

Some machine learning algorithms such as support vector machine [10], decision tree [11], logistic regression [12] have also been applied to hydrology forecasting. The ANN (Artificial Neural Networks, ANN) models have been widely used in hydrology because of its good performance in solving nonlinear problems [13], [14]. C.L. Wu et al. proposed an ANN model by combining multiple data processing techniques [15]. Ramli Adnan added Kalman filtering technique to the ANN model to correct the error [16]. However, these models can achieve good results in short lead time predictions which the lead time is less than two days [17], but there are large errors in long lead time predictions which the lead time is longer than two days.

In general, traditional hydrological models could yield a good prediction result by precisely modeling the flood formation process, but a large number of parameters need to be manually calibrated. Instead, data-driven models can readily achieve the short lead time flood-forecasting based on historical data, but its performance in long lead time hydrology forecasting is weak.

In order to balance the performance gap between these two types of models, this paper proposed a novel hydrology forecasting model by using the proposed convolution regression algorithm based on the basic principle of flood formation. This model can mainly reflect the influence of rainfall on the flow of the basin section. It superimposes the effect of unit hourly rainfall on the future n-hour flow changes, which means the output of the model is a series of data rather than a single value. This makes it possible to predict the process of a flood, peak value and its arrival time which are the most important in flood forecasting [18].

The most significant innovation of this article is that it is the first time to add the discrete convolution functions to the regression problem. Deducing and explaining its solution steps in detail. Due to the characteristics of the convolution function, not all samples are required to be structurally consistent. The number of the inputs and outputs of each sample just need to meet the same quantitative relationship. This feature makes the selection of samples very flexible and the flood processes with different duration can be selected to train the model. This feature can be also applied in the field of signal processing [19]. As a result, the contributions of this paper can be generalized as follows:

1) The first main contribution of this paper is proposed a novel convolution regression algorithm, derived its closed-form expression and gave its iterative solution by using a machine learning method.
2) The second contribution is that a flood process selection algorithm is proposed to extract the flood processes in historical data.
3) Finally, The flood forecasting model based on the proposed CRML is validated by historical flooding events. The results showed that the model has high accuracy in predicting flood peak and its arrival time.

II. METHODS

A. STUDY AREA AND DATA PROCESS

The data used in this paper are rainfall data and flow data from January 1, 2011 to September 7, 2018 of 50 rainfall stations and one hydrological station (Xixian hydrological Station) in the Xixian Basin, Henan Province, China as shown in Figure 1. Both rainfall and hydrological stations collect data in hours with some missing data.

In the historical data of Xixian basin, there are a lot of missing rainfall data. The Inverse Distance Weighting method(IDW) is one of the traditional methods which is the most commonly used approach for estimation of missing rainfall data [20].The formula of weighting method is given

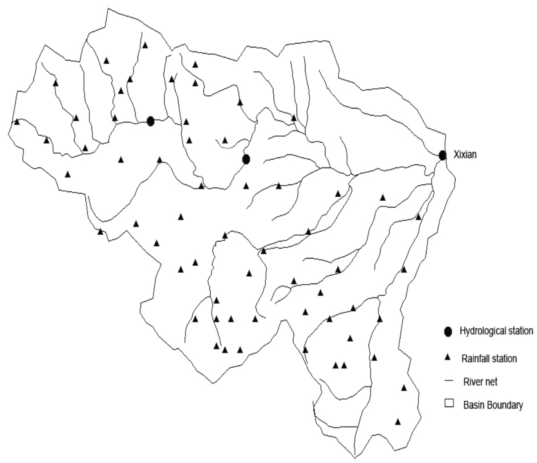


FIGURE 1. **Distribution map of the rainfall and hydrological stations in Xixian Basin, Xixian Hydrological Station is on the right-most side of the map.**

as Equation 1.

$$R_t = \sum_{\substack{i=1 \\ i \neq t}}^{N} \alpha_i R_i \tag{1}$$

$$\alpha_i = \frac{d_{it}^{-p}}{\sum\limits_{\substack{i=1 \\ i \neq t}}^{N} d_{it}^{-p}} \tag{2}$$

where R_t is the estimated value of the missing data at the target station t; N is the number of the neighboring stations; R_i is the observation at the i_{th} neighboring station and α_i is the weight of the i_{th} neighboring station with constraint $\sum_{i=1}^{N} \alpha_i = 1$. In the Equation 2, the d_{it} is the distance between the target station t and the i_{th} neighboring station. Greater values of p assign greater influence to values closest to the target station. The value of p usually ranges from 1.0 to 6.0 and the most commonly used value for p is 2. In this study, the value for p is 2, the number of neighboring station is 3.

The hourly rainfall data of each rainfall station is then weighted and summed according to the Thiessen polygon [21]. The Thiessen polygon was proposed by the Dutch climatologist Thiessen. Thiessen uses the rainfall intensity of a rainfall station in each Thiessen polygon to represent the average rainfall intensity of this polygonal region. The ratio between the area of the Thiessen polygon and the total area of the basin is used as the weight of the rainfall station. Through data filling and superposition, the hourly rainfall data of the Xixian Basin is obtained. For the flow missing data from the hydrology station, the interpolation methods can be used because of its continuity. The missing flow data is filled by quadratic interpolation [22] in this paper.

The important notations in this paper are listed in TABLE 1.

TABLE 1. **Notations.**

Lable	Notation
α_i	the weight of the i_{th} neighboring station
d_{it}	the distance between the target station t and the i_{th}
W_t	the flow value at time t
σ_j	the convolution coefficient
r_{t-j+1}	the rainfall value at time t-j+1
S_i	the sum of the square error
T_i	the duration of a flood process
$W^i_{pred(t)}$	the predict flow value of flood process i at the time t
$W^i_{real(t)}$	the real flow value of flood process i at the time t
m	the number of samples
E	the loss function
K_{lk}	the independent variable of the normal function
V_l	the dependent variable of the normal function
σ_j^n	the value of σ_j in n_{th} iteration
η	the learning rate
$\nabla\sigma_j^n$	the gradient of σ_j^n
α	the attenuation coefficient
$iter$	the current iteration number
$decay\_steps$	the attenuation speed
N	the length of a sample
Nr	the net rainfall
R	the rainfall
Sh	the shape parameter of a certain basin
CN	the Soil Conservation Service's curve number
$r(\delta)$	the correlation coefficient of rainfall and flow
δ	the time interval between rainfall and flow
i	the i_{th} value of rainfall or flow
$\overline{\mathrm{W}}$	the average value of flow data
$\overline{\mathrm{R}}$	the average value of rainfall data
tol	the minimum drop error in gradient descent algorithm

B. PROPOSED CRML

In this chapter, the proposed convolution regression algorithm is deduced, and the closed-form solution and iterative solution for solving the convolution coefficients are given. The hydrology forecasting problem can be simplified as the superposition of the influence of hourly rainfall on the future flow changes. The convolution formula [23] shown in Equation 3 can be used to get the flow at any time in the future:

$$W_t = \sum_{j=1}^{n} \sigma_j r_{t-j+1} \tag{3}$$

where W_t is the flow at the moment t, σ_j is the convolution coefficients, r_{t-j+1} is the rainfall at time $t-j+1$.

