Hindawi
Advances in Meteorology
Volume 2017, Article ID 5019646, 17 pages
https://doi.org/10.1155/2017/5019646

Research Article

Applications of Cluster Analysis and Pattern Recognition for Typhoon Hourly Rainfall Forecast

Fu-Ru Lin,[1] Nan-Jing Wu,[2] and Ting-Kuei Tsay[1]

[1] *Department of Civil Engineering, National Taiwan University, Taipei 10617, Taiwan*
[2] *Department of Civil and Water Resources Engineering, National Chiayi University, Chiayi 60004, Taiwan*

Correspondence should be addressed to Nan-Jing Wu; njwu@mail.ncyu.edu.tw

Received 30 June 2016; Revised 8 January 2017; Accepted 7 February 2017; Published 21 March 2017

Academic Editor: Soni M. Pradhanang

Based on the factors of meteorology and topography, it is assumed that there exist some certain patterns in spatial and temporal rainfall distribution of a watershed. A typhoon rainfall forecasting model is developed under this assumption. If rainfall patterns can be analyzed and recognized in terms of individual watershed topography, only the spatial rainfall distribution prior to a specific moment is needed to forecast the rainfall in the next coming hours. It does not need any other condition in meteorology and climatology. Besides, supplement techniques of missing rainfall gage data are also considered to build an all-purpose forecast model. By integrating techniques of cluster analysis and pattern recognition, present proposed rainfall forecasting model is tested using historical data of Tamsui River Basin in Northern Taiwan. Good performance is validated by checking on coefficient of correlation and coefficient of efficiency.

1. Introduction

Typhoon rainfall forecast is extremely important since it is the basic requirement in flood routing simulation using a hydrologic model, allowing an extension of the lead-time of the river flow forecasting computations. It is particularly needed in small- and medium-sized mountainous basins [1]. In Taiwan, due to the high mountains and steep river slopes, heavy rainfalls, especially during typhoon events, have frequently led to serious disasters, such as flooding, landslide, or debris flows. In order to reduce loss of life and major economic impacts, the government has invested a great deal of manpower and budgets to build the disaster warning systems in which rainfall forecast plays a key role. It provides rainfall input data to forecast the surface runoff outflow of a watershed. This outflow or the gaged water depth at the outlet of the watershed is also needed as the information for the upstream boundary condition of unsteady river flow computations [2–4]. Quite often, whenever a typhoon has occurred, undesired conditions may occur when gaged rainfall data do not transmit into database system at all for further computational uses. Furthermore, lack of immediate rainfall data may affect the accuracy in real-time flood forecasting or other systems. In order to deal with such situation, the authority should not only assure the stability of an observation system and its transmission instruments but also build an all-purpose rainfall forecast model to manage the situation of lost data at any moment and provide reasonably accurate and efficient forecast data.

Traditionally, rainfall forecasting is based mainly on numerical fluid dynamic models [5]. This classical approach attempts to model the fluid and thermal dynamic systems for grid-point time series prediction based on boundary meteorological data. The simulation often requires intensive computations involving complex differential equations and computational algorithms. Besides, the accuracy is bounded by certain constraints such as the adoption of incomplete boundary conditions, model assumptions, grid resolutions, and numerical instabilities. Furthermore, because of the high variability in space and time, typhoon rainfall is one of the most difficult elements for the hydrologic cycle to forecast. The highly nonlinear and extremely complex physical process of typhoon rainfall also leads to a lot of difficulties in constructing a physically based mathematical model [6].

Radar data and satellite images were also used to forecast the rainfall [7, 8]. Unfortunately, the relationship between rainfall and the outputs from satellite and radar images was not clear while the outputs do not allow a satisfactory assessment of rain intensities [1]. Another reason was that due to ground occultation and altitude effects, the radar detection was particularly difficult in mountainous regions [9, 10].

In recent decades, the research using artificial intelligence has gained scientific attention. The Artificial Neural Network (ANN) is one of the most representatives of these achievements. Researches using ANNs were sequentially reported. Luk et al. [11] assumed that the spatial rainfall distribution at a specific moment is bounded with the records of the relevant rainfall gages in the lasted time interval. By using a backward propagation neural network (BP-ANN), they successfully built a model for forecasting the rainfall pattern in the next coming 15 minutes. The same concept was used to build another rainfall forecasting model by applying other kinds of neural networks such as feedforward neural network, partial recurrent network, and time delayed neural network [12]. Toth et al. [1] compared the accuracy of the short-term rainfall forecasts obtained with three time series analysis techniques, such as linear stochastic autoregressive moving average (ARMA) models, artificial neural networks (ANNs), and the nonparametric nearest-neighbors method. Chang et al. [13] compared and discussed three types of multistep-ahead (MSA) methods using previous rainfall and river stage for flood forecasting. Lin et al. [14] used a novel kind of neural network called support vector machines (SVMs) to construct typhoon rainfall forecasting models. They used these models with and without typhoon characteristics to forecast the rainfall. Because all the rainfall or flood forecasting models mentioned above regard the gaged rainfall records in the last period of time as the input data, these models might not work properly when data gaps occur. The model could not carry on further computations unless the lost data can be estimated correctly.

When a storm or frontal surface is approaching, the rainfall patterns in the windward area may be quite different from those in the leeward area, due to the topographical effects. As the storm or frontal surface moves during the typhoon period, the rainfall patterns may alter drastically at a specific gage location. This implies that the spatial and temporal distribution of the rainfall are influenced by some information of meteorology and topography. Because the topography does not change with time and also storms or frontal surfaces usually move along some certain paths, the trends of spatial-temporal rainfall distribution could be bounded within some specific patterns. Based on the consideration of these meteorological and climatological factors, it is assumed in this paper that there exist certain patterns in spatial and temporal rainfall distribution for a particular river basin. An unsupervised pattern recognition method, which has powerful ability of fault tolerance, is applied. The clustered construction can identify the coordinate data from the remainder data even if the input data are incomplete or have data gaps. The results from the recognized patterns are model outputs. These model outputs are used as the input for river runoff or elevation forecasting at the outlet of the basin. This paper brings up the pattern recognition and cluster analysis in statistics to classify the rainfall distribution in space and time from historical data of similar meteorological and climatological conditions. This study intends to build an all-purpose model with good accuracy and reliability for typhoon hourly rainfall forecast. The model holds good for its design function even with data gaps in rainfall data.

2. Methodology

2.1. Cluster Analysis. Cluster analysis is the general logic, formulated as a procedure, by which we objectively group the entities together on the basis of their similarities and differences [15]. The objective of data clustering is to employ certain clustering algorithms to identify clusters consisting of similar data within a dataset. The original dataset is thus decomposed into disjoint clusters, with each cluster having a center to represent the cluster. We can use the cluster centers to represent the original dataset to achieve the following two goals, namely, data compression, and computation reduction. In general, clustering algorithms can be divided into two types: (1) hierarchical clustering and (2) nonhierarchical clustering (or called partition clustering). Two sorts of hierarchical clustering could be found. They are agglomerative and divisive ones. For agglomerative hierarchical clustering, the number of clusters is increased from one until the desired number of clusters is reached. On the other hand, for divisive hierarchical clustering, the number of clusters is decreased from the size of the dataset until the desired number of clusters is reached. For nonhierarchical clustering approaches, the number of clusters is fixed in advance. And then a number of iterations are performed to identify the best clusters with their cluster centers [16].

Many empirical results indicate that the point of adding nonrandomly selected, nonhierarchical clustering method is better than the hierarchical clustering method [17]. Meanwhile, in nonhierarchical clustering the number of clusters should be predetermined and its starting from a randomly initial partition may cause optimization locally. Therefore, some algorisms such as two-stage cluster or two-step cluster were developed by using one or two algorisms above to increase their advantage and decrease their shortcoming. The Statistical Product and Service Solutions (SPSS) two-step cluster will be used in this paper, and below is mainly drawn from "the support document of SPSS and IBM knowledge center" [18, 19], for completeness. The SPSS two-step clustering component is a scalable cluster analysis algorithm designed to handle very large datasets and is well-known for recent years. The procedure of the cluster is divided into two steps. In the first step, the records were preclustered into many small subclusters by a sequential clustering approach. Thus, the records were scanned one by one and decided if the current record should merge with the previously formed clusters or start a new cluster based on the distance criterion. A modified cluster feature (CF) tree which consists of levels of nodes was implemented. In the second step, subclusters resulting from the first step were taken as input and then were grouped into the desired number of clusters by agglomerative hierarchical clustering method.

