**Accepted Manuscript**

**International Journal of Pattern Recognition and Artificial Intelligence**



| | |
|---|---|
| Article Title: | Water level prediction of rainwater pipe network using an SVM-based machine learning method |
| Author(s): | Hao Wang, Lixiang Song |
| DOI: | 10.1142/S0218001420510027 |
| Received: | 11 January 2019 |
| Accepted: | 28 March 2019 |
| To be cited as: | Hao Wang, Lixiang Song, Water level prediction of rainwater pipe network using an SVM-based machine learning method, *International Journal of Pattern Recognition and Artificial Intelligence*, doi: 10.1142/S0218001420510027 |
| Link to final version: | https://doi.org/10.1142/S0218001420510027 |



World Scientific
www.worldscientific.com

*Type of the Paper (Article, Review, Communication, etc.)*

# Water level prediction of rainwater pipe network using an SVM-based machine learning method

**Hao Wang [1], Lixiang Song [2]***

[1]. China Institute of Water Resources and Hydropower Research, Beijing 100038, China
[2]. Department of Water Resources and Environment, Pearl River Hydraulic Research Institute, Guangzhou 510611, China
* Correspondence: Lixiang Song, slx.hust@live.cn

Academic Editor:
Received: date; Accepted: date; Published: date



**Abstract:** Model accuracy and running speed are the two key issues for flood warning in urban areas. Traditional hydrodynamic models, which have a rigorous physical mechanism for flood routine, have been widely adopted for water level prediction of rainwater pipe network. However, with the amount of pipes increase, both the running speed and data availability of hydrodynamic models would be decreased rapidly. To achieve a real-time prediction for the water level of the rainwater pipe network, a new framework based on a machine learning method was proposed in this article. The spatial and temporal autocorrelation of water levels for adjacent manholes was revealed through theoretical analysis, and then a support vector machine (SVM)-based machine learning model was developed, in which the water levels of adjacent manholes and rivers-near-by-outlets at the last time step were chose as the independent variables, and then the water levels at the current time step can be computed by the proposed machine learning model with calibrated parameters. The proposed framework was applied in Fuzhou city, China. It turns out that the proposed machine learning method can forecast the water level of the rainwater pipe network with good accuracy and running speed.

**Keywords:** urban flood; real-time prediction; support vector machine; machine learning methods

## 1. Introduction

As a consequence of global warming and rain islands phenomena resulting from urbanization, heavy rainfall with short duration is more likely to occur. Floods caused by heavy rainfall are becoming more and more severe in urban areas [1]. Once Floods occurs in the urban areas, it will cause tremendous social loss. For instance, on July 10, 2004, a rainfall event occurred in Beijing (capital of China) caused 2 m flooding in more than 10 overpass bridges and result in serious traffic jams [2]. In the past decades, urban flood disaster has attracted considerable attention, and much work has been focused on the establishment of an effective early warning system [3, 4]. The timely flood warning should be very useful for prevention of waterlogging disasters. For example, the traffic police could close the roads before inundation occurs if the accurate predictions were timely provided. So model accuracy and running speed are the two key issues for flood warning in urban areas.

In the recent decade, Hydrodynamic model based on two-dimensional shallow water equation has been widely used in flood prediction and has a strict physical mechanism of flood process [5-7]. As the terrain in urban areas is very complex with many irregular buildings and narrow streets, computational mesh refinement should be required for hydrodynamic modelling. However, the running speed would decrease markedly with the increasing number of compute grids [8].

Generally speaking, the urban flood is swift, which is defined as less than 3 hours before the peak of hydrology. So the time consumption of hydrodynamic modelling should be reduced to a half hour to provide a practical warning service. With the development of computer hardware, varies of the high-performance calculating technique were employed to enhance the efficiency of computing, such as models of parallel computation that using multiple CPU cores and graphics processing unit (GPU) cards [8-13]. Those parallel models have a great tradeoff between computational accuracy and running time, and thus can be used for forecasting and warning of the flood on real-time.

From the perspective of the practical application, detailed boundary conditions, including rainfall, tidal levels of rivers near the city, flash floods of mountains crossing the city, and real-time states of facilities including pumps, sluices, rainwater pipe network, and manholes, should all be provided for urban floods simulation in tidal areas. For example, blocking of manholes or pipes would induce a local inundation during a heavy rainfall [14]. If the blocking was not considered by hydrodynamic models, the local inundation would not be precisely predicted. Furthermore, the accuracy of predictions of different boundary conditions can directly influence the final accuracy of inundation prediction. In another word, the model accuracy was decreasing step by step from the rainfall prediction to the final inundation prediction. So accurate prediction of urban inundation is a complex task in the framework of hydrodynamic modeling, since a fair amount of data is often not available.

Another type of model for flood prediction is based on data-driven methods, which do not require detailed boundary conditions. This type of model can deal with nonlinear problems effectively and satisfactorily by machine learning method, and is not restricted by integral and/or differential mathematical operations [15, 16]. So, compared to hydrodynamic models, these machine learning based models have some significant advantages, including relatively simple implementation, rapid running speed, rapid convergence, and strong adaptability [15-19]. With the development of waterlogging monitoring system, real-time water level monitoring of rainwater pipe network is becoming increasingly prevalent in modern cities. So a large amount of useful real-time monitoring data can be used by machine learning methods for water level prediction of the rainwater pipe network, which is vital for urban flood warning.

Varies of machine learning methods, including artificial neural networks (ANN), support vector machines (SVM) and adaptive neuro-fuzzy inference system (ANFIS), have been successfully applied to hydrological context over the last decades [20-24]. The SVM is a new method based on artificial intelligence developed on the basis of statistical learning theory. Misra et al. [24] compared the ANN and SVM to make a prediction for runoff and sediment in different time scales, and concluded that both ANN and SVM could accurately simulate runoff and sediment yield on time scale of the day, week, and month. Tayfur et al. [16] adopted the genetic algorithm (GA), the ant colony optimization (ACO), and the particle swarm optimization (PSO) ways to train the ANN, and concluded that the above ANN, GA, ACO, and PSO models are useful measures for flood hydrography prediction. Yu et al. [25] builted a real-time, river water level prediction model by using the SVM, and results showed that the model can make the prediction for flood level one-to-six-hours ago. Noymanee et al. [26] adopted the Bayesian linear model for forecasting of potential extreme floods and reconstruction of historical rivers floods. SVM is prevalent for machine learning, and it has been widely used, since it has many attractive features including less prone to over-fitting, strong ability of fitting and generalization, and the nice ability to process infinite or nonlinear characteristics in high dimensional eigenspace. With those advantages, SVM is considered to be the most advanced machine learning method with good generalization ability [27].