How to measure the error between the predicted and actual values is the key to measure the performance of the regression task. In this paper, the sum of square error [24] is used as the performance metric. The sum of square error has a good geometric meaning, which represents the Euclidean distance. For any sample in this problem, the sum of square error can be expressed as:

$$S_i = \sum_{t=1}^{T_i} (W^i_{pred(t)} - W^i_{real(t)})^2 \tag{4}$$

where S_i is the sum of square error of the i_{th} sample, T_i is the length of the sample, $W^i_{pred(t)}$ is the predicted value of the flow of the sample i at the time t while the $W^i_{real(t)}$ is expressed as the actual value. Then our task can be transformed into the following objective function:

$$\begin{aligned}\sigma^* &= \underset{\sigma}{\text{argmin}} \sum_{i=1}^{m} S_i \\ &= \underset{\sigma}{\text{argmin}} \sum_{i=1}^{m}\sum_{t=1}^{T_i}(w^i_{pred(t)} - w^i_{real(t)})^2 \\ &= \underset{\sigma}{\text{argmin}} \sum_{i=1}^{m}\sum_{t=1}^{T_i}(\sum_{j=1}^{n}\sigma_j r_{t-j+1} - w^i_{real(t)})^2 \end{aligned} \tag{5}$$

In the above objective function, the first summation formula represents the total errors of m training samples, and the second summation formula represents the sum of square error between the actual value and the predicted value for each training sample. This formula is similar to the objective function in linear regression, except that we introduce a convolution formula. In order to solve this objective function, a loss function was designed as shown in Equation 6:

$$\begin{aligned}E &= \frac{1}{2m}\sum_{i=1}^{m}\sum_{t=1}^{T_i}(w^i_{pred(t)} - w^i_{real(t)})^2 \\ &= \frac{1}{2m}\sum_{i=1}^{m}\sum_{t=1}^{T_i}(\sum_{j=1}^{n}\sigma_j r_{t-j+1} - w^i_{real(t)})^2 \end{aligned} \tag{6}$$

The method of solving the parameters σ to minimize the error E is called the least squares method [25]. To achieve E minimum, you are required to:

$$\begin{aligned}\frac{\partial E}{\partial \sigma_l} &= \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{q}(\sum_{j=1}^{n}\sigma_j r^i_{t-j+1} - q^i_{real(t)})r^i_{t-l+1} \\ &= 0, \quad (l = 1, 2, \ldots, n)\end{aligned} \tag{7}$$

Then we can get the normal equations of its least squares method.

$$\begin{cases} K_{11}\sigma_1 + K_{12}\sigma_2 + \ldots + K_{1n}\sigma_n = V_1, \\ K_{21}\sigma_1 + K_{22}\sigma_2 + \ldots + K_{2n}\sigma_n = V_2, \\ \ldots\ldots \\ K_{n1}\sigma_1 + K_{n2}\sigma_2 + \ldots + K_{nn}\sigma_n = V_n. \end{cases} \tag{8}$$

$$K_{lk} = \sum_{i=1}^{m}\sum_{t=1}^{q} r^i_{t-l+1} r^i_{t-k+1}, \quad \{l, k = 1, 2 \ldots, n\} \tag{9}$$

$$V_l = \sum_{i=1}^{m}\sum_{t=1}^{q} W^i_{real(t)} r^i_{t-l+1}, \quad \{l = 1, 2, \ldots, n\} \tag{10}$$

Convert it to a matrix form:

$$K\sigma = V \tag{11}$$

The optimal closed-form solution [26] for convolution coefficients can be obtained:

$$\sigma = K^{-1}V \tag{12}$$

The above equations prove that the proposed method can converge to an optimal solution. However, it is difficult to get that optimal solution directly. Gradient descent algorithm is a kind of iterative method and often used to solve the least squares problem [27]. The gradient descent ensures that in each iteration, the parameters are updated in the opposite direction of the loss function gradient. Then the iteration formula of σ_j can be expressed as:

$$\begin{aligned}\sigma_j^{n+1} &= \sigma_j^n - \eta\nabla\sigma_j^n \\ &= \sigma_j^n - \eta\frac{\partial E}{\partial \sigma_j^n}\end{aligned} \tag{13}$$

where η is the learning rate, which is used to control the step size of each σ_j iteration. The learning rate determines how fast the parameters move to the optimal values. If the learning rate is too large, it is likely to exceed the optimal value; conversely, if the learning rate is too small, the efficiency of the optimization is low, and the algorithm may not converge for a long time [28]. In this paper, in order to improve the efficiency of the algorithm operation, a more flexible learning rate setting method-exponential decay method [29] is added to achieve exponential decay of the learning rate. In this way, the model uses a large learning rate at the beginning of training to quickly obtain a suboptimal solution, and then gradually reduce the learning rate as the iteration progresses, making the model more stable in the later stage of training. The learning rate exponential decay function is designed in this paper as follows:

$$Decay\_rate(iter) = \alpha^{\left\lfloor \frac{iter}{decay\_steps} \right\rfloor} \tag{14}$$

where $\alpha \in (0, 1)$ is the attenuation coefficient. *iter* is the current iteration number and *decay_steps* is the attenuation speed. The ratio is rounded down so that the decay rate becomes a step function.

Using the learning rate decay function in Equation 14, the iterative formula can be modified to:

$$\sigma_j^{n+1} = \sigma_j^n - \eta\alpha^{\left\lfloor \frac{n}{decay\_steps} \right\rfloor}\nabla\sigma_j^n \tag{15}$$

According to the above theory, the pseudo code of the convolution regression algorithm can be seen in Algorithm 1: where $x_k = \{x^1, x^2, \ldots, x^n\}$ and $y_k = \{y^1, y^2, \ldots, y^m\}$. The stop condition is set to reach the maximum number of iterations or the minimum drop error is less than *tol*.