2.2. Pattern Recognition. The concept of "recognition" comes from the main theory of artificial neural networks. When new input data comes out, one can determine the category and the output corresponding to that category immediately. The network structure requires powerful ability of fault tolerance. A clustered construction, even if the input data is incomplete, can still identify the coordinate data from the remainder data and show which category it belongs to. The key point of pattern recognition in this study is the winner-take-all (WTA) network. For a group of artificial neurons, the neurons compete with each other. The weight is given as 1 to the winner neuron, the one who is closest to the input data, and 0 to all others. This process is known as the winner-take-all.

In this paper, a "pattern" is a multivariable time-space series. The rainfall record of some lasted time interval at a specific moment of several gages is combined as an input vector and the dataset collected from numerous storm events is divided into some specific groups. This way, not only the characteristics of rainfall within the space, such as topography (windward, leeward, altitude, etc.), but also the "behavior" that they change over time, can be obtained. With these procedures, a model of typhoon rainfall forecast can be established. The so-called "pattern" is referred to as the rainfall distribution in time and space with respect to a certain typhoon category, and "recognition" is the information available to the corresponding classification categories.

Assume that there is a group of statistical samples. Each sample is composed of n values and expressed as a mathematical vector of n components:

$$\vec{x}_i = \left[x_{i,1} \ \ x_{i,2} \ \ \cdots \ \ x_{i,n}\right]^T, \tag{1}$$

where i is the serial number of a specific sample.

Firstly, the cluster analysis is preceded. In order to divide these samples into several certain patterns, the neural network structure of winner-take-all (WTA) is employed to describe the distribution of samples. The pattern which any specific sample belongs to can be expressed as

$$P\left(\vec{x}_i\right) = \sum_{j=1}^{m} j f\left(\left\|\vec{x}_i - \vec{c}_j\right\|\right), \tag{2}$$

where $P(\vec{x}_i)$ is a natural number that expresses the pattern to which the ith sample belongs, m denotes the numbers of classification, and $f(\|\vec{x}_i - \vec{c}_j\|)$ is a binary function, which is the radial basis function (RBF) used in WTA neural network

$$f\left(\left\|\vec{x}_i - \vec{c}_j\right\|\right) = \begin{cases} 1 & \text{if } \left\|\vec{x}_i - \vec{c}_j\right\| = \min\left\{\left\|\vec{x}_i - \vec{c}_k\right\|, \ k = 1, 2, \ldots, m\right\} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\|\vec{x}_i - \vec{c}_j\| = \sqrt{\sum_{l=1}^{n}(x_{i,l} - c_{j,l})^2}$ in which $x_{i,l}$ denotes the lth component of $\vec{x}_i$ and $\vec{c}_j$ represents the center of the jth cluster, which resulted from the approach of two-step clustering (see Section 2.1):

$$\vec{c}_j = \left[c_{j,1} \ \ c_{j,2} \ \ \cdots \ \ c_{j,n}\right]^T. \tag{4}$$

After completion of classifying the statistical samples, for any new input, $\vec{x}$, one can find which pattern it belongs to by checking with this formula:

$$P\left(\vec{x}\right) = \sum_{j=1}^{m} j f\left(\left\|\vec{x} - \vec{c}_j\right\|\right). \tag{5}$$

Furthermore, a relation between the input data and output data needs to be constructed. Consider an output data $\vec{y}_i$ which is composed of n_{out} values; each $\vec{x}_i$ corresponds to a specific $\vec{y}_i$. The output vector is expressed as

$$\vec{y}_i = \left[y_{i,1} \ \ y_{i,2} \ \ \cdots \ \ y_{i,n_{\text{out}}}\right]^T. \tag{6}$$

Here i is the serial number of a specific sample as previously defined. After all the samples have been clustered, one can find the kth component of the output vector $\vec{y}$ corresponded to an input vector $\vec{x}$ by the following formula:

$$y_k \cong \hat{y}_{P(\vec{x}),k} = \sum_{j=1}^{m} \hat{y}_{j,k} f\left(\left\|\vec{x} - \vec{c}_j\right\|\right), \tag{7}$$

where $\hat{y}_{j,k}$ represents the kth component of the jth output pattern. If each sample belongs to a certain cluster and the distances among them are very small, one can determine $\hat{y}_{j,k}$ by using the average value to represent the whole values of output data:

$$\hat{y}_{j,k} = \frac{\sum_{i=1}^{N} y_{i,k} f\left(\left\|\vec{x}_i - \vec{c}_j\right\|\right)}{\sum_{i=1}^{N} f\left(\left\|\vec{x}_i - \vec{c}_j\right\|\right)}, \tag{8}$$

where N is the total number of samples.

When new data are added, one can find the cluster centers as described previously, identify to which pattern this sample belongs, and may use the relationship between input and output to predict the corresponding output.

2.3. Model Setup. In practice, the input data, $\vec{x}_i$, are composed of spatial and temporal information and can be expressed as follows:

$$\vec{x}_i = \left[p_1(t) \ \ p_2(t) \ \ \cdots \ \ p_{n_R}(t) \ \ p_1(t-1) \ \ p_2(t-1) \ \ \cdots \ \ p_{n_R}(t-1) \ \ \cdots \ \ p_1(t-n_l) \ \ p_2(t-n_l) \ \ \cdots \ \ p_{n_R}(t-n_l)\right]^T, \tag{9}$$

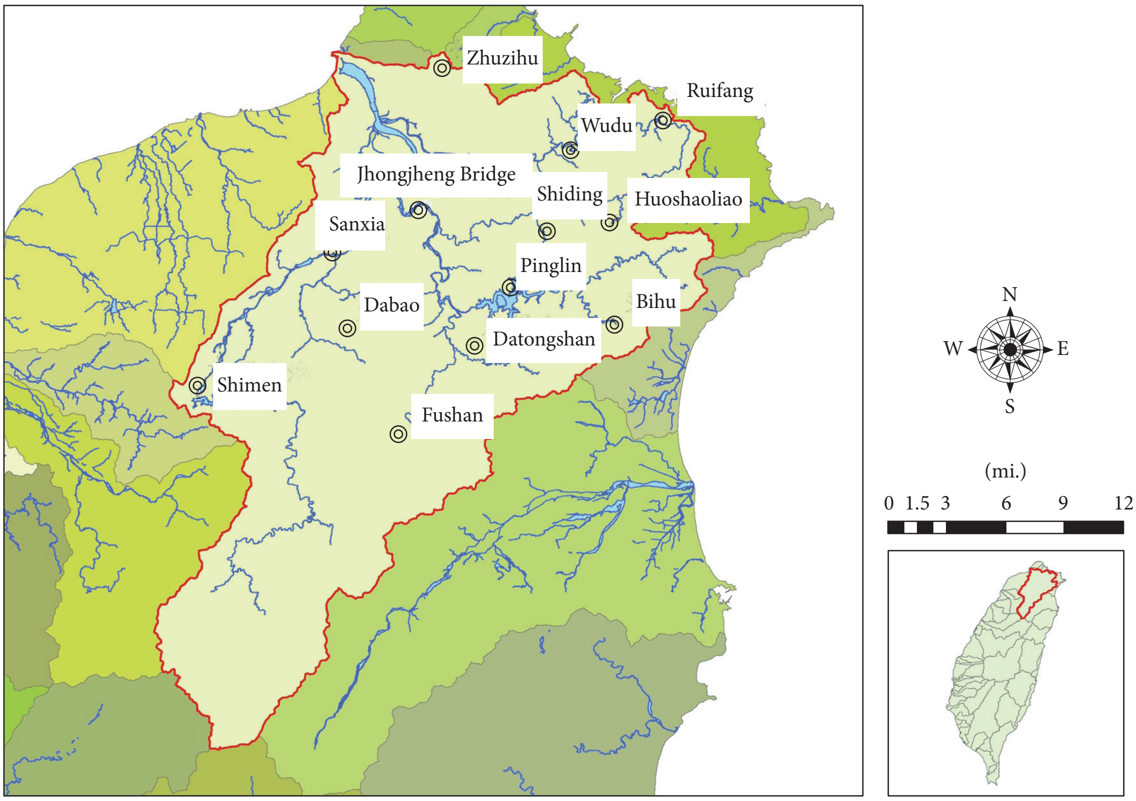


Figure 1: Location of Tamsui River Basin and the 13 rain gages.