Although machine learning methods have been increasingly used for discharge and water level predictions in the hydrological context, few works have been done on urban floods prediction, especially the water level prediction of rainwater pipe network. In this work, we have introduced the SVM to urban floods application of water level prediction. This new model framework was based on hydrodynamic theoretical analysis, and the input indexes of the SVM reflect the hydrodynamic relation among adjacent manholes as well as rivers-near-by-outlets. The proposed SVM model was verified through a practical application in Fuzhou city, China.

The organization of the study is: the governing equations of floods in rainwater pipe network were given in section 2, and then the numerical procedure is presented to reveal the spatial and temporal autocorrelation of water levels for adjacent manholes. Section 3 presents the SVM-based model for water level prediction of rainwater pipe network. In Section 4, the validity of the model is verified by a case study in Fuzhou City, China. And last, section 5 gives a brief conclusion.

## 2. Hydrodynamic Model

### *2.1. Governing equations*

Saint venant equation represents the conservation of mass and momentum of the unsteady flow in the pipeline, which can be expressed as:

$$\frac{\partial Z}{\partial t}+\frac{1}{B}\frac{\partial Q}{\partial x}=\frac{q}{B} \tag{1}$$

$$\frac{\partial Q}{\partial t}+gA\frac{\partial Z}{\partial x}+\frac{\partial}{\partial x}(\beta uQ)+g\frac{|Q|Q}{c^2AR}=0 \tag{2}$$

in which, $x$ means the longitudinal coordinate; $t$ means the time; $Z$ means the cross-sectional water elevation; $B$ means the width of channel cross-section at the free water-surface level; $Q$ is the cross-sectional flow discharge; $q$ means the lateral discharge per unit channel length; $A$ is wetted cross-sectional area of the river course; $g$ means gravitational acceleration; $u$ is the cross-sectionally averaged velocity; $\beta$ is the coefficient for correcting the nonuniform velocity distribution; $R$ means the hydraulic radius; c denotes the coefficient of Chezy, given by,

$$c=\frac{R^{1/6}}{n} \tag{3}$$

where $n$ means Manning roughness coefficient.

The confluence condition of mass conservation is needed at a node for rainwater pipe network modelling. The confluence condition is given by,

$$\sum Q=\sum_{j=1}^{k_{\text{in}}}Q_{i,j}-\sum_{j=1}^{k_{\text{out}}}Q_{i,j}=\frac{\partial V}{\partial t}=A_s\frac{\partial H}{\partial t} \tag{4}$$

where the subscripts $i$ and $j$ represent the node number and the conduit number, respectively; $k_{\text{in}}$ and $_{\text{out}}$ are the numbers of inflow and outflow conduits of the $i$th node, respectively; $V$ denotes the node assemble volume; $A_s$ denotes the node assembly surface area; $H$ denotes the hydraulic water head in the node.

The 1D St. Venant equations (1), (2) and the confluence condition (4) with appropriate initial and boundary conditions construct the model of 1D rainwater pipe network flows [28].

### *2.2. Solution Method*

Combining the continuity and momentum equations (1) and (2), the momentum equation of a conduit can be obtained as follows,

$$\frac{\partial Q}{\partial t}=2U\frac{\partial A}{\partial t}+U^2\frac{\partial A}{\partial x}-gA\frac{\partial H}{\partial x}-gn^2\frac{Q|Q|}{AR^{4/3}} \tag{5}$$

It can be seen that equation (5) can be applied to calculate the flow time trajectory for conduits.

Besides, from equation (4), the node continuity equation can be written as,

$$\frac{\partial H}{\partial t}=\frac{\sum Q}{A_s} \tag{6}$$

In the equations (5) and (6), the finite difference approximations take the place of the spatial and temporal derivatives, and then a finite difference form of the linkage momentum equation and

nodal continuity equation can be directly deduced. Furthermore, using the implicit backward Euler method, the discrete form of the linkage momentum equation can be re-written as follows,

$$Q^{t+\Delta t} = \frac{Q^t + \Delta Q_{inertia} + \Delta Q_{pressure}}{1 + \Delta Q_{friction}} \tag{7}$$

in which, $\Delta Q_{inertia}$ is the inertial term, $\Delta Q_{pressure}$ is pressure term, and $\Delta Q_{friction}$ is a friction term.

The discrete form in nodal continuity equation, for non-outfall nodes, can be re-written as,

$$H^{t+\Delta t} = H^t + \frac{\frac{\Delta t}{2}(\sum Q^t + \sum Q^{t+\Delta t})}{A_s^{t+\Delta t}} \tag{8}$$

For outfall nodes, the hydraulic head of water in the node is given by the boundary condition,

$$H^{t+\Delta t} = H_{outfall} \tag{9}$$

From the above equations, it can be seen that the hydrodynamic models calculate new values of its state variables in a series of time steps, in which each step of the system receives a new set of external inputs [28]. For each non-outfall nodes, the simulation process can use the following general mathematical equation sets to represent that is worked out at each time step,

$$H_i^{t+\Delta t} = f(H_i^t,\ H_j^t,\ I^t,\ P) \tag{10}$$

in which, the subscript $i$ denotes the index of nodes; $H_j$ denotes the hydraulic heads of water in the nodes which are direct neighbors of the $i$th node. $I^t$ denotes an input vector at time $t$; $P$ denotes a constant parameter vector.

Equation (10) reveals the spatial and temporal autocorrelation of water levels for adjacent manholes, which is vital for constructing the input indexes of the SVM.

## 3. SVM-based Machine Learning Methods

SVM is prevalent for machine learning [32-36], and it has been widely used, such as fault checking of rolling element bearings [27], flood disaster loss evaluation [29], and so on. Support vector regression (SVR) is an application of the SVM in regression learning. In the following sections, the relevant theories of SVR will be briefly introduced, which have been adopted by Huang et al. [29] for the multi-index comprehensive evaluation.

### *3.1. ε-SVR regression algorithm*

Assuming a training set $\mathbf{S}=[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$, where $\mathbf{x}_i=(x_{i1}, x_{i2}, \ldots, x_{im})$ represents a vector of imput; $y_i$ denotes the corresponding output vector; $n$ means the sample number; $m$ denotes the number of input vector dimension. The input space $\mathbf{X}$ was mapped to a high dimensional feature space by SVR algorithm through non-linear function $\varphi(\mathbf{X})$, and SVR finds a linear regression function in the high dimensional feature space, $f(\mathbf{x})=\mathbf{w}^{\mathrm{T}}\cdot\varphi(\mathbf{X})+b$, which is optimally fitting to the real outputs y in the range of error of $\varepsilon$, where $\mathbf{w}$ and b are parameters of regression function. The corresponding kernel function is defined as $K(\mathbf{x}_i, \mathbf{x}) = \varphi(\mathbf{x}_i)^{\mathrm{T}}\cdot\varphi(\mathbf{x})$.