C. PHYSICAL MEANING OF THE MODEL

The model proposed in this paper was inspired by the unit hydrograph theory first mentioned by Sherman in 1932 [30]. The unit hydrograph theory is derived from the physical assumption that all floods due to rainstorms of a given duration runoff in the same amount of time. Zoch *et al.* proposed a general physical theory of streamflow based on the assumption that at any time the streamflow is proportional to the amount of rainfall remaining with the soil at that time which is called net rainfall based on the work of Sherman. All unit hydrograph procedures are based on two fundamental: Invariance and superposition, which means flood can be

Algorithm 1 Convolution Regression Algorithm

Input: The training sets $D = \{(x_k, y_k)\}_{k=1}^{m}$;
The convolution coefficient σ_i;
The learning rate η;
The attenuation coefficient α;
The decay steps *dec*;
The maximum number of iterations *iter*;
The minimum decline error *tol*;

Output: convolution coefficients;

Random initialization convolution coefficients σ_i;
repeat
  for all $(x_k, y_k) \in D$ **do**
    Calculate the output of the current sample $\hat{w}_i$ based on the current convolution coefficients according to Equation 3;
    Calculate the gradient term $\nabla\sigma_j$ according to Equation 7;
    Update the parameters σ_j according to Equation 15;
  end for
until The stop condition is reached

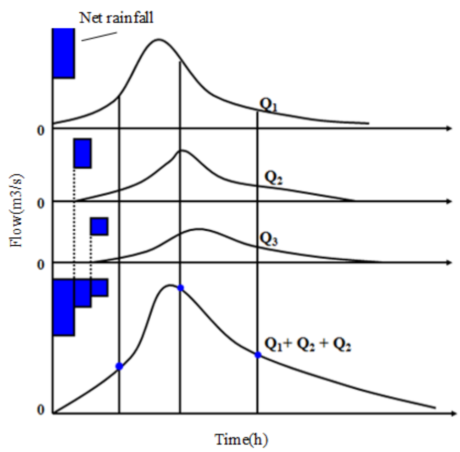


FIGURE 2. The fundamental of the unit hydrograph.

superimposed by the unit net rainfall times the unit hydrograph as shown in Figure 2.

However, this theory has a few limitations. One of the biggest limitations is that it must be associated with a production function in order to provide an estimate of net rainfall [31]. But over the years researchers and practitioners have mostly used empirical methods [32] such as use the Soil Conservation Service(SCS)s curve number (CN) method [33] to obtain the net rainfall. The CN method is an empirical equation to calculate net rainfall by using a shape parameter *Sh* based on soil, vegetation, land use and soil moisture prior to a rainfall event [34]:

$$Nr = \begin{cases} \dfrac{(R - 0.2Sh)^2}{P + 0.8Sh}, & R > 0.2Sh \\ 0, & R \le 0.2Sh \end{cases} \tag{16}$$

where Nr(mm) is net rainfall, R(mm) is rainfall. The shape parameter *Sh* is obtained from:

$$Sh = \frac{25400}{CN} - 254 \tag{17}$$

where CN ranges from 0 to 100. The CN is determined from land cover and management, and from the hydrologic soil group using a table from the SCS handbook [35]. This traditional method has many parameters which need to calibrated by experienced experts. To solve this problem, the model based on the proposed CRML learns the convolution coefficients which has the similar meaning to unit hydrograph from the historic data directly and ignores the soil, topography, and vegetation of a certain basin. The proposed model not only retains the physical meaning of the traditional model, but also is simpler and more objective than the traditional model.

III. SIMULATION OF THE CRML

In order to verify the effectiveness of the algorithm, this section simulates the proposed CRML. We assume that the convolution coefficients can be generated by the Equation 18 in this paper to better demonstrate the effectiveness of the algorithm.

$$\sigma_t = 0.5 * \sin(2\pi t) + \frac{1}{2}t^2 \tag{18}$$

where the $t \in [0, 1]$, and the step is 0.01 so that we can get 100 convolution coefficients σ. Then we can get the training and testing sets by generating a series of numbers randomly as the input X and then convolving X with the convolution coefficients as the output Y. For the purpose of verifying the robustness of the algorithm, we added white noise to the output of each training sample as shown in Figure 3.

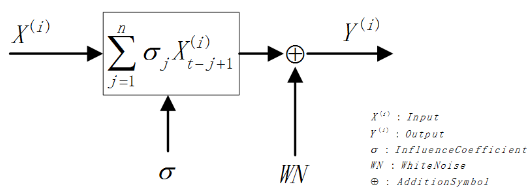
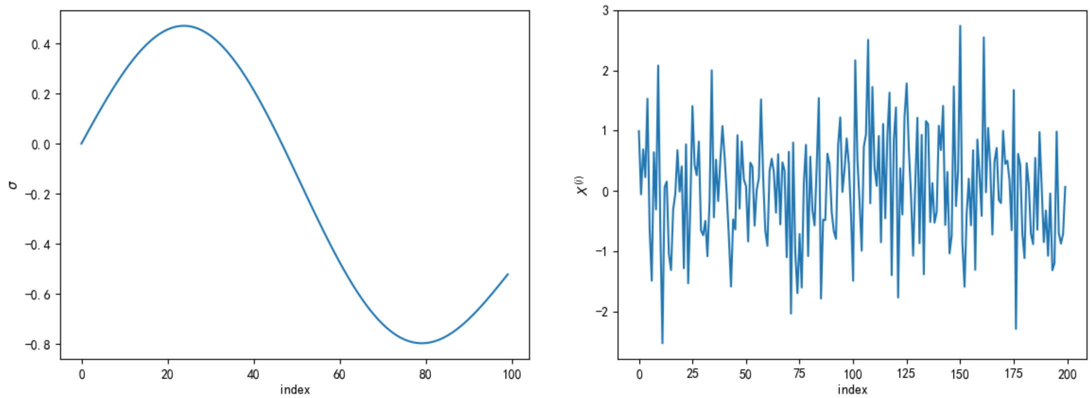


FIGURE 3. Generating a training sample by add white noise to the output.