where i is the serial number of a specific sample and n_l is the number of time steps considered in the input pattern. The subscript of p (i.e., $1, 2, \ldots, n_R$) is the serial number of the rain gage while $p_1(t)$ denotes the rainfall data at time t in rain gage number 1 and $p_1(t - n_l)$ denotes the rainfall data at previous n_l time steps in the same rain gage. The values of $p_2(t)$, $p_2(t-1), \ldots, p_2(t-n_l)$, $p_3(t-1), \ldots, p_{n_R}(t-n_l)$ are all defined in a similar way. And the output data, $\vec{y}_i$, can be expressed as follows:

$$\vec{y}_i = \left[p_1(t+1) \quad p_2(t+1) \quad \cdots \quad p_{n_R}(t+1)\right]^T. \qquad (10)$$

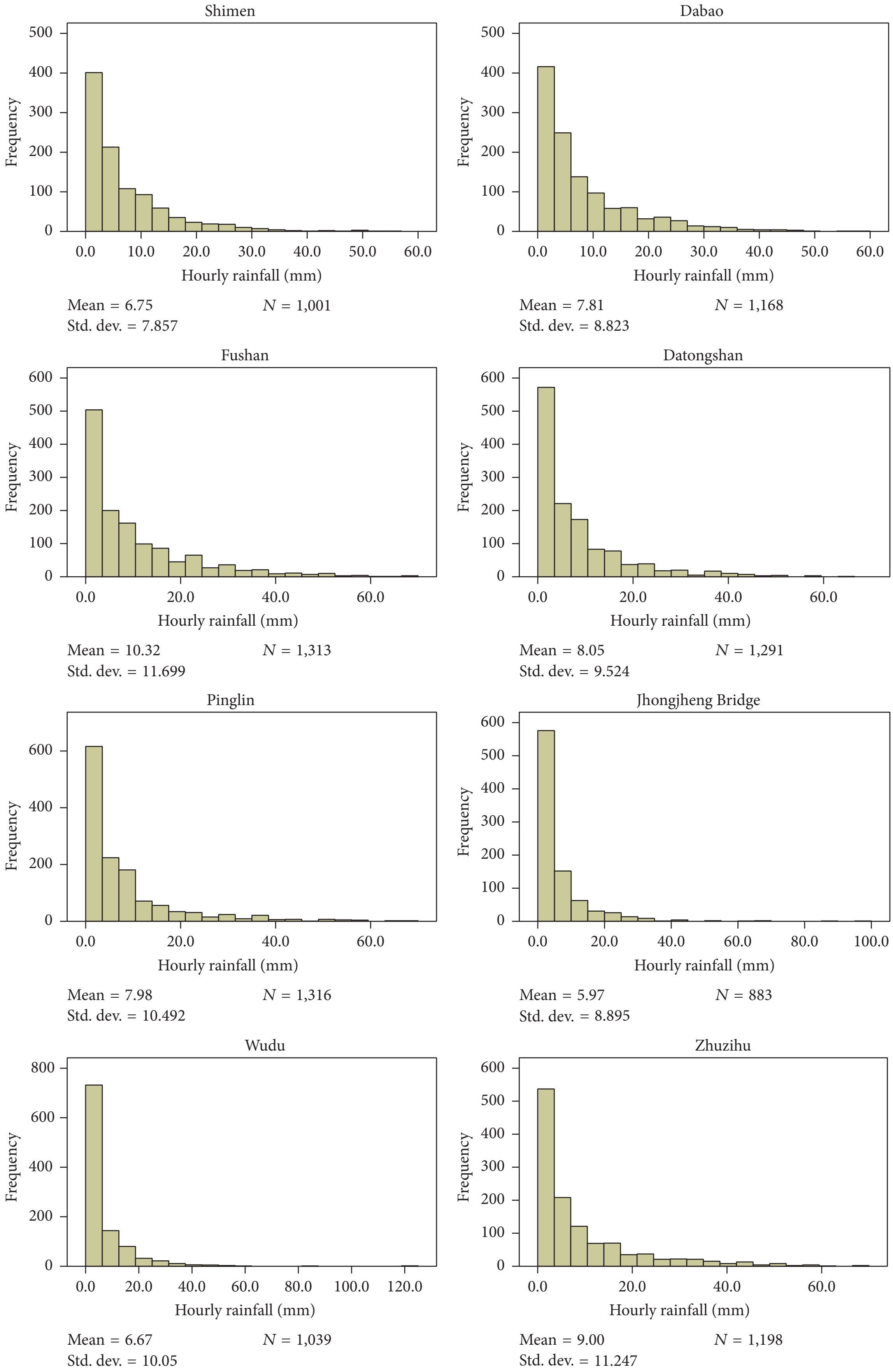
In this paper, rainfall records of last (specifically, $n_l = 1$) and present hour are used as the input data to forecast gaged rainfall in the next hour. So, $\overrightarrow{x}_i$ contains $2 \times n_R$ components and $\vec{y}_i$ contains n_R components.

3. Applications

3.1. Study Area. In this paper, the feasibility of this method is tested to the rainfall forecasting in Tamsui River Basin in the Northern Taiwan. Tamsui River runs through Taipei, the capital city of Taiwan, and has a total drainage area of approximately 2726 km^2. Due to the peculiar topography, the three mainly tributaries, Keelung River, Dahan Stream, and Sintain Stream, converge in Taipei Basin in which there usually are severe damage during storms and typhoons. Because of concentration of population (population 6.5×10^6) and developed urban and suburban areas, government has invested a great deal of manpower and budgets to build the flood warning system. So there are abundant historical observations of rainfall data. However, when typhoon occurs, the transmittal system becomes poor, resulting in missing rainfall data. Furthermore, lack of immediate rainfall data may affect the accuracy in flood forecasting. In order to deal with this situation, one should not only ensure the stability of observation system and transmission instrument but also build an all-purpose forecast model to manage the situation of lost data at any moment.

There are many rain gages in Tamsui River Basin. Some of them, belonging to Water Resources Agency, are operationally stable and experience fewer situations of lost data. Therefore, in this paper hourly rainfall data of these rain gages are used to forecast the gaged rainfall in the next hour. There are total 16 rainfall gages in Tamsui River Basin which belong to Water Resources Agency. Three of them were set up after 2001; the other 13 gages have more than 20 years of historical data. Locations of Tamsui River Basin and these 13 rain gages are shown in Figure 1. Frequency diagrams and information of hourly rainfall of the rain gages in Tamsui River Basin during typhoon events are shown in Figure 2.

3.2. Calibration and Validation of Dataset. After removing the events with incomplete data, total of 32 typhoon events which occurred during 1995–2015 were analyzed for this



Mean = 6.75 N = 1,001
Std. dev. = 7.857

Mean = 7.81 N = 1,168
Std. dev. = 8.823

Mean = 10.32 N = 1,313
Std. dev. = 11.699

Mean = 8.05 N = 1,291
Std. dev. = 9.524

Mean = 7.98 N = 1,316
Std. dev. = 10.492

Mean = 5.97 N = 883
Std. dev. = 8.895

Mean = 6.67 N = 1,039
Std. dev. = 10.05

Mean = 9.00 N = 1,198
Std. dev. = 11.247

Figure 2: Continued.