To optimize the generalization error of regression, a loss function should be defined. In this work, the $\varepsilon$-insensitive loss function [29] is adopted as the loss function. The $\varepsilon$-insensitive loss function tolerates errors in the range of $\varepsilon$, while makes a penalty for the sample that is out of range of $\varepsilon$. The linear form of the $\varepsilon$-insensitive loss function is given by

$$L_\varepsilon(y_i) = \max\{0,\ \left|y_i - [\boldsymbol{w}^{\mathrm{T}}\cdot\varphi(\boldsymbol{x}_i) + b]\right| - \varepsilon\} \tag{11}$$

It should be noted that if the difference between target value $y$ and the regression estimation function value with learning constructs is less than $\varepsilon$, then $y$ = f(x). Based on the $\varepsilon$-insensitive loss

function that the existence of regression value error is allowed, an SVR problem can be represented as the following convex quadratic programming by introducing two positive relaxation variables [30],

$$\begin{aligned} &\min_{\mathbf{w},b,\xi,\xi^*} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\cdot\mathbf{w}+C\sum_{i=1}^{N}(\xi_i+\xi_i^*) \\ &\text{s.t.} \quad \begin{cases} y_i-[\mathbf{w}^{\mathrm{T}}\cdot\varphi(\mathbf{x}_i)+b]\le\varepsilon+\xi_i \\ [\mathbf{w}^{\mathrm{T}}\cdot\varphi(\mathbf{x}_i)+b]-y_i\le\varepsilon+\xi_i^* \quad i=1,2,...,N \\ \xi_i\ge 0 \quad \xi_i^*\ge 0 \end{cases} \end{aligned} \tag{12}$$

in which, $\xi_i, \xi_i^*$ are relaxation variables; $C$ is the regularization parameter controlling the degree of punishment to samples whose error exceeds $\varepsilon$.

The first term of the objective function of equation (12) denotes the universal approximation and generalization properties in the whole data space, and the second term denotes the fitting error of the regression model in the training sample space. Represented by the equation (12) in accordance with the structural risk minimization principle, the SVR can minimize the experienced risk and model complexity simultaneously, thus should preserve the generalization ability in case of limited samples.

By introducing Lagrange multiplying factors, $\lambda_i, \lambda_i^*$, the optimized problem of equation (12) is the same with the following problem based on the Wolfe dual principle [31],

$$\begin{aligned} &\max_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*} \quad \sum_{i=1}^{N}y_i(\lambda_i-\lambda_i^*)-\varepsilon\sum_{i=1}^{N}(\lambda_i+\lambda_i^*)-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\lambda_i-\lambda_i^*)(\lambda_j-\lambda_j^*)\varphi(\mathbf{x}_i)^{\mathrm{T}}\cdot\varphi(\mathbf{x}_j) \\ &s.t. \quad \begin{cases} \sum_{i=1}^{N}(\lambda_i-\lambda_i^*)=0 \\ 0\le\lambda_i,\lambda_i^*\le C \end{cases} \quad i=1,2,...,N \end{aligned} \tag{13}$$

The optimal Lagrange multiplying factors, $\tilde{\lambda}_i, \tilde{\lambda}_i^*$, can be computed by resolving equation (13), and then the parameter of the regression function, $b$, can be calculated through using Karush-Kuhn-Tucker condition [29].

There are only a few non-zero of $\hat{\lambda}_i=\tilde{\lambda}_i-\tilde{\lambda}_i^*$ that corresponding samples named support vectors. According to the functional theory, the kernel function $K(\mathbf{x}_i, \mathbf{x}) = \varphi(\mathbf{x}_i)^{\mathrm{T}}\cdot\varphi(\mathbf{x})$ can mapping the input space $\mathbf{X}$ to a high dimensional eigenspace, and then generating an inner product in high dimensional eigenspace. So, regression function could be given as,

$$f(\mathbf{x})=\sum_{i=1}^{N}\hat{\lambda}_iK(\mathbf{x}_i,\mathbf{x})+b^* \tag{14}$$

in which, $\mathbf{x}_i$ denotes the support vector of training samples set, and $N$ denotes the number of support vectors (SV). The parameter of $b^*$ is given by,

$$b^*=\frac{1}{N}\left\{\sum_{0<\lambda_i<C}[y_i-\sum_{\mathbf{x}_j\in\mathrm{SV}}(\lambda_j-\lambda_j^*)K(\mathbf{x}_j,\mathbf{x}_i)-\varepsilon]+\sum_{0<\lambda_i^*<C}[y_i-\sum_{\mathbf{x}_j\in\mathrm{SV}}(\lambda_j-\lambda_j^*)K(\mathbf{x}_j,\mathbf{x}_i)+\varepsilon]\right\} \tag{15}$$

### *3.2. SVM-based water level prediction model*

#### *3.2.1 Input index of rainwater pipe network*

According to the theoretical analysis in Section 2, it can be concluded that the water level of a manhole in the next time step is determined by the water level of the manhole and its adjacent manholes or rivers-near-by-outlets at the current time step, as well as the rainfalls as the input

condition. So the input index of SVM for water level prediction of rainwater pipe network is constructed as follows,

$$\mathbf{x}_i = (Z_i,\ Z_{i1},\ \ldots\ ,\ Z_{im},\ r_i)^t \tag{16}$$

in which, $\mathbf{x}_i$ denotes the input vector of the $i$th node; $Z_i$ denotes the water level of the $i$th node; $Z_{i1}$, …, $Z_{im}$ denote the water level of the adjacent manholes of the $i$th node; $r_i$ denotes the rainfall.

The output of the SVM is the water level of the manhole at the next time step, i.e.,

$$y_i = Z_i^{t+\Delta t} \tag{17}$$

in which, the superscript $t$ denotes the current time; and $\Delta t$ denotes the time step of the model. The time step is determined by the accuracy of rainfall forecast data and the time interval of measuring.

*3.2.2 Selection of training samples*

Commonly, training samples can be obtained in two ways. The first approach is using historical data of the rainwater pipe network, and the second is using simulated data from hydrodynamic models. The former demands the historical data with a certain amount of water levels of the rainwater pipe network, while the latter one requires a hydrodynamic model with calibrated parameters. Since the real-time water level monitoring of rainwater pipe network is becoming increasingly prevalent in modern cities, historical data is available, and is adapted to generate a training sample set in this paper. Besides, the simulated data are simultaneously used for training samples generation, as some nodes are lack of historical data. The training data should be selected to preserve the diversity as well as the fitting and generalization prediction capabilities of SVM.