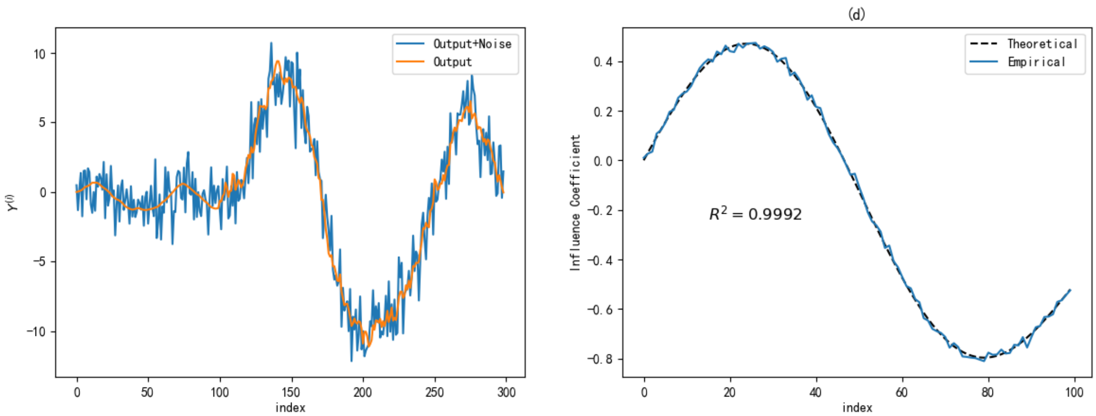
In addition, in order to reflect that the proposed algorithm doesn't need to maintain the consistency of the samples which have the same length of input X and the output Y. Samples with different length of X and Y are generated in the simulation, and the statistics are shown in TABLE 2. Then set the learning rate as 0.005 and the total iterations as 200, the result of the simulation can be shown in Figure 4 and TABLE 3. The results were discussed by using three statistical measures: Root Mean Square Error (RMSE), mean absolute error (MAE), and determination coefficient (R^2) given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{19}$$



(a) The theoretical convolution coefficients

(b) The input X of a training example.



(c) The output Y of a training sample.

(d) The theoretical and empirical of convolution coefficients

FIGURE 4. The simulation results. (a) The theoretical convolution coefficients generated by the Equation 18. (b) The input X of a training sample generated by randomly. (c) The output Y of the a training example by convolving the convolution coefficients shown in (a) with the input X shown in (b). (d) The theoretical and empirical of convolution coefficients.

TABLE 2. The statistics of training and testing sets.

Types	input_X	output_Y	Training samples	Testing samples
1	50	149	250	50
2	75	174	250	50
3	100	199	250	50
4	150	249	250	50
5	200	299	250	50

TABLE 3. Statistical performance of the simulation.

Types	Training			Testing		
	RMSE	MAE	R2	RMSE	MAE	R2
1	1.5257	1.1945	0.6682	0.0814	0.0640	0.9995
2	1.4937	1.1855	0.8895	0.0766	0.0626	0.9998
3	1.2554	0.9906	0.7955	0.0818	0.0619	0.9984
4	1.4400	1.1630	0.8920	0.097	0.0724	0.9991
5	1.4670	1.1551	0.8678	0.1132	0.0830	0.9997

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{20}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y_i})^2} \tag{21}$$

The convolution coefficients can be obtained after training as shown in Figure 4 (d), compared the empirical results with the theoretical results by using the R^2. The R^2 was 0.9992 which means the calculated results match the real results perfectly. The error of the training sets comes from the white noise, while the testing sets's R^2 score all above 0.99.

This simulation verified that the proposed algorithm in this paper is effective and robust. In the next section, we will implement this algorithm on hydrology forecasting.

IV. RESULTS AND DISCUSSION ON HYDROLOGY FORECASTING

A. CREATE TRAINING SET AND TESTING SET

Based on the above analysis, we can easily convert the flood process into the superposition of the hourly rainfall effect on the flow. For training the convolution coefficients in the proposed CRML, we generated training set and testing set in the historical rainfall and flow data in Xixian basin from January 1, 2010 to September 7, 2018.

In hydrology forecasting, people pay more attention to the peak of the flood and its arrive time [18]. However, there is a large amount of non-flood data in the historical data which would cause error in training. Therefore, a flood progress selection algorithm was proposed in this paper and expressed at Algorithm 2.

Algorithm 2 Flood Progress Selection Algorithm

Input: The historical rainfall data *Rainfall*;
The historical streamflow data *Flow*;
The historical average streamflow *avg_flow*;
The sliding window size *sw_size*;

Output: Selected data set $\hat{D}\{x, y\}$;

```
Initialize the selected data set D̂{x, y};
set threshold thr = 2 * avg_flow;
for all index in D do
    if mean(Flow[index : index + sw_size]) ≥ thr then
        end = index + sw_size + 1;
        while mean(Flow[end − sw_size:end]) ≥ thr do
            end = end + 1;
        end while
        add Flow[index − sw_size:end] in D̂[y];
        add Rainfall[index − sw_size:end] in D̂[x];
        index = end + 1;
    else
        index = index + 1
    end if
end for
```

In the Algorithm 2, we used a sliding window to extract flood progresses in the historical data. When the average streamflow of a sliding window greater than the double of the historical average streamflow, it can be considered as the start of a flood process. Then slide the window until the average streamflow in the sliding window is less than twice of the historical average streamflow, it can be considered as the end of a flood process. We extracted the rainfall and flow data between the start and end of a flood process as the input X and the output Y. By using the proposed flood process selection algorithm, a set of samples that only consist of flood processes are generated. Random selection of 10 flood processes from the sample set is utilized to verify the effectiveness of the model while the rest samples are divided into training set and testing set by 7:3. In order to make the model converge as quickly as possible, we normalized the data set. Normalization [36] treats data as dimensionless data ranging between 0 and 1 as shown in Equation 22.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{22}$$

B. PARAMETERS OF THE MODEL

In the hydrology forecasting model based on the proposed CRML, the most important parameter is the number of the convolution coefficients. According to the analysis in the section II, the number of the convolution coefficients represents the duration of unit hourly rainfall influences to flow changes. However, the time at which the flow changes in different region are affected by rainfall is also different. Thats because the flow changes in a river is not only affected by rainfall, but also related to the underlying surface conditions of the basin [37] such as basin topography, size, shape, slope, and vegetation conditions. Through the correlation analysis [38], the relationship between the rainfall process and the flow process of a certain basin is obtained. The correlation coefficient is a statistical indicator designed by statistician Carl Pearson and is the amount of linear correlation between the variables [39], as shown in Equation 23.

$$r(\delta) = \frac{\sum_i [(W_{i+\delta} - \overline{W}) * (R_i - \overline{R})]}{\sqrt{\sum_i (W_{i+\delta} - \overline{W})^2} * \sqrt{\sum_i (R_i - \overline{R})^2}} \tag{23}$$

where W is the flow data, $\overline{W}$ is the average value of the flow data; R is the rainfall data, $\overline{R}$ is the average value of the rainfall data; δ is the time interval between the rainfall data and the flow data which means the rainfall process is δ hours earlier than flow process. The trend of the correlation coefficient between rainfall and flow within different time interval of Xixian basin as shown in Figure 5.