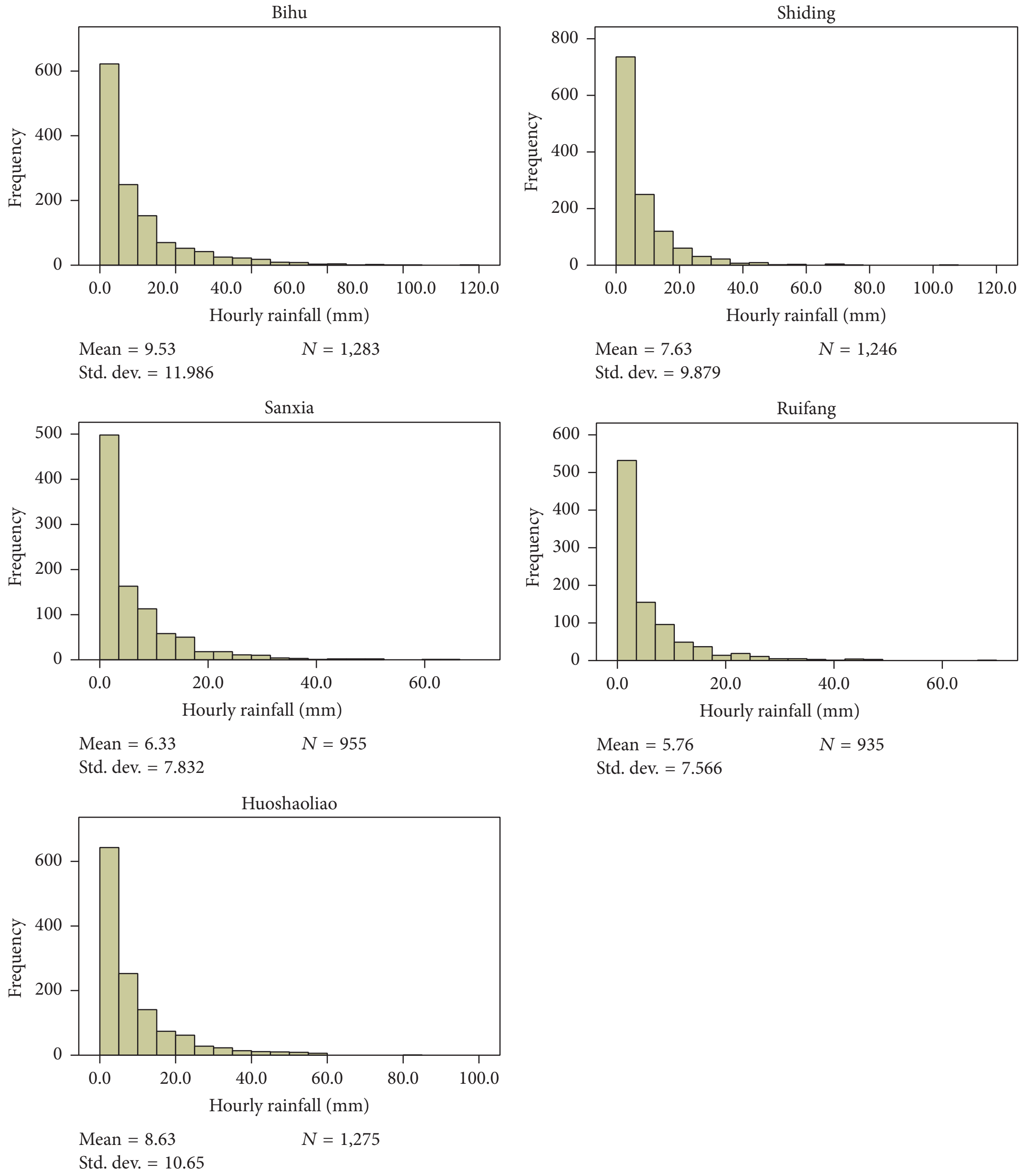


FIGURE 2: Frequency diagrams and information of hourly rainfall of the rain gages in Tamsui River Basin during typhoon events.

study. Each of them caused intense rainfall in Tamsui River Basin when the typhoon centers passed the vicinity of Northern Taiwan. The collected events are separated into two sets of data, calibration and validation, as listed in Table 1. There are 2175 samples for calibration. Dataset was entered into SPSS by using two-step clustering component for classification and log-likelihood option was chosen as the distance measurement. The number of clusters is determined by how many samples we have and at least how many samples should be in a cluster. Although choosing more categories may produce more accurate results, overfitting could also happen if the dataset is divided into too many clusters. Considering each cluster should be 10 samples at least and clusters are as many as possible, the dataset is divided into 22 clusters. Thus, twenty-two clusters are chosen for the number of clusters in this application. Table 2 shows the result of the classification. In this paper, the hourly rainfall data of Typhoon Soudelor (2015) and Typhoon Dujuan (2015) were chosen as the validation dataset.

3.3. Supplement of Missed Rainfall Data. The model has capacity to automatically fill in any missing data within the

Table 1: The list of typhoon events collected in this study for the model establishment.

Year	Event	Data condition	Choose or not		
1996	HERB	●	◎	●	With complete data
1997	WINNIE	●	◎	○	Some data lost
1997	AMBER	●	◎	◎	Choose
1998	YANNI	○			
1998	ZEB	○			
2000	KAI-TAK	○			
2000	BILIS	●	◎		
2000	PRAPIROON	○			
2000	XANGSANE	●	◎		
2001	TORAJI	●	◎		
2001	NARI	●	◎		
2002	SINLAKU	●	◎		
2004	MINDULLE	●	◎		
2004	AERE	○	◎ validation		
2005	HAITANG	●	◎		
2005	MATSA	●	◎		
2005	TALIM	●	◎		
2005	LONGWANG	●	◎		
2006	BILIS	●	◎		
2006	KAEMI	●	◎		
2007	WUTIP	●	◎		
2007	SEPAT	●	◎		
2007	WIPHA	●	◎		
2007	KROSA	●	◎		
2008	KALMAEGI	●	◎		
2008	FUNG-WONG	●	◎		
2008	SINLAKU	●	◎		
2008	JANGMI	●	◎		
2009	MORAKOT	●	◎		
2012	TEMBIN	●	◎		
2012	SAOLA	●	◎		
2013	TRAMI	●	◎		
2013	KONG-REY	●	◎		
2013	FITOW	●	◎		
2014	MATMO	●	◎		
2015	SOUDELOR	●	◎ validation		
2015	DUJUAN	●	◎ validation		

gages. The basic concept of supplement is to arrange the data of several gages in the catchment in sequence hours to a mathematical vector, and historical rainfall records were divided into m clusters. The winner-take-all neural network is used to build the relationship between samples and m clusters as well. When part of the input vector data is missing, one can still use the remaining information to determine the cluster. Figure 3 shows the flow chart of vector transform when losing data. Due to lack of data in some stations, the n-component vector will be transformed to an n'-component vector and the computation will be proceeded in remaining n' components. By using the procedure of Figure 3, one can transform $\overrightarrow{x}$ and $\overrightarrow{c}_j$ into $\overrightarrow{\xi}$ and $\overrightarrow{\varsigma}_j$, and pattern number can be judged by the following formula:

$$P\left(\overrightarrow{\xi}\right) = \sum_{j=1}^{m} jf\left(\left\|\overrightarrow{\xi} - \overrightarrow{\varsigma}_j\right\|\right). \tag{11}$$

Table 2: Results of the classification of rainfall sample in Tamsui River Basin.