*3.2.3 Machine learning steps*

Step1: Generating training samples

Nodes of the rainwater pipe network can be classified as monitored-node and common-node. The monitored-node has the real-time water level monitoring data, while the common-node has no measured data. So both the measured data and the simulated data should be used to generate training samples. For a whole process of rainfall with initial state (i.e. the water levels of all nodes), the training samples can be generated iteratively.

Step2: Conduct normalization processing

The input index should be normalized by

$$u_j^k = U(x_j^k) \tag{18}$$

in which, $u_j^k$ is an index of the normalization; $U$ is normalization function related to the bottom and top elevations of the node. The value of $u$ is bounded from 0 to 1.

It should be noted that the nodal water level will be set as top elevations when overflow occurs, since the measured data is bounded from nodal bottom elevation to top elevation.

Step3: Model training and evaluation

Take Normalized sample $(u_j^k, y^k)$ as input data for the SVM, and select model parameters ($C$, $\varepsilon$) and the type of kernel function, the optimal parameters $(\hat{\boldsymbol{\lambda}}, b)$ can be estimated through training, and then the SVM can be used for water level prediction by equation (14). Finally, the predicted normalization results of SVM model are converted to original values of water level by using the inverse function of Equation (18).

High nonlinearity exists in the water level of the rainwater pipe network. However, since SVM has the ability to deal with infinite or nonlinear features in a high dimensional feature space, as well

as the time step of water level prediction is relatively short, the proposed machine learning method can predict the water level of rainwater pipe network with good accuracy.

## 4. Case Study

In this part, the validity of the SVM model is verified by taking Fuzhou City, China as an example.

### *4.1. The study area*

Fuzhou is the capital of Fujian Province of China (Figure 1). The elevation varies from 0 m to 1000 m, and there are mountains locating in the north of Fuzhou. Until now, mountain floods have passed through Fuzhou to Minjiang River. There are 107 rivers in Fuzhou. The city is separated into two areas, Jiangbei area, and Jiangnan area, by the Minjiang River (Figure 2). The Minjiang River empties into the East China Sea, and is a tidal river that discharges floods of Fuzhou. Due to the heavy rain caused by the typhoon, Fuzhou suffers severe waterlogging disasters in recent years (Figure 3). In recent years, the rainwater pipe network of Fuzhou have been developed rapidly, and the drainage standard is about 1 to 5-year return period rainstorm (Figure 1).

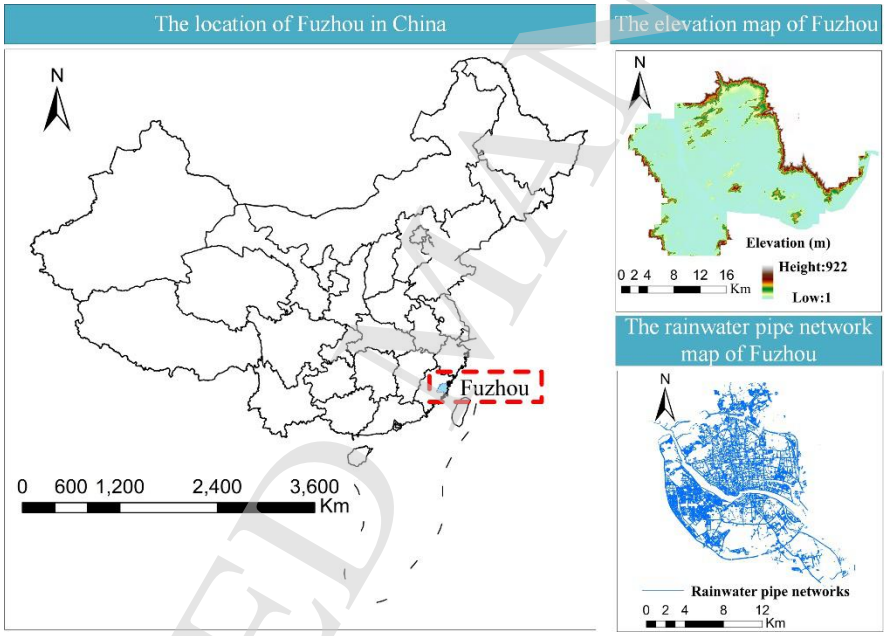



**Figure 1.** Fuzhou's location

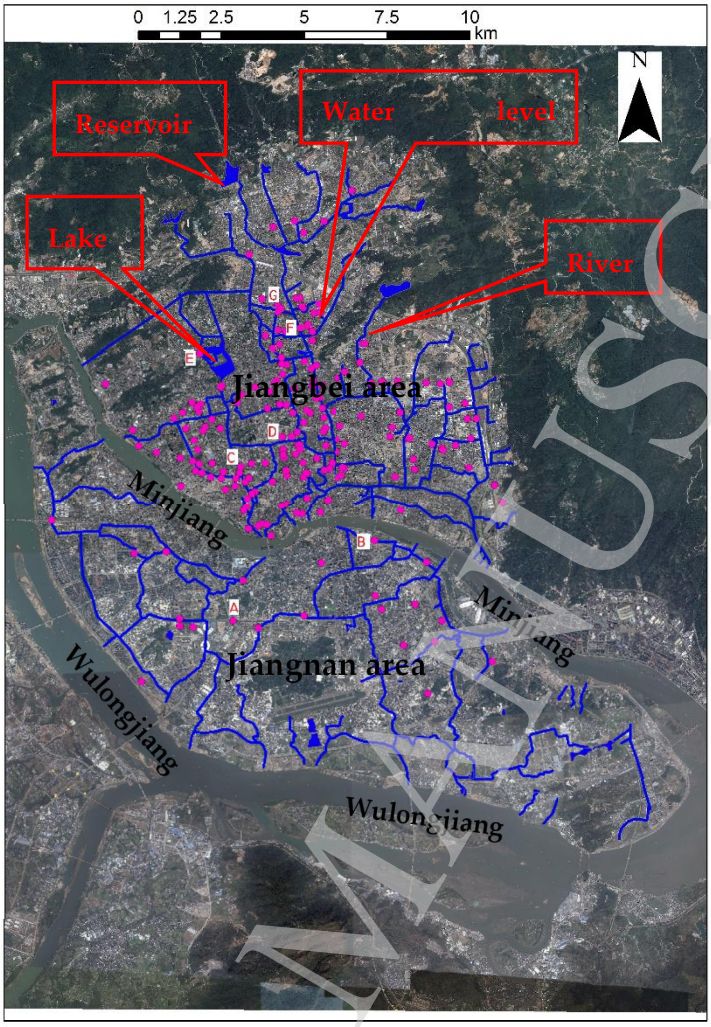



**Figure 2.** The research area: Fuzhou City, China.

**Figure 3.** The historical inundation area in Fuzhou City.

*4.2. Data preprocess*

Real-time data are collected by the water level sensors which are installed under the manhole covers. There are about 220 water level sensors of the rainwater pipe network (Figure 2). The positions of data are identified by the IDs of manhole covers. There are two kinds of data parameters. The first one records data, which contains IDs of manhole covers, time, and water levels; the second one is device data, which contains IDs of sensors, positions, and IDs of manhole covers. So the record data and device data are connected only by the IDs of manhole covers.