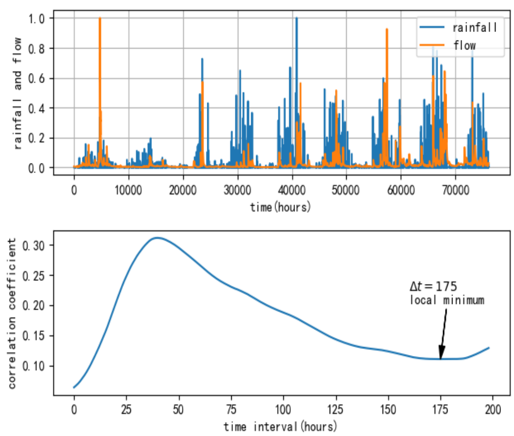


FIGURE 5. Correlation coefficient between rainfall and flow.

The Figure 5 shows that the effect of unit hourly rainfall on flow changes is raising at first and then lower for nearly 200 hours and at the 40th hour reaches the maximum, at the 175th hour, reaches the local minimum. In other words, unit hourly rainfall can effectively affect the future flow changes for 175 hours in Xixian basin, which means the number of convolution coefficients can be set as 175. The other parameters of the algorithm 1 are based on multiple experiments, and the settings can be shown in Table 4.

Where the learning rate η is 0.005, the attenuation coefficient α is 0.95, the *decay_steps* is set as 10 while the total iterations are set as 200, and the minimum decline error

TABLE 4. Convolution regression algorithm parameter table.

Parameter	Value
η	0.005
α	0.95
$Decay\_steps$	10
$iter$	200
tol	1e-6

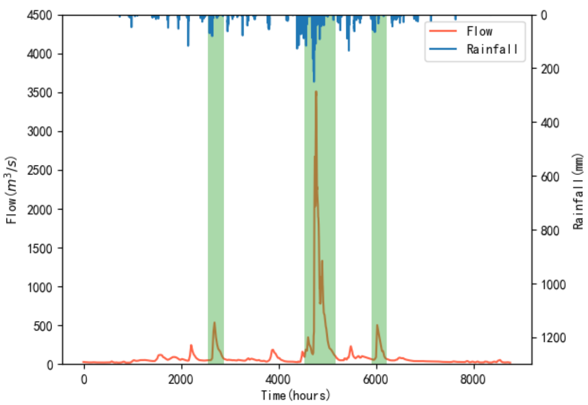


FIGURE 6. The effect of the flood process selection algorithm in 2010 of Xixian basin.

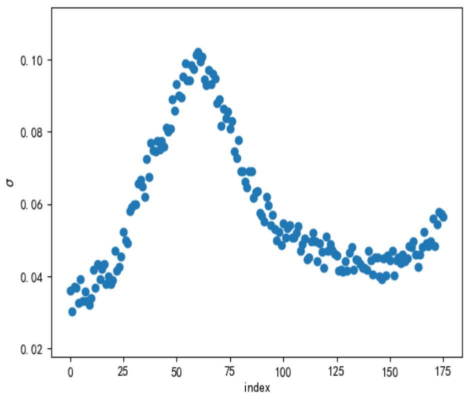


FIGURE 7. The convolution coefficients obtained by training.

is 0.000001. In the algorithm 2, the historical average streamflow of the Xixian basin was $84.97m^3/s$, and the size of the sliding window was 175. The effect of the algorithm 2 can be shown in Figure 6, the curve covered by the green rectangle is the selected flood processes.

C. DISCUSS THE RESULTS

In this subsection, we discussed the results of the experiment and analyzed the effects of the model. After training by using the parameters given in subsection B, the convolution coefficients as shown in Figure 7 can be obtained.

According to the Figure 7, the trend of convolution coefficients is in line with the above analysis that the unit hourly rainfall effects on the change of flow are first increased and then decreased. Using the trained influence coefficient,

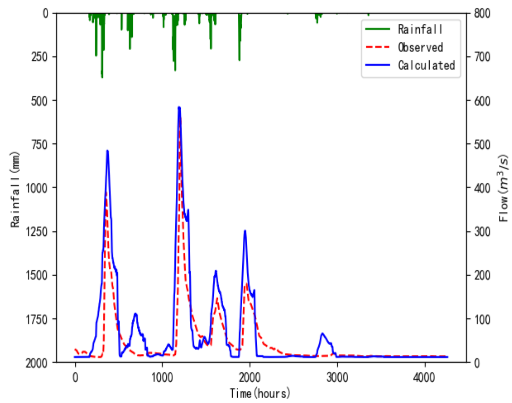


FIGURE 8. Comparison of observed and calculated flow for Xixian in 2018.

TABLE 5. The mode based CRML performance statistics for Xixian.

Types	Coefficient of determination r^2	Peak flow (m^3/s)	Relative error(%)	Time error (hours)
Observed	0.71	583.55	4.2	18
Calculated		560		

the model was validated using the data of 2018 in Xixian basin as shown in Figure 8 and performance statistics in Table 5.

As can be seen from the Figure 8 and Table 5, the coefficient of determination between observed and calculated values is 0.71, and the relative errors (the ratio of the absolute error to the observed) of estimated maximum peak flow is 4.2% and its arrive time error is 18 hours.

To better illustrate the effects of the model, we extracted 10 flood processes to verify the effectiveness of the model. The results are shown in Table 6 and Figure 9.

According to *Standard for hydrological information and hydrological forecasting* [40], the permissible error of the flood peak forecast is 20% of the measured flood peak flow; the permissible error of the flood peak arrive time is 30% of the interval from the start of the forecast to the time when the peak of the measured flood peak appears. According to the results shown in Table 6 and Figure 9, the model proposed in this paper can accurately predict the trend of floods. The qualified rate of flood peak flow is 90%, and the qualified rate of peak time is 100%, which are in line with hydrology forecasting claim.

D. DISCUSS THE FORECAST ERRORS

The error of the hydrology forecasting model based on convolution regression algorithm proposed in this paper mainly consists of three aspects: (1) human factors such as flow changes caused by reservoir storage; (2) weather forecast errors, especially rainfall forecast error; (3) Model error caused by data processing. Among them, the error caused by human factors has the greatest impact, because the process of humanitys participation cannot be reflected in the data; and the model studies the influence of rainfall on the

TABLE 6. Performance statistics on 10 historical floods by using the proposed model.