Serial number		Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Number of samples
(1)	$p(t)$	0.15	0.20	0.53	0.37	0.45	0.15	0.26	0.50	0.56	0.36	0.11	0.23	0.52	908
	$p(t-1)$	0.17	0.24	0.53	0.36	0.41	0.15	0.27	0.53	0.51	0.29	0.15	0.25	0.51	
	$p(t+1)$	0.35	0.43	0.82	0.60	0.74	0.26	0.37	0.65	0.91	0.56	0.28	0.34	0.78	
(2)	$p(t)$	1.54	1.63	1.94	1.84	2.46	1.11	1.27	2.30	3.19	2.26	1.69	1.14	2.86	362
	$p(t-1)$	1.65	1.77	1.92	1.82	2.26	1.20	1.37	2.50	2.94	2.37	1.84	1.16	2.69	
	$p(t+1)$	1.92	2.30	3.15	2.48	2.72	1.17	1.32	2.41	3.58	2.42	1.88	1.20	3.18	
(3)	$p(t)$	12.03	10.09	4.76	3.33	3.03	1.44	1.25	2.72	3.92	3.28	5.75	0.51	2.28	73
	$p(t-1)$	12.13	8.45	4.87	2.88	2.83	1.04	0.67	2.85	4.49	2.61	3.93	0.33	1.72	
	$p(t+1)$	7.33	7.14	5.79	4.34	4.66	1.97	1.84	4.01	4.93	3.44	4.22	0.95	2.92	
(4)	$p(t)$	7.76	6.67	7.49	5.69	4.13	4.20	4.29	11.67	4.49	5.60	5.78	2.29	5.62	41
	$p(t-1)$	11.69	10.78	9.76	7.29	4.89	5.62	5.33	12.62	5.71	6.64	11.33	1.84	5.93	
	$p(t+1)$	7.12	7.78	8.88	6.80	5.17	5.15	4.63	12.20	5.93	6.59	6.24	2.76	5.95	
(5)	$p(t)$	0.85	2.39	9.93	4.35	2.70	0.60	0.78	1.37	2.80	2.23	0.61	0.40	1.68	138
	$p(t-1)$	0.83	2.33	10.14	4.58	2.86	0.65	0.80	1.75	2.60	2.07	0.72	0.38	1.62	
	$p(t+1)$	1.74	2.78	8.41	4.16	3.17	0.70	1.28	2.05	3.48	2.88	0.96	1.00	2.64	
(6)	$p(t)$	1.37	1.90	2.87	2.65	5.36	1.66	4.92	6.35	5.80	4.35	1.37	5.35	8.51	118
	$p(t-1)$	1.27	2.06	3.42	2.89	5.94	2.72	6.16	7.43	6.28	5.23	1.64	6.33	10.05	
	$p(t+1)$	1.91	2.91	3.27	3.16	4.31	2.14	3.92	5.16	4.66	4.20	1.82	4.24	6.48	
(7)	$p(t)$	1.38	3.81	5.81	6.81	17.43	9.71	18.81	10.95	13.38	20.86	3.52	16.10	23.38	18
	$p(t-1)$	1.33	4.62	7.14	7.90	25.24	8.33	21.90	13.43	18.90	21.24	3.67	16.52	28.90	
	$p(t+1)$	2.50	4.78	8.72	8.83	13.78	4.67	21.94	8.89	11.17	13.50	7.22	11.39	16.94	
(8)	$p(t)$	3.47	5.96	13.30	9.11	6.35	1.72	2.13	4.60	7.93	5.89	2.89	1.32	5.91	151
	$p(t-1)$	3.31	5.43	11.59	8.26	5.88	1.56	2.11	4.09	7.62	6.13	2.43	1.46	6.30	
	$p(t+1)$	3.96	5.89	13.11	8.24	7.30	2.35	3.05	6.03	8.64	6.60	2.93	2.44	7.58	
(9)	$p(t)$	3.41	7.15	13.25	9.84	11.98	3.90	7.00	13.67	13.16	7.54	3.10	5.49	14.36	58
	$p(t-1)$	3.82	6.39	12.08	9.33	9.75	2.62	4.18	11.00	11.95	5.48	2.79	3.48	11.25	
	$p(t+1)$	5.50	8.76	14.60	11.48	11.40	4.69	8.02	14.48	14.28	9.09	4.55	6.24	13.50	
(10)	$p(t)$	1.70	3.70	3.35	2.95	4.55	0.30	0.15	0.70	25.45	1.45	11.40	1.00	2.20	22
	$p(t-1)$	2.00	2.20	2.95	2.90	3.80	0.45	0.15	1.00	23.95	3.10	10.50	0.75	2.40	
	$p(t+1)$	6.00	6.09	4.05	5.73	6.55	1.05	0.82	2.86	20.91	2.32	13.23	2.18	3.95	
(11)	$p(t)$	8.82	17.82	25.79	24.33	16.08	10.44	9.05	18.23	14.77	15.82	7.62	5.67	13.10	41
	$p(t-1)$	6.51	11.33	21.67	14.74	11.67	5.62	6.23	14.82	11.67	10.00	3.67	5.49	11.23	
	$p(t+1)$	11.44	14.59	24.17	19.71	17.68	11.49	8.95	15.29	15.63	14.63	8.17	5.98	13.20	
(12)	$p(t)$	7.50	14.96	19.46	19.58	28.58	9.00	17.00	19.62	29.50	19.54	7.38	15.65	32.42	25
	$p(t-1)$	6.77	11.00	19.77	16.50	11.12	4.08	6.31	10.62	12.54	8.46	6.23	6.81	12.50	
	$p(t+1)$	12.08	15.44	21.48	16.96	20.40	8.52	11.96	17.60	19.48	18.88	8.64	10.08	20.40	
(13)	$p(t)$	10.69	11.00	26.24	14.07	9.17	5.55	3.34	9.03	8.83	7.10	4.90	3.07	8.07	28
	$p(t-1)$	11.62	21.21	30.86	25.62	16.69	14.52	13.21	24.00	12.24	16.66	8.55	8.76	14.59	
	$p(t+1)$	6.54	11.89	22.14	15.43	15.29	7.43	7.46	14.43	13.39	12.25	5.21	6.32	14.21	

Table 2: Continued.

Serial number		Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Number of samples
(14)	$p(t)$	19.08	15.58	26.04	12.13	8.04	2.25	3.79	8.29	17.21	9.79	12.79	3.25	6.79	23
	$p(t-1)$	14.58	18.79	30.21	14.17	9.50	2.96	2.50	6.38	17.04	8.46	10.46	2.83	5.83	
	$p(t+1)$	11.52	12.83	22.57	12.61	9.13	3.26	5.57	11.87	14.70	13.35	6.78	4.35	10.78	
(15)	$p(t)$	26.53	23.68	7.84	14.68	10.68	13.37	13.58	23.95	7.74	12.42	25.74	9.42	11.05	20
	$p(t-1)$	31.84	29.95	14.47	23.84	13.58	8.84	10.63	20.32	14.05	11.58	31.63	7.84	12.21	
	$p(t+1)$	19.85	20.30	8.50	12.90	12.30	13.60	12.00	20.45	8.65	13.70	18.05	10.15	11.90	
(16)	$p(t)$	7.27	17.32	8.27	9.23	6.14	20.14	7.09	20.50	3.55	9.00	16.68	4.09	6.55	22
	$p(t-1)$	8.05	11.82	7.23	8.18	7.23	23.64	6.59	18.27	3.68	12.59	17.27	3.73	8.00	
	$p(t+1)$	11.14	14.91	8.05	9.50	4.41	14.77	6.95	15.95	6.14	8.27	16.41	3.73	6.50	
(17)	$p(t)$	14.43	22.36	25.71	29.43	35.64	37.00	50.07	36.43	34.93	46.50	17.21	29.07	42.36	11
	$p(t-1)$	13.71	23.79	28.36	32.21	46.86	40.93	48.14	33.14	45.14	53.50	17.79	27.29	47.93	
	$p(t+1)$	17.36	23.09	34.64	27.73	26.91	23.09	39.18	34.91	33.00	29.91	17.91	20.64	36.64	
(18)	$p(t)$	15.93	25.50	41.14	31.93	33.93	24.71	16.64	27.86	19.86	30.14	5.50	18.21	22.50	13
	$p(t-1)$	9.64	17.07	33.14	24.14	22.86	14.79	11.79	17.57	8.43	20.36	1.71	14.21	16.50	
	$p(t+1)$	13.92	17.46	36.92	21.38	23.77	9.77	14.92	25.46	16.31	18.38	12.54	15.62	18.15	
(19)	$p(t)$	14.11	22.44	33.30	33.30	41.89	16.70	22.96	35.22	44.59	28.37	14.81	16.26	37.52	27
	$p(t-1)$	12.81	21.48	32.11	30.74	37.30	14.11	21.67	33.37	37.67	24.70	13.81	16.89	32.56	
	$p(t+1)$	18.56	24.26	36.26	30.00	28.15	16.41	29.11	37.22	36.74	28.30	16.78	17.11	35.37	
(20)	$p(t)$	23.17	30.17	30.61	25.04	9.30	5.74	14.61	12.26	24.61	23.43	22.52	11.61	19.09	19
	$p(t-1)$	19.09	24.78	25.43	19.74	14.09	7.74	21.61	18.13	28.26	29.61	15.30	18.78	26.61	
	$p(t+1)$	19.63	22.37	22.95	20.47	9.16	5.37	17.16	10.89	20.84	20.74	18.00	12.95	20.47	
(21)	$p(t)$	7.29	10.71	11.13	13.04	17.67	9.25	9.58	13.33	27.79	11.33	6.92	9.58	22.25	26
	$p(t-1)$	7.54	16.17	14.29	18.63	29.46	11.00	10.54	12.46	42.92	16.58	7.75	10.17	30.79	
	$p(t+1)$	7.08	12.19	10.50	13.31	15.08	7.35	8.27	11.62	26.73	9.50	6.58	9.00	20.46	
(22)	$p(t)$	8.71	13.68	28.84	16.10	24.29	7.52	19.58	33.68	25.65	18.45	9.42	12.90	18.16	31
	$p(t-1)$	9.77	15.71	31.61	19.19	26.32	8.39	20.06	34.45	29.68	20.48	11.19	11.77	19.26	
	$p(t+1)$	8.68	15.94	28.00	20.55	24.65	10.68	19.90	31.68	25.90	18.68	9.23	13.74	19.77	

Rainfall unit: mm.

Table 3: Coefficients of correlation and efficiency in the dataset of calibration.

	Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Average
CC	0.66	0.82	0.84	0.84	0.82	0.73	0.81	0.84	0.81	0.82	0.75	0.75	0.84	0.79
CE	0.5	0.58	0.65	0.62	0.56	0.49	0.56	0.64	0.54	0.57	0.5	0.5	0.6	0.55

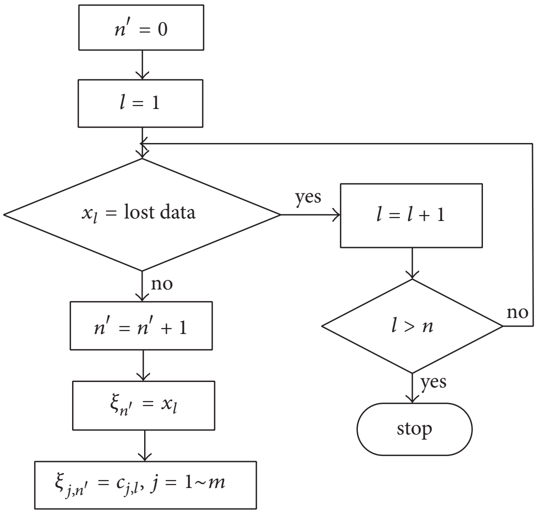


Figure 3: Flow chart of vector transform when data missing occurs.