To preserve the diversity of the training data as well as the fitting and generalization prediction capabilities of SVM, the training data is selected from historical rainstorm event with different rainfall intensity from 2012 to 2016. For the sake of simplification, historical data from 1 June 2017 to 31 October 2017 is used for model validation in this work, and more than 9, 000, 000 records are available. The original data would be unstructured, and should be preprocessed. For example, in the case of changing equipment, the IDs of manhole covers would be changed, so the original data should be repaired by considering that one position would have multi- IDs of manhole covers. Another particular case is that the time interval of different water level sensors would be variable. So the original water levels records should be interpolated to structured data with the consistent time interval. The minimum time interval of the original data is 5 min. So the unified time interval is chosen as 5 min, and the linear interpolation method is adopted in this work.

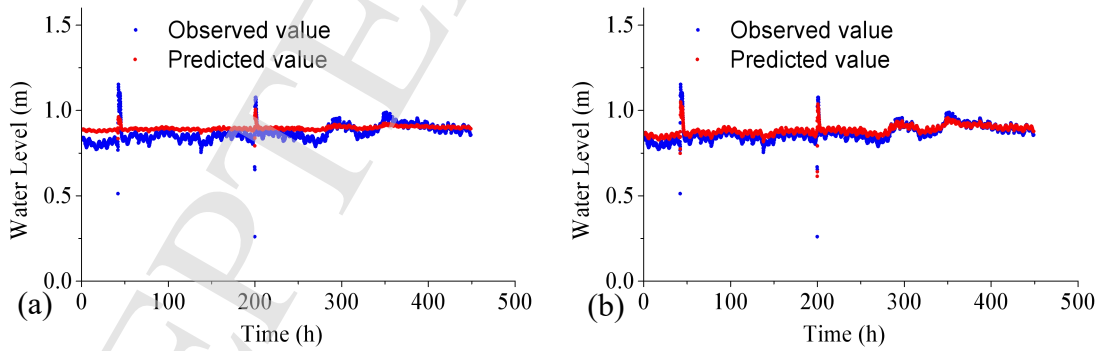
*4.3. Parameters estimation*

As the unified time interval is 5 min, the time step of water level prediction is also set to 5 min. However, the time period of rainfall and water level used by the SVM as input data should be quantitatively analyzed, since there is a time lag between rainfall-induced overland flow and pipe flows. In this work, four cases with different time periods are considered (Table 1).

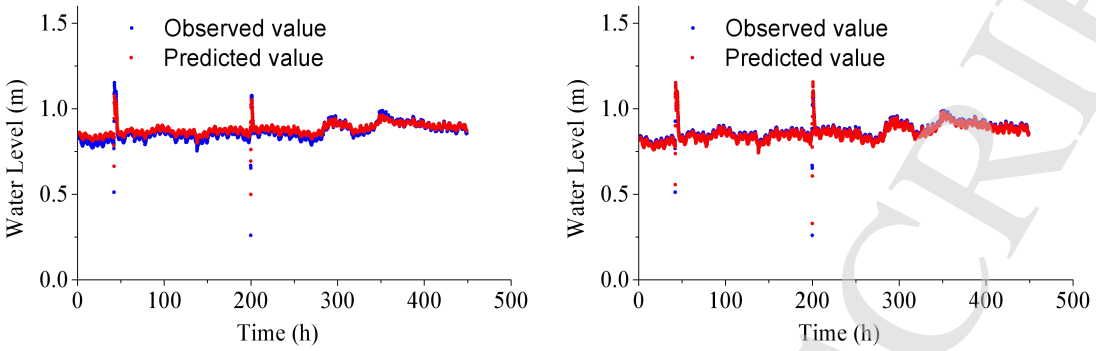
**Table 1.** Four cases with different time periods.

| Time period (min) | Input data | | Predicted data |
|---|---|---|---|
| | rainfall | water level | water level |
| Case 1 | 20 | 10 | 5 |
| Case 2 | 10 | 15 | 5 |
| Case 3 | 20 | 15 | 5 |
| Case 4 | 10 | 10 | 5 |

Figure 4 presents the results comparisons, at station E (Figure 2), between the observed and predicted water levels for the above four cases. From Figure 4 it can be seen that using the past 10 min rainfall and the past 10 min water levels (Case 4) can achieve the highest accuracy for water level prediction in the near future of 5 min. So parameters of Case 4 are adapted for water level prediction in this work.




(a) (b)




(c) (d)

**Figure 4.** Results comparisons between observed and predicted water levels. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4.

*4.4. Results and discussion*

In this work, only results of the water level sensors A-G (Figure 2) are presented for simplification. Figure 5 presents the comparisons between measured and predicted water levels. Table 2 gives the quantitative analysis on accuracy, in which the mean values, the mean absolute error (MAE), the root mean square error (RMSE), the linear correlation coefficient (LCC), and the Nash-Sutcliffe efficiency coefficient (NSE) is given by,

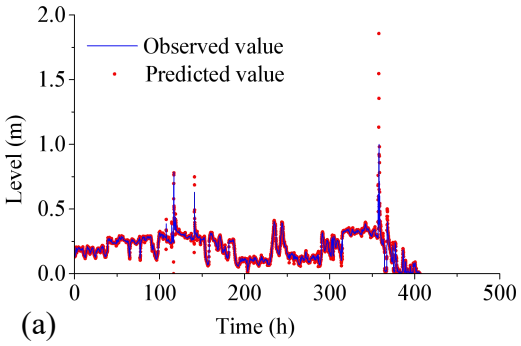
$$MAE = \frac{1}{N}\sum_{i=1}^{N} | x_i - \hat{x}_i | \tag{19}$$