Event Number	Event Period	Observed Peak	Calculated Peak	Peak Error	Allowed Error	Satisfied	Peak Arrive Time Error	Allowed Error	Satisfied
1	2010-07-16 04:00-2010-07-25 06:00	3510	3580.8	-70.8	702	Y	-10	21.6	Y
2	2012-09-06 09:00-2012-09-13 17:00	2010	975.7	1034.3	402	N	-21	24.9	Y
3	2013-07-18 01:00-2013-07-24 04:00	389	392.9	-3.93	77.8	Y	-19	30.9	Y
4	2013-08-23 01:00-2013-08-30 16:00	560	486.9	73	112	Y	12	27.3	Y
5	2014-08-29 17:00-2014-09-05 01:00	1070	1111	-41	214	Y	-14	24	Y
6	2015-06-26 16:00-2015-07-03 05:00	1810	1955.9	-145.9	362	Y	-17	21	Y
7	2016-06-28 04:00-2016-07-03 07:00	1990	1723.1	266.8	398	Y	-18	31.2	Y
8	2016-07-15 21:00-2016-07-25 15:00	3250	3600.2	-350.2	650	Y	-32	38.7	Y
9	2017-07-06 16:00-2017-07-11 16:00	2150	1782.9	367	430	Y	-25	28.2	Y
10	2018-07-28 08:00-2018-08-03 14:00	680	813.9	-133.9	136	Y	0	18	Y

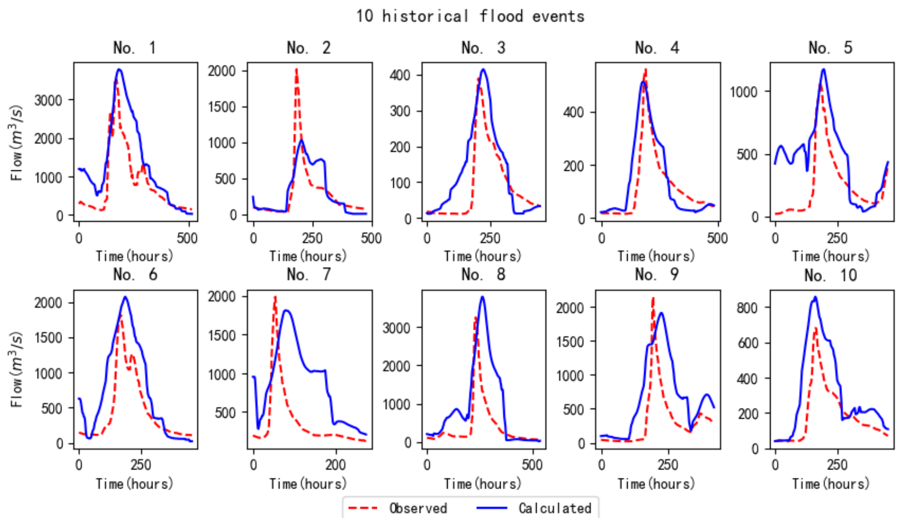


FIGURE 9. Validating the model by using the 10 historical flood events.

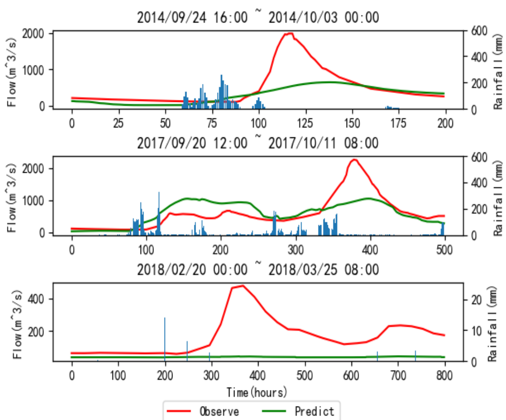


FIGURE 10. Three datasets of abnormal flooding.

flow changes. If the flow changes is independent of rainfall, it will greatly affect the accuracy of the model. As shown in Figure 10, it can be noted that the corresponding rainfall is not enough to cause the flow changes of this scale. Therefore, during these time periods, the flow change is influenced by other unidentified stronger factors than rainfall such as reservoirs storage or release of water.

V. CONCLUSION

In summary, this paper designs a novel convolution regression algorithm, which introduces the convolution function into the regression problem, and gives the closed-form solution for the convolution coefficient and the gradient-based, exponential-based iterative solution step of attenuation. Because of the characteristics of convolution function, the output of the algorithm is a series of data rather than a single value. This means that we can predict the flood process, especially the peak value of the flood and its arrival time which are most important in hydrology forecasting. On the other hand, the samples used for training are not necessary to satisfy the consistency. We can extract flood processes with different durations for training models without any other data segmentation.

In order to apply the proposed CRML to hydrology forecasting, a flood process selection algorithm based on the sliding window was proposed in this paper. After using this algorithm, a set of samples consisting of flood processes was generated. The convolution coefficients are obtained after several rounds of iterations with reasonable parameters. Finally, the hydrology forecasting model based on the CRML was verified by 10 flood events, the result shows that it has high accuracy in flood peak and its arrival time, and complies with the requirements of hydrology forecasting.

The hydrology forecasting model based on CRML can effectively predict short and long lead-time floods due to its physical meaning. The key difference of the proposed model to previous work is that this model does not need to calibrate complex parameters compared with the traditional hydrological forecasting models, but only needs to learn convolution coefficients from historical data . This greatly simplifies the flood forecasting work and improves the forecasting efficiency. Meanwhile, this model has the ability to predict flood process directly by using the rainfall data and the trained convolution coefficients without more complex calculations and has good effect in long lead-time flood forecasting.

In addition to uncontrollable factors such as manual intervention, rainfall forecasting errors, etc., data processing methods such as quadratic interpolation, smoothing and other model errors, etc. will affect the result of the model. Therefore, further work needs to be done in the processing and extraction of data. Future work will start with quantitative analysis of historical data, design more scientific and effective experimental methods, and determine the parameters of the model. Our future work will also investigate the combination of advanced communication technology [41]–[45] with artificial intelligence [46]–[48], to introduce real-time flooding measurement monitoring, intelligent parameters filtering and adaptive model selection, into the flooding forecasting as well as controlling.