Then the pattern of the lost data can be supplement by the following formula:

$$x_l \cong c_{P,l}, \qquad (12)$$

where subscript P is pattern number and l denotes the lth component of $\overrightarrow{x}$ and $\vec{c}_P$. Note that this $\overrightarrow{x}$ is a new input vector when the rainfall is applied.

Another typhoon event, Aere in 2004 with large missing data in Sanxia, was chosen for demonstrating the performance of incomplete input data.

3.4. Validation and Performance Measures. To evaluate the model performance, two indices which are commonly used are employed here.

Coefficient of correlation (CC) is as follows:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N}(p_i - \overline{p})(o_i - \overline{o})}{\sqrt{\sum_{i=1}^{N}(p_i - \overline{p})^2 \times \sum_{i=1}^{N}(o_i - \overline{o})^2}}. \qquad (13)$$

Coefficient of efficiency (CE) is as follows:

$$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{\sum_{i=1}^{N}(o_i - \overline{o})^2}. \qquad (14)$$

In (13) and (14), p is the forecast value and o is the observation value. i is the serial number of sample, and N is the amount of samples included in a certain event. The over bar indicates the average quantities:

$$\overline{o} = \frac{1}{N}\sum_{i=1}^{N} o_i \qquad (15a)$$

$$\overline{p} = \frac{1}{N}\sum_{i=1}^{N} p_i. \qquad (15b)$$

4. Result and Discussion

Table 3 illustrates the results of coefficient of correlation and coefficient of efficiency in the dataset of calibration. It is apparent from the information supplied that they showed the consistency among the 13 rain gages. The correlation coefficient exceeds 0.66 (the lowest at Shimen), while the highest is up to 0.84. The highest coefficient of efficiency is 0.65 (Fushan) and the average is 0.55.

Typhoon Soudelor was the most intense tropical cyclone to develop in the Northern Hemisphere in 2015 (category 5 super typhoon scaled by SSHWS). When it passed through Taiwan, torrential rains and destructive winds caused widespread damage and disruptions, especially in north area. According to Central Emergency Operation Center, at least eight people were killed and four were missing in Taiwan, in addition to 437 injured. Agricultural losses across the island were estimated at NT$2.2 billion (US$66.7 million) by August 11. A record-breaking 4.29 million households lost power on the island. Figure 4 shows the observed hourly rainfall data and the simulation results during Typhoon Soudelor (2015). As we can see, the observed hourly rainfall data are up to 87 mm in Zhuzihu rain gage. In addition, the model output showed a good agreement between simulated and observed data. The coefficient of correlation and efficiency values are shown in Table 4. The coefficient of correlation exceeds 0.68 (the lowest at Shimen), while the highest is up to 0.89 (Zhuzihu). The highest coefficient of efficiency is 0.74 (Shiding and Ruifang) and the average is 0.62.

Typhoon Dujuan was the second most intense tropical cyclone of the Northwest Pacific Ocean in 2015 (category 4 typhoon scaled by SSHWS). Three people were killed and 376 were injured in Taiwan. Figure 5 shows the observed hourly rainfall data and the simulation results during Dujuan Typhoon (2015). It also indicated that the observed data and the simulated data were quite close. The coefficient of correlation and efficiency values are shown in Table 5. The coefficient of correlation exceeds 0.69 (the lowest at Shimen), while the highest is up to 0.88 (Jhongjheng Bridge). The highest coefficient of efficiency is 0.77 (Jhongjheng Bridge) and the average is 0.65.

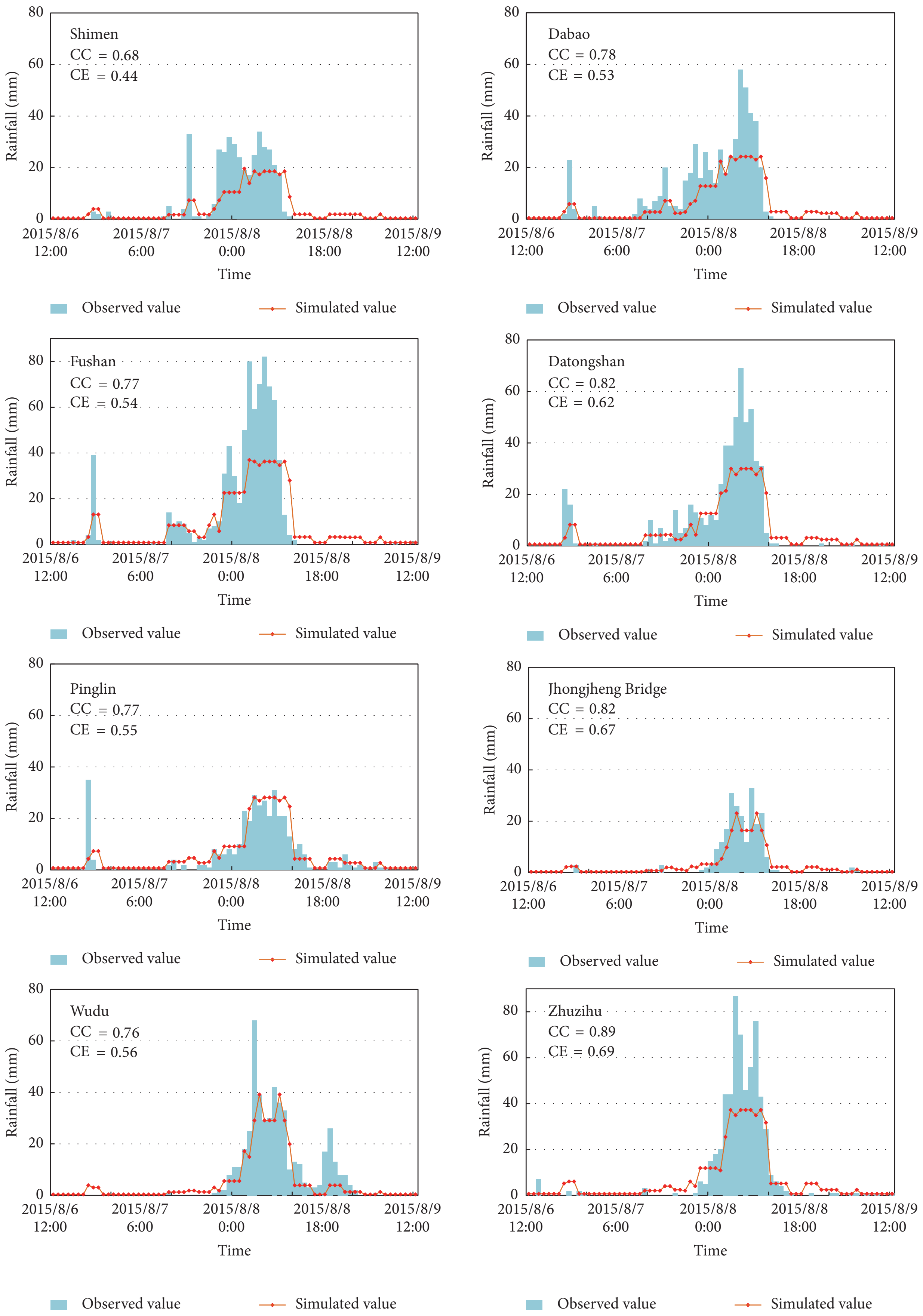


FIGURE 4: Continued.