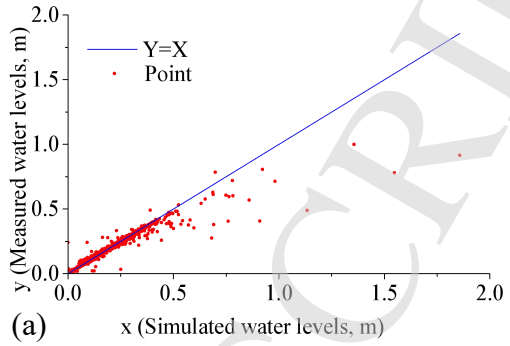
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{x}_i)^2} \tag{20}$$

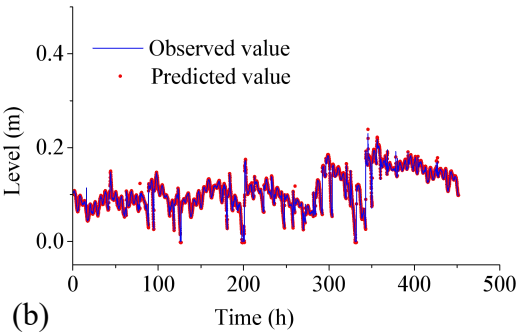
$$LCC = \frac{Cov(x, \hat{x})}{\sqrt{Var[x]Var[\hat{x}]}} \tag{21}$$

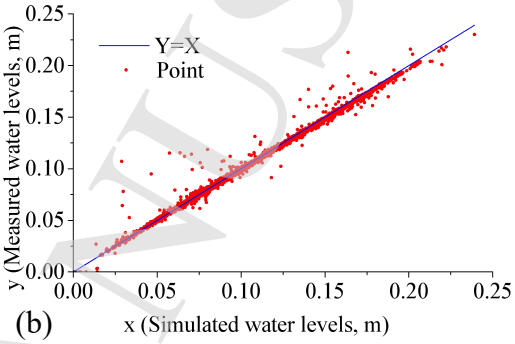
$$NSE = 1 - \frac{\sum_{i=1}^{N} (\hat{x}_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{22}$$

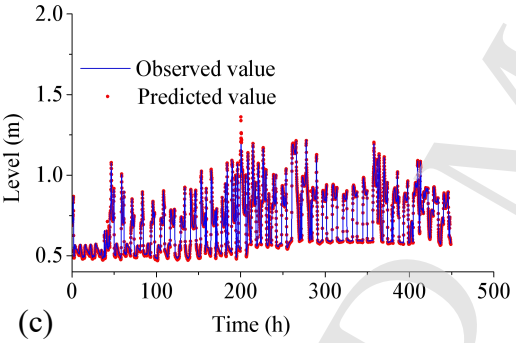
in which, $x$, $\hat{x}$ denote the measured and predicted water levels, respectively; $\bar{x}$ denotes the averaged value of the measured water levels; $N$ denotes the number of records; $C_{ov}$ denotes the covariance; $V_{ar}$ denotes the variance. Those four indexes are widely adopted for quantitative analysis of model accuracy. An NSE approaching to 1 means that the reliability of the model is very well.