REFERENCES

[1] F. A. Ruslan, Z. M. Zain, R. Adnan, and A. M. Samad, "Flood water level prediction and tracking using particle filter algorithm," in *Proc. IEEE 8th Int. Colloq. Signal Process. Appl.*, Mar. 2012, pp. 431–435.

[2] Y. Sun, W. Bao, P. Jiang, X. Ji, S. Gao, Y. Xu, Q. Zhang, and W. Si, "Development of multivariable dynamic system response curve method for real-time flood forecasting correction," *Water Resour. Res.*, vol. 54, no. 7, pp. 4730–4749, Jul. 2018.

[3] A. D. Feldman, *Hydrologic Model. system HEC-HMS: Tech. reference manual*. US Army Corps of Engineers, Hydrologic Engineering Center, 2000.

[4] R. Zhao, P. Wang, and F. Hu, "Relations between parameter values and corresponding natural conditions of xinanjiang model," *J. Hohai Univ., Natural Sci.*, vol. 20, no. 1, pp. 52–59, 1992.

[5] Z. Liu, M. L. Martina, and E. Todini, "Flood forecasting using a fully distributed model: Application of the topkapi model to the upper Xixian catchment," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 9, no. 4, pp. 347–364, Oct. 2005.

[6] E. S. Cooper, S. L. Dance, J. Garcia-Pintado, N. K. Nichols, and P. J. Smith, "Observation impact, domain length and parameter estimation in data assimilation for flood forecasting," *Environ. Model. Softw.*, vol. 104, pp. 199–214, Jun. 2018.

[7] C. A. Doswell, H. E. Brooks, and R. A. Maddox, "Flash flood forecasting: An ingredients-based methodology," *Weather Forecasting*, vol. 11, no. 4, pp. 560–581, Dec. 1996.

[8] R. Scitovski and S. Maričić, and S. Scitovski, "Short-term and long-term water level prediction at one river measurement location," *Croatian Oper. Res. Rev.*, vol. 3, no. 1, pp. 80–90, Dec. 2012.

[9] W. Palash, Y. Jiang, A. S. Akanda, D. L. Small, A. Nozari, and S. Islam, "A streamflow and water level forecasting model for the ganges, brahmaputra, and meghna rivers with requisite simplicity," *J. Hydrometeorol.*, vol. 19, no. 1, pp. 201–225, Jan. 2018.

[10] M. S. Tehrany, B. Pradhan, and M. N. Jebur, "Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS," *J. Hydrol.*, vol. 512, pp. 332–343, May 2014.

[11] T. M. Shafapour, B. Pradhan, and M. N. Jebur, "Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS," *J. Hydrol.*, vol. 504, no. 8, pp. 69–79, Nov. 2013.

[12] A. M. Youssef, B. Pradhan, M. N. Jebur, and H. M. El-Harbi, "Landslide susceptibility mapping using ensemble bivariate and multivariate statistical models in Fayfa area, Saudi Arabia," *Environ. Earth Sci.*, vol. 73, no. 7, pp. 3745–3761, Apr. 2015.

[13] R. J. Abrahart, F. Anctil, P. Coulibaly, C. W. Dawson, N. J. Mount, L. M. See, A. Y. Shamseldin, D. P. Solomatine, E. Toth, and R. L. Wilby, "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting," *Prog. Phys. Geography, Earth Environ.*, vol. 36, no. 4, pp. 480–513, Jun. 2012.

[14] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions," *Environ. Model. Softw.*, vol. 25, no. 8, pp. 891–909, Aug. 2010.

[15] C. Wu, K. W. Chau, and C. Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," *J. Hydrol.*, vol. 389, no. 1, pp. 146–167, Jul. 2010.

[16] R. Adnan, F. A. Ruslan, A. M. Samad, and Z. M. Zain, "Artificial neural network modelling and flood water level prediction using extended Kalman filter," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng.*, Nov. 2012, pp. 535–538.

[17] B. W. Golding, "Long lead time flood warnings: Reality or fantasy?" *Meteorol. Appl.*, vol. 16, no. 1, pp. 3–12, Mar. 2010.

[18] M. Campolo, P. Andreussi, and A. Soldati, "River flood forecasting with a neural network model," *Water Resour. Res.*, vol. 35, no. 4, pp. 1191–1197, Apr. 1999.

[19] J. Virieux, "P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method," *Geophysics*, vol. 51, no. 4, pp. 889–901, Apr. 1986.

[20] J. Suhaila, M. D. Sayang, and A. A. Jemain, "Revised spatial weighting methods for estimation of missing rainfall data," *Asia–Pacific J. Atmos. Sci.*, vol. 44, no. 2, pp. 93–104, May 2008.

[21] F. R. Fiedler, "Simple, practical method for determining station weights using thiessen polygons and isohyetal maps," *J. Hydrol. Eng.*, vol. 8, no. 4, pp. 219–221, Jul. 2003.

[22] X. Chen, C. Mei, B. Xu, K. Yu, and X. Huang, "Quadratic interpolation based teaching-learning-based optimization for chemical dynamic system optimization," *Knowl.-Based Syst.*, vol. 145, pp. 250–263, Apr. 2018.

[23] A. Brandt, "A signal processing framework for operational modal analysis in time and frequency domain," *Mech. Syst. Signal Process.*, vol. 115, pp. 380–393, Jan. 2019.

[24] T. Beckers, J. Umlauft, and S. Hirche, "Mean square prediction error of misspecified Gaussian process models," in *Proc. IEEE Conf. Decision Control (CDC)*, Dec. 2018, pp. 1162–1167.

[25] G. De Luca, J. R. Magnus, and F. Peracchi, "Weighted-average least squares estimation of generalized linear models," *J. Econometrics*, vol. 204, no. 1, pp. 1–17, May 2018.

[26] S. Zheng, X. Cai, C. Ding, F. Nie, and H. Huang, "A closed form solution to multi-view low-rank regression," in *Proc. 29th AAAI Conf. Artif. Intell.*, Feb. 2015, pp. 1973–1979.

[27] C. Tan, S. Ma, Y.-H. Dai, and Y. Qian, "Barzilai-borwein step size for stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 685–693.

[28] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 464–472.

[29] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," Dec. 2012, *arXiv:1212.5701*. [Online]. Available: https://arxiv.org/abs/1212.5701

[30] J. C. Dooge, "A general theory of the unit hydrograph," *J. Geophys. Res.*, vol. 64, no. 2, pp. 241–256, Feb. 1959.

[31] M. Sivapalan, "Process complexity at hillslope scale, process simplicity at the watershed scale: Is there a connection?" *Hydrol. Processes*, vol. 17, no. 5, pp. 1037–1041, Apr. 2003.