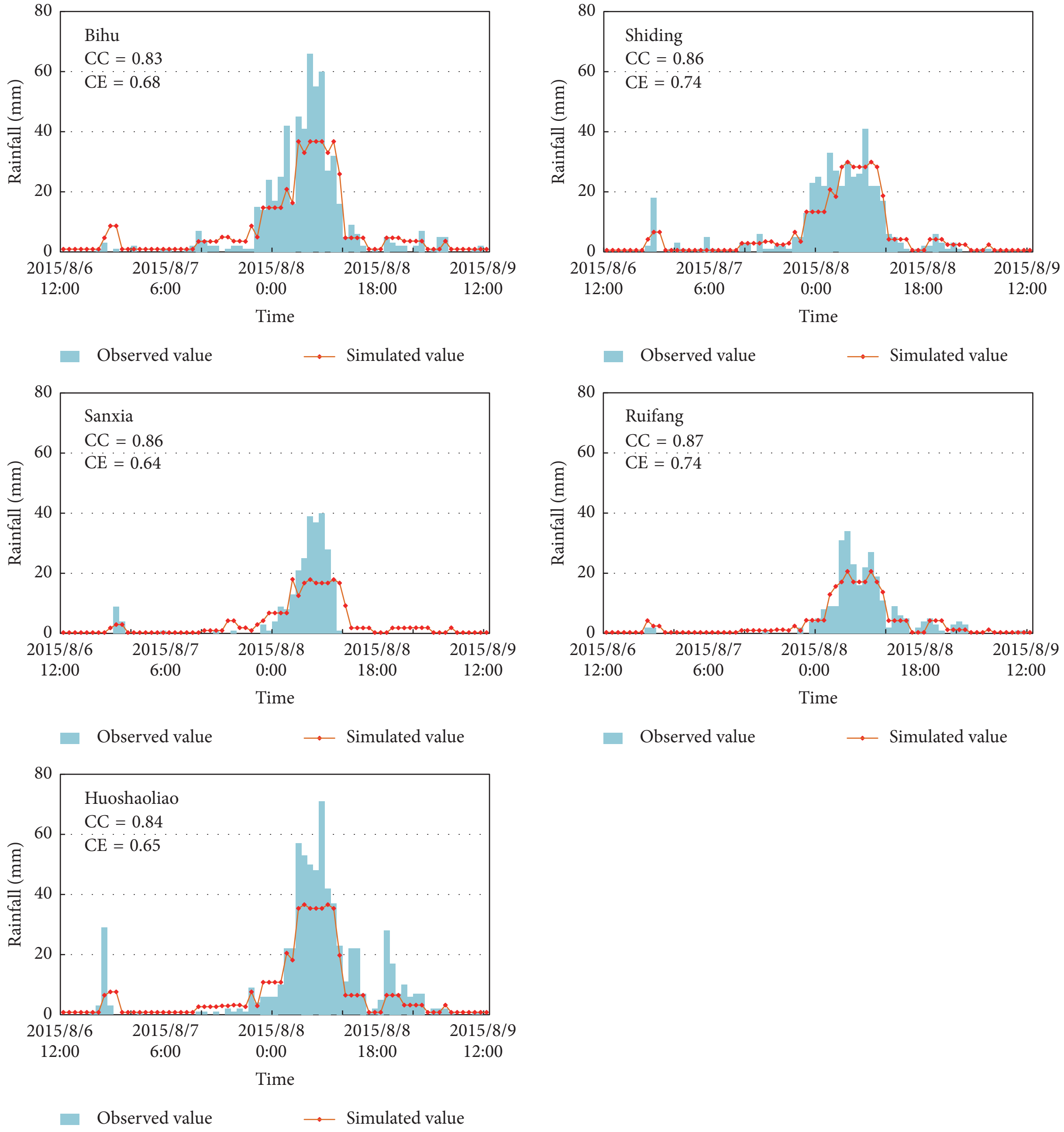


Figure 4: Results of hourly rainfall forecast during Typhoon Soudelor (2015).

Table 4: Coefficients of correlation and efficiency during Typhoon Soudelor (2015).

	Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Average
CC	0.68	0.78	0.77	0.82	0.77	0.82	0.76	0.89	0.83	0.86	0.84	0.87	0.84	0.81
CE	0.44	0.53	0.54	0.62	0.55	0.67	0.56	0.69	0.68	0.74	0.64	0.74	0.65	0.62

Table 5: Coefficients of correlation and efficiency during Typhoon Dujuan (2015).

	Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Average
CC	0.69	0.82	0.77	0.85	0.84	0.88	0.83	0.86	0.79	0.81	0.84	0.86	0.85	0.82
CE	0.45	0.62	0.57	0.73	0.71	0.77	0.68	0.71	0.62	0.64	0.7	0.66	0.67	0.65

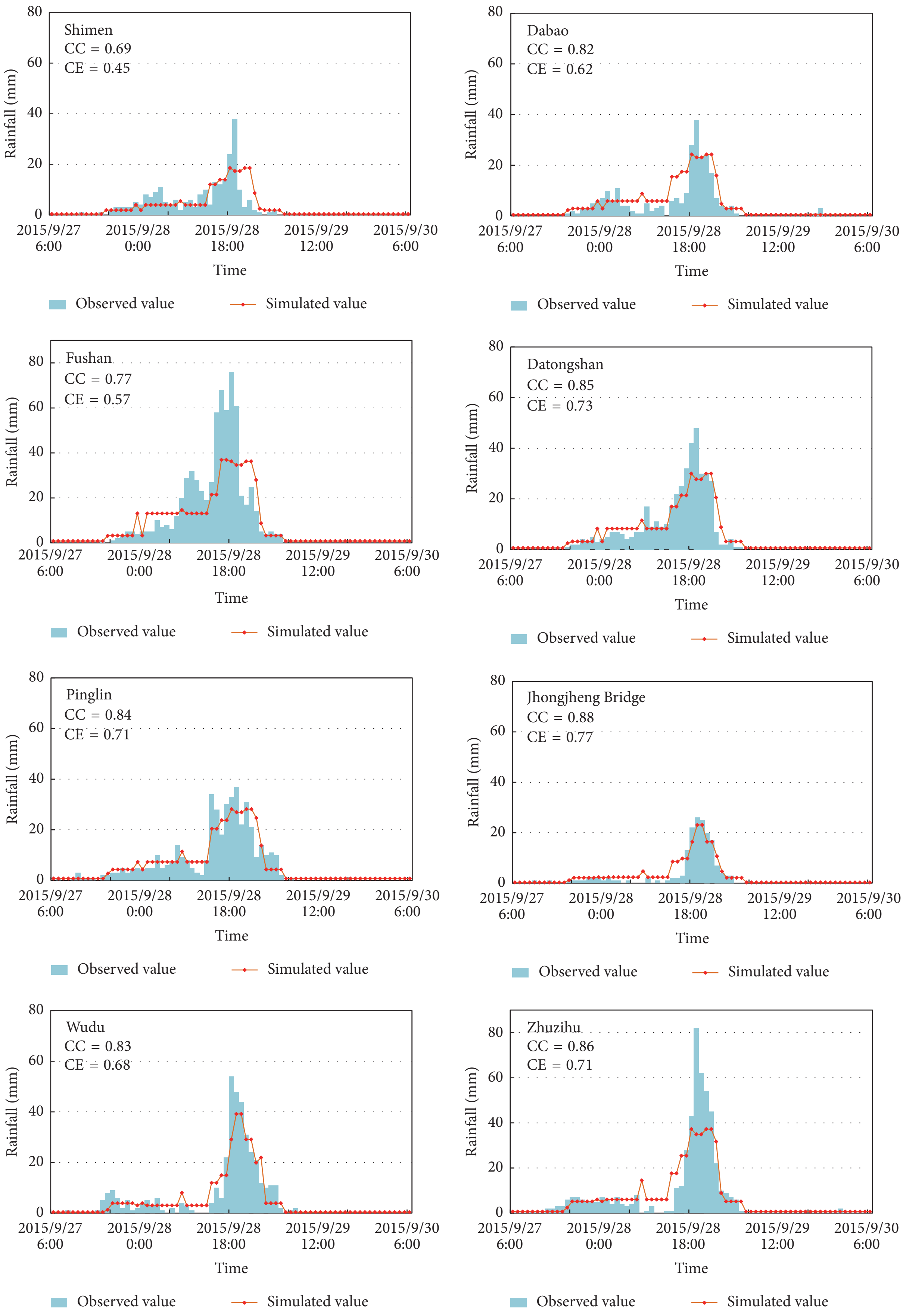


FIGURE 5: Continued.