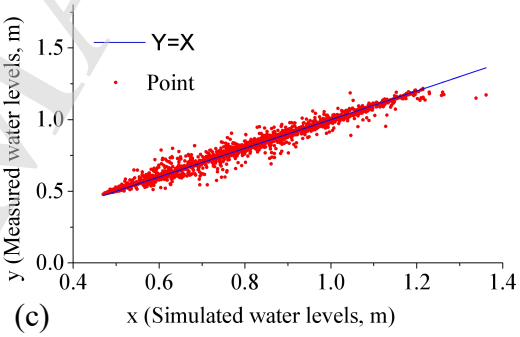


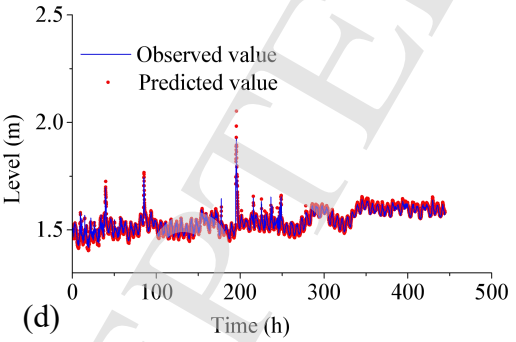



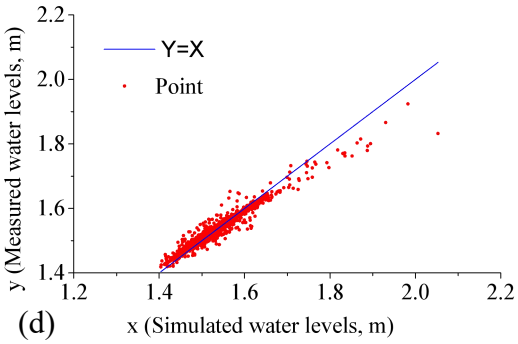



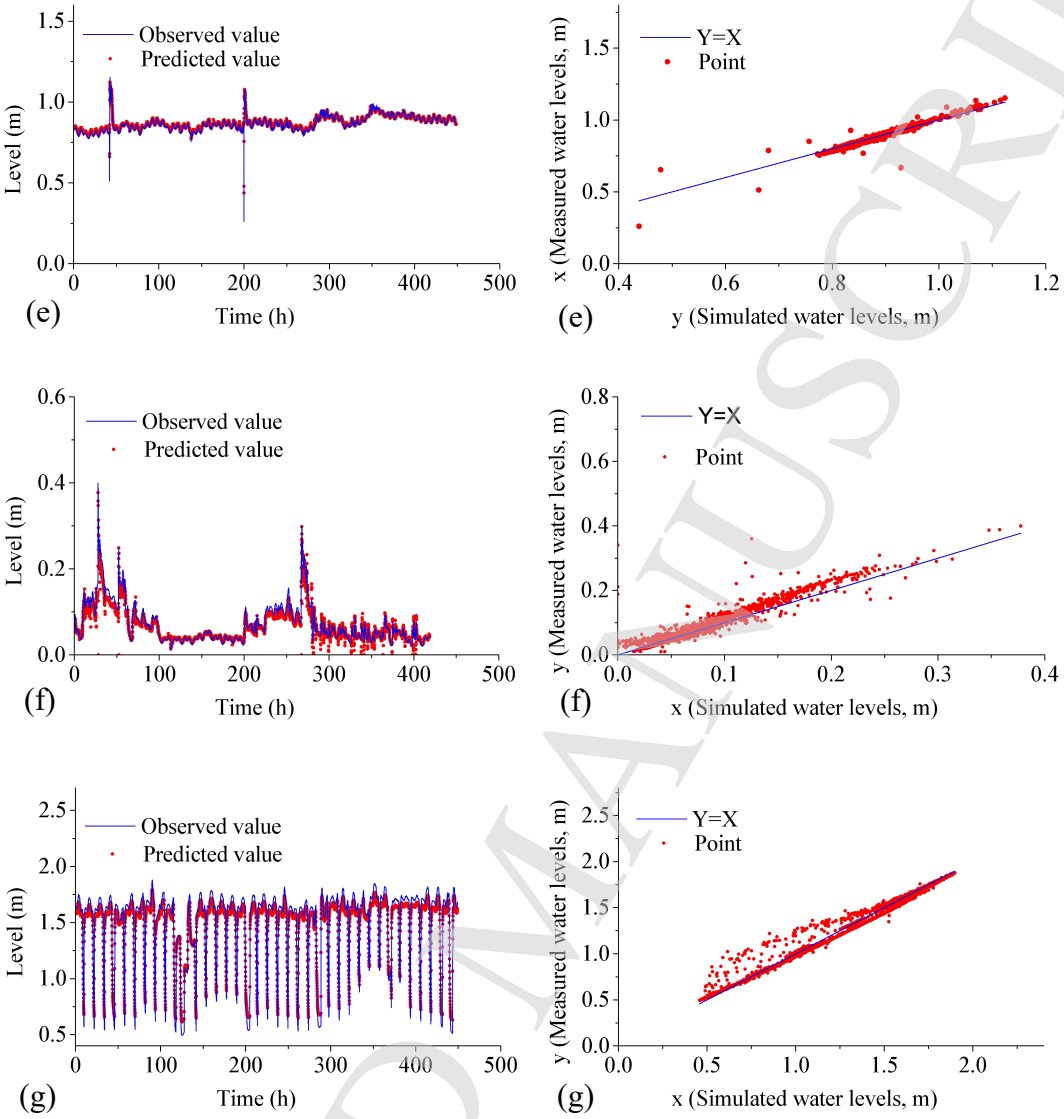



**Figure 5.** The comparisons between measured and predicted water levels. (a)~(g): Water level sensor A~G.

From table 2, it can be seen that the mean values of measured and predicted water level is consistent; the MAE of A-G varies from 1.47 mm to 19.17 mm; the RMSE varies from 3.52 mm to 36.75 mm; the LCC of A-G is close to 1, and the NSE of A-G varies from 0.912 to 0.992. Results show that the suggested machine learning means can predict the water level of the rainwater pipe network with good accuracy.

It should be noted that, Table 2 indicates again that using the past 10 min rainfall and the past 10 min water levels (Case 4) can achieve high accuracy for water level prediction in the near future of 5 min. The main reason might be the time lag between rainfall-induced overland flow and pipe flows, which is determined by the density of manholes, the ground topography, underlying surface condition, and so on.

**Table 2.** The quantitative analysis of model accuracy.

| Position | Mean value(m) | | MAE (equation 19) (mm) | RMSE (equation 20) (mm) | LCC (equation 21) | NSE (equation 22) |
|---|---|---|---|---|---|---|
| | Measured | Predicted | | | | |
| **A** | 0.202 | 0.205 | 5.67 | 27.34 | 0.972 | 0.925 |
| **B** | 0.108 | 0.108 | 1.47 | 3.52 | 0.996 | 0.992 |
| C | 0.704 | 0.700 | 11.12 | 20.38 | 0.994 | 0.986 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **D** | 1.541 | 1.542 | 6.82 | 10.58 | 0.989 | 0.953 |
| **E** | 0.867 | 0.872 | 6.46 | 9.77 | 0.988 | 0.954 |
| **F** | 0.072 | 0.067 | 7.69 | 13.89 | 0.969 | 0.912 |
| **G** | 1.501 | 1.505 | 19.17 | 36.75 | 0.995 | 0.988 |

## 5. Conclusions

In this work, a new framework based on a machine learning method was proposed to achieve a real-time prediction for the water level of rainwater pipe network. The Equation (10) was derived to reveal the spatial and temporal autocorrelation of water levels for adjacent manholes, and then a support vector machine (SVM)-based model was developed by choosing the water levels of adjacent manholes and rivers-near-by-outlets at the last time step as the input indexes.

The proposed framework was applied in Fuzhou City, China. It was found that using the past 10 min rainfall and the past 10 min water levels can achieve the highest accuracy for water level prediction in the near future of 5 min. So the time step of water level prediction is 5 min in this work. Results comparison show that the mean values of measured and predicted water level is in consistent; the MAE of A-G varies from 1.47 mm to 19.17 mm; the RMSE varies from 3.52 mm to 36.75 mm; the LCC of A-G is close to 1, and the NSE of A-G varies from 0.912 to 0.992. So, it can be concluded that the proposed machine learning method can predict the water level of the rainwater pipe network with good accuracy.

It's worth noting in the current work that the suggested model, was applied to predict water levels in the near future of 5 min. Indeed, 5 min would be too short for flood rescue; however, it would be useful for city traffic dispatch management. To prolong forecast period, accurate rainfall forecast data should be used, and this is the further work that we plant to do in the near future. Besides, a wavelet decomposition approach, one of pre-processing data methods, would be required to increase the model accuracy using sub-time series decomposed from original input and target time-series .

**Acknowledgments:** This work was funded by a grant from the National Natural Science Foundation of China (Project No. 51809297).

**Author Contributions:** Conceptualization, Lixiang Song; Data curation, Hao Wang; Formal analysis, Hao Wang; Funding acquisition, Lixiang Song; Investigation, Hao Wang; Methodology, Lixiang Song; Project administration, Hao Wang; Resources, Hao Wang; Software, Hao Wang; Supervision, Hao Wang; Validation, Hao Wang; Visualization, Hao Wang; Writing – original draft, Hao Wang; Writing – review & editing, Hao Wang.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Mailhot A, Duchesne S, Caya D, et al. Assessment of future change in intensity-duration-frequency (IDF) curves for southern quebec using the canadian regional climate model. Journal of Hydrology, 2007, 347(1–2): 197-210.
2. Pan A, Hou A, Tian F, et al. Hydrologically enhanced distributed urban drainage model and its application in Beijing City. Journal of Hydrologic Engineering, 2012, 17(6): 667-678.
3. Yang L, Smith J, Baeck M, et al. Flash flooding in small urban watersheds: Storm event hydrologic response. Water Resources Research, 2016, 52: 4571-4589.
4. Elfeki A, Masoud M, Niyazi B. Integrated rainfall–runoff and flood inundation modeling for flash flood risk assessment under data scarcity in arid regions: Wadi Fatimah basin case study, Saudi Arabia. Natural Hazards, 2017, 85: 87-109.
5. Song L, Zhou J, Guo J, et al. A robust well-balanced finite volume model for shallow water flows with wetting and drying over irregular terrain. Advances in Water Resources, 2011, 34(7): 915-932.
6. Guinot V, Sanders B, Schubert J. Dual integral porosity shallow water model for urban flood modeling. Advances in Water Resources, 2017, 103: 16-31.