[32] R. Rigon, M. Bancheri, G. Formetta, and A. de Lavenne, "The geomorphological unit hydrograph from a historical-critical perspective," *Earth Surf. Processes Landforms*, vol. 41, no. 1, pp. 27–37, Jan. 2016.

[33] S. L. Dingman, *Physical Hydrology*, vol. 7458. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.

[34] M. G. Huang, "The necessities of levying upon the education texes in China from the taxation theories in the western countries," *J. Yibin Teachers College*, vol. 5, no. 2, pp. 40–45, 2005.

[35] *SCS National Engineering Handbook. Section 4. Hydrology*. United States Dept. Agricult., Soil Conservation Service, Washington, DC, USA, 1985.

[36] J. Sulaiman and S. H. Wahab, "Heavy rainfall forecasting model using artificial neural network for flood prone area," in *Proc. IT Converg. Secur.* Singapore: Springer, 2018, pp. 68–76.

[37] K. Khosravi, B. T. Pham, K. Chapi, A. Shirzadi, H. Shahabi, I. Revhaug, I. Prakash, and D. T. Bui, "A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at haraz watershed, northern iran," *Sci. Total Environ.*, vol. 627, pp. 744–755, Jun. 2018.

[38] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Inf. Process. Manage.*, vol. 54, no. 3, pp. 359–369, May 2018.

[39] A. Ly, M. Marsman, and E.-J. Wagenmakers, "Analytic posteriors for Pearson's correlation coefficient," *Statistica Neerlandica*, vol. 72, no. 1, pp. 4–13, Feb. 2018.

[40] J. Sun, J. Zhang, J. Wang, J. Liang, S. Zhang, S. Chen, J. Le, X. Rui, S. Ge, and L. Li, "Standard for hydrological information and hydrological forecasting," Chinese Standard, Chinese Ministry of Water Resources, 2008.

[41] C. Chen, H. Jinna, T. Qiu, M. Atiquzzaman, and Z. Ren, "CVCG: Cooperative V2V-aided transmission scheme based on coalitional game for popular content distribution in vehicular ad-hoc networks," *IEEE Trans. Mobile Comput.*, to be published.

[42] C. Chen, Q. Pei, and X. Li, "A GTS allocation scheme to improve multiple-access performance in vehicular sensor networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1549–1563, Mar. 2016.

[43] C. Chen, X. Liu, H.-H. Chen, M. Li, and L. Zhao, "A rear-end collision risk evaluation and control scheme using a Bayesian network model," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 264–284, Jan. 2019.

[44] C. Chen, L. Liu, X. Du, X. Wei, and C. Pei, "Available connectivity analysis under free flow state in VANETs," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 1, pp. 1–19, Aug. 2012.

[45] C. Chen, Y. Jin, Q. Pei, and N. Zhang, "A connectivity-aware intersection-based routing in vanets," *EURASIP J. Wireless Commun. Netw.*, vol. 2014, no. 1, pp. 1–14, Mar. 2014.

[46] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in SAR images," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5530–5538, Dec. 2018.

[47] R. Lai, J. Guan, Y. Yang, and A. Xiong, "Spatiotemporal adaptive nonuniformity correction based on BTV regularization," *IEEE Access*, vol. 7, pp. 753–762, 2019.

[48] Z. Duan, Y. Yang, K. Zhang, Y. Ni, and S. Bajgain, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31820–31827, 2018.

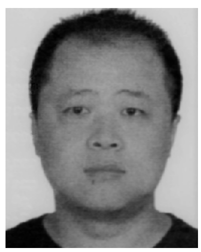
CHEN CHEN (M'09–SM'18) received the B.Eng., M.Sc., and Ph.D. degrees in electrical engineering and computer science (EECS) from Xidian University, Xi'an, China, in 2000, 2006, and 2008, respectively, where he is currently an Associate Professor with the Department of EECS and also the Director of the Xi'an Key Laboratory of Mobile Edge Computing and Security and the Intelligent Transportation Research Laboratory. He was a Visiting Professor in EECS with the University of Tennessee and in CS with the University of California. He serves as the General Chair, the PC Chair, the Workshop Chair, or a TPC Member for a number of conferences. He has authored or coauthored two books and over 80 scientific articles in international journals and conference proceedings. He has contributed to the development of five copyrighted software systems and invented 50 patents. He is also a Senior Member of the China Computer Federation (CCF) and a member of the ACM and the Chinese Institute of Electronics.

QIANG HUI received the B.Eng. degree in communication engineering from Nanjing Tech University, Nanjing, China, in 2017. He is currently pursuing the master's degree with Xidian University. His major is transportation information engineering and control. His research interests include wireless communication and computer engineering.

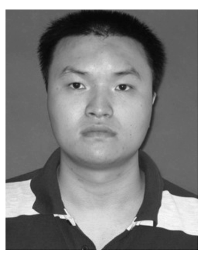
QINGQI PEI (SM'15) received the B.Eng., M.Sc., and Ph.D. degrees in computer science and cryptography from Xidian University, in 1998, 2005, and 2008, respectively, where he is currently a Professor and a member of the State Key Laboratory of Integrated Services Networks. His research interests include cognitive networks, data security, and physical layer security. He is also a Professional Member of the ACM and a Senior Member of the Chinese Institute of Electronics and the China Computer Federation.

YANG ZHOU received the B.Eng. degree in communication engineering from Xidian University, Xian, China, in 2000. Since 2000, he has been with Goldenwater Information Technology Development Company Ltd., Beijing, China. His research interests include wireless communication and computer engineering.

BIN WANG received the B.Eng., M.Sc., and Ph.D. degrees in electrical engineering and computer science (EECS) from Xidian University, Xi'an, China, in 2003, 2006, and 2011, respectively. Since 2011, he has been with the School of Communication Engineering, Xi'an University of Science and Technology. His research interests include wireless communication and computer engineering.



NING LV received the B.Eng. degree in electrical engineering and automation from Xi'an Jiaotong University, in 2000, and the M.Sc. degree in circuit and system engineering from Xidian University, Xi'an, China, in 2004, where he is currently pursuing the Ph.D. degree in pattern recognition and intelligent system.



JI LI received the B.Eng. degree in computer science and technology and the M.Sc. degree in software engineering from Beihang University, Beijing, China, in 2002 and 2007, respectively. He has been with Goldenwater Information Technology Development Company Ltd., Beijing, China, where he is involved in consulting, designing, and researching of water resources information construction for a long time.