7. Singh J, Altinakar M, Ding Y. Numerical modeling of rainfall-generated overland flow using nonlinear shallow-water equations. ASCE Journal of Hydraulic Engineering, 2015, DOI: 10.1061/(ASCE)HE.1943-5584.0001124.
8. Sanders B, Schubert J, Detwiler R. ParBreZo: a parallel, unstructured grid, Godunov-type, shallow-water code for high-resolution flood inundation modeling at the regional scale. Advances in Water Resources, 2010, 33(12): 1456-1467.
9. Lai W, Khan A A. A parallel two-dimensional discontinuous galerkin method for shallow-water flows using high-resolution unstructured meshes. Journal of Computing in Civil Engineering, 2016, 31(3): 04016073.
10. Liang Q, Xia X, Hou J. Catchment-scale high-resolution flash flood simulation using the GPU-based technology. Procedia Engineering, 2016, 154: 975-981.
11. Zhang S, Yuan R, Wu Y, et al. Implementation and efficiency analysis of parallel computation using OpenACC: a case study using flow field simulations. International Journal of Computational Fluid Dynamics, 2016, 30(1): 79-88.
12. Zhang S, Yuan R, Wu Y, et al. Parallel computation of a dam-break flow model using OpenACC applications. Journal of Hydraulic Engineering, 2017, 143(1): 04016070.
13. Hu X, Song L. Hydrodynamic modeling of flash flood in mountain watersheds based on high-performance GPU computing. Natural Hazards, 2018, 91(2): 567-586.
14. Chang T J, Wang C H, Chen A S. A novel approach to model dynamic flow interactions between storm sewer system and overland surface for different land covers in urban areas. Journal of Hydrology, 2015, 524: 662-679.
15. ASCE Task Committee. Artificial neural networks in hydrology. I: Preliminary concepts. Journal of Hydrologic Engineering , 2000, 5: 115-123.
16. Tayfur G, Singh V P, Moramarco T, et al. Flood hydrograph prediction using machine learning methods. Water, 2018, 10, 968; doi:10.3390/w10080968.
17. Lima A R, Hsieh W W, Cannon A J. Variable complexity online sequential extreme learning machine, with applications to streamflow prediction. Journal of Hydrology, 2017, 555: 983-994.
18. Lima A R, Cannon A J, Hsieh W W. Forecasting daily streamflow using online sequential extreme learning machines. Journal of Hydrology, 2016, 537: 431-443.
19. Rasouli K, Hsieh W W, Cannon A J. Daily streamflow forecasting by machine learning methods with weather and climate inputs. Journal of Hydrology, 2012, 414-415: 284-293.
20. Agarwal A, Mishra S K, Ram S, Singh J K. Simulation of runoff and sediment yield using artificial neural networks. Biosystems Engineering, 2006, 94(4): 597-613.
21. Mustafa M R, Isa M H, Rezaur R B. Artificial neural networks modeling in water resources engineering: Infrastructure and applications. International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering, 2012, 6(2): 128-136.
22. Nayebi M, Khalili D, Amin S, Zand-Parsa S. Daily stream flow prediction capability of artificial neural networks as influenced by minimum air temperature data. Biosystems Engineering, 2006, 95(4): 557-567.
23. Gong Y, Zhang Y, Lan S, Wang H. A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. Water Resources Management, 2016, 30(1): 375-391.
24. Misra D, Oommen T, Agarwal A, Mishra S K, Thompson A M. Application and analysis of support vector machine based simulation for runoff and sediment yield. Biosystems Engineering, 2009, 103: 527-535.
25. Yu P S, Chen S T, Chang I F. Support vector regression for real-time flood stage forecasting. Journal of Hydrology, 2006, 328: 704-716.
26. Noymanee J, Nikitin N O, Kalyuzhnaya A V. Urban Pluvial Flood Forecasting using Open Data with Machine Learning Techniques in Pattani Basin. Procedia Computer Science, 2017, 119: 288-297.
27. Zhang X, Zhou J, Wang C, Li C, Song L. Multi-class support vector machine optimized by inter-cluster distance and self-adaptive deferential evolution. Applied Mathematics & Computation, 2012, 218(9): 4973-4987.
28. Lewis A. Rossman L A. Storm Water Management Model Reference Manual Volume II – Hydraulics. U.S. Environmental Protection Agency, 2017, EPA/600/R-17/111.
29. Huang Z, Zhou J, Song L, et al. Flood disaster loss comprehensive evaluation model based on optimization support vector machine. Expert Systems with Applications, 2010, 37(5): 3810-3814.

30. Chen S, Yu P. Pruning of support vector networks on flood forecasting. Journal of Hydrology, 2007, 347(1-2): 67-78.
31. Huang Z, Zhou J, Song L, et al. Improved Support Vector Machine and its Application. Applied Mechanics & Materials, 2010, 20-23: 147-153.
32. Joachims, Thorsten. Making large-scale support vector machine learning practical. Advances in Kernel Methods 1999.
33. Tong Z , Xiao Z , Liu H , et al. A Novel Parallel LSA-SVM Algorithm Based on Semantic Distance for Blog. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(09):18.
34. Famouri M , Taheri M , Azimifar Z . Fast Linear SVM Validation Based on Early Stopping in Iterative Learning. International Journal of Pattern Recognition and Artificial Intelligence, 2015:1551013.
35. Alimjan G , Sun T , Jumahun H , et al. A Hybrid Classification Approach Based on Support Vector Machine and K-Nearest Neighbor for Remote Sensing Data. International Journal of Pattern Recognition and Artificial Intelligence, 2017, 31(10).
36. Alimjan G , Sun T , Liang Y , et al. A New Technique for Remote Sensing Image Classification Based on Combinatorial Algorithm of SVM and KNN. International Journal of Pattern Recognition and Artificial Intelligence, 2017:1859012.

Hao Wang, Ph.D, research on urban drainage system optimization, email: wanghao123612@163.com.

Lixiang Song, Ph.D, research in computational hydrodynamics. Email: slx.hust@live.cn