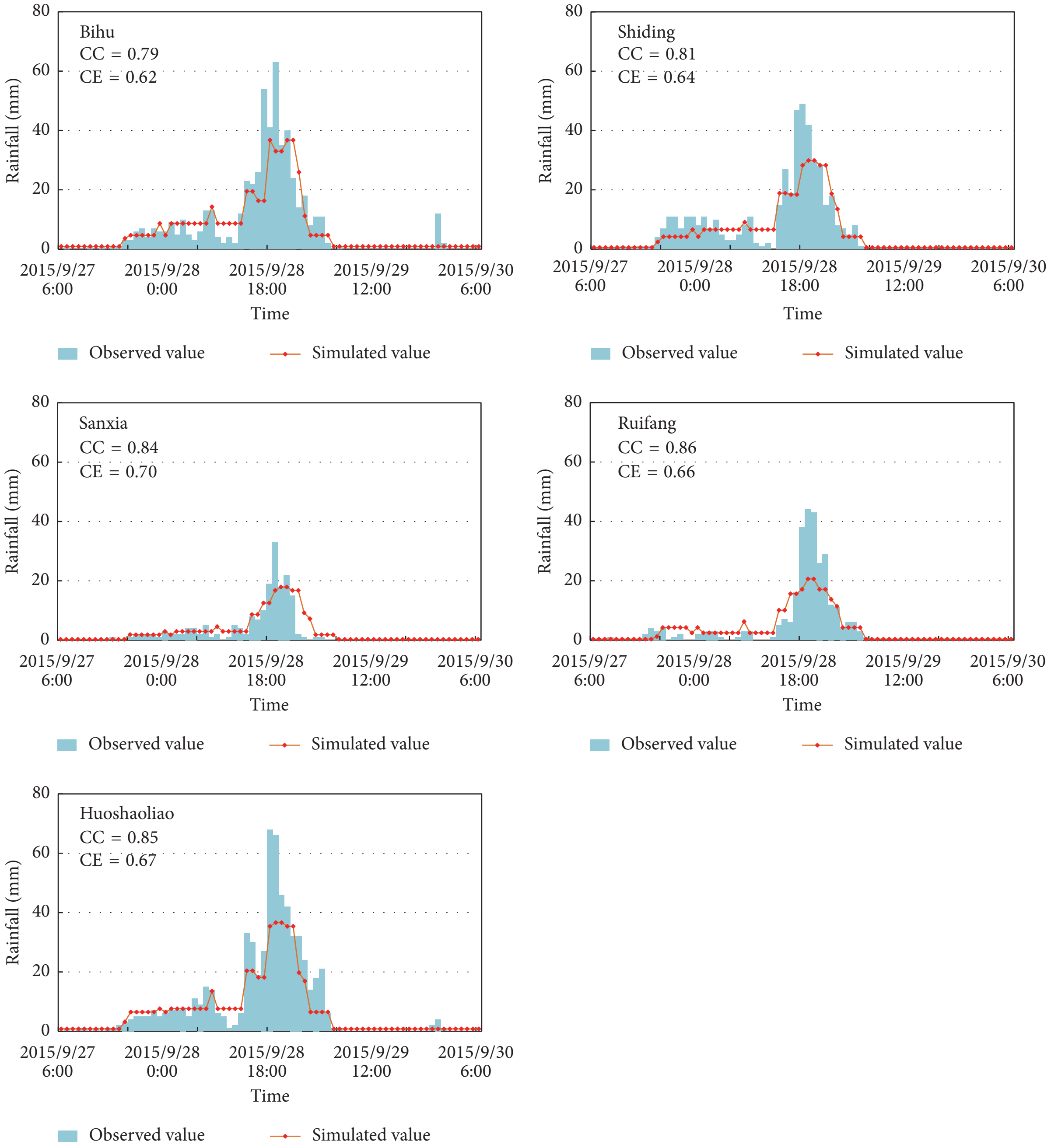


Figure 5: Results of hourly rainfall forecast during Typhoon Dujuan (2015).

Figure 6 shows the observed data and results of simulation during Typhoon Aere (2004). Because of missing data, there is no observed rainfall data at Sanxia station during Typhoon Aere. By the procedure of supplement, data of remaining stations could still be simulated and compared to the observed data. The coefficient of correlation and efficiency values are shown in Table 6. The simulation of the coefficient of correlation exceeds 0.61 (the lowest at Ruifang), while the highest is up to 0.89 (Fushan). The highest coefficient of efficiency is 0.76 (Fushan), while the average is 0.55.

Compared to previous studies, in Luk et al. [11], they used BPNN and successfully built a model for forecasting the rainfall pattern in the next coming 15 minutes. Normalized mean squared error (NMSE) was chosen as the performance indicator and was about 0.63 to 0.65. In Luk et al. [12] they also used the same concept to build another rainfall forecasting model by applying other kinds of neural networks such as multilayer feedforward neural network, partial recurrent network, and time delayed neural network. Normalized mean squared error was about 0.63 to 0.67 forecasting the rainfall pattern in the next coming 15 minutes. In Lin et al. [14], they used SVM-based models with and without typhoon characteristics to forecast the rainfall. The coefficients of efficiency are 0.44 and 0.43, respectively. In this study the average

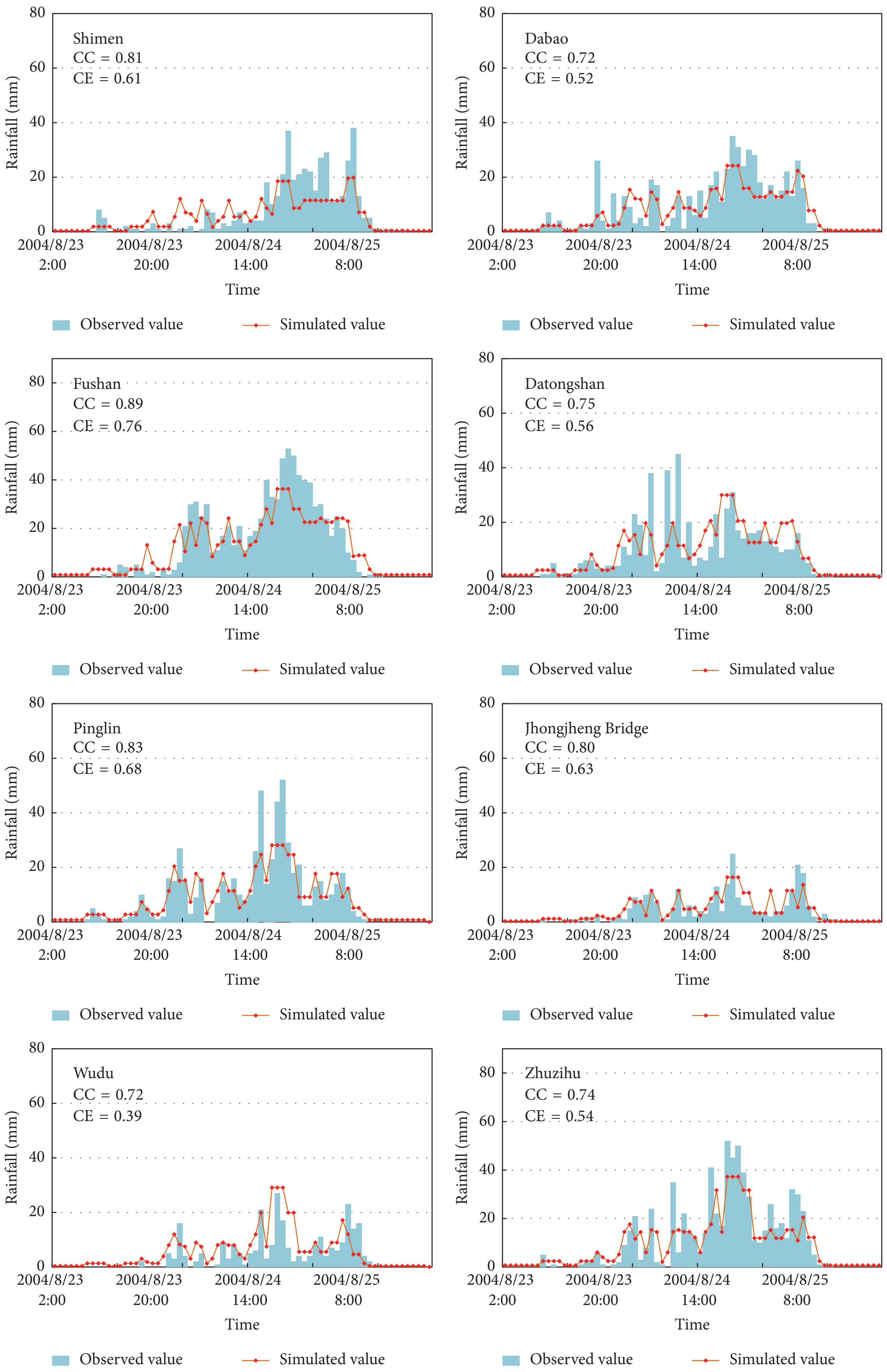


FIGURE 6: Continued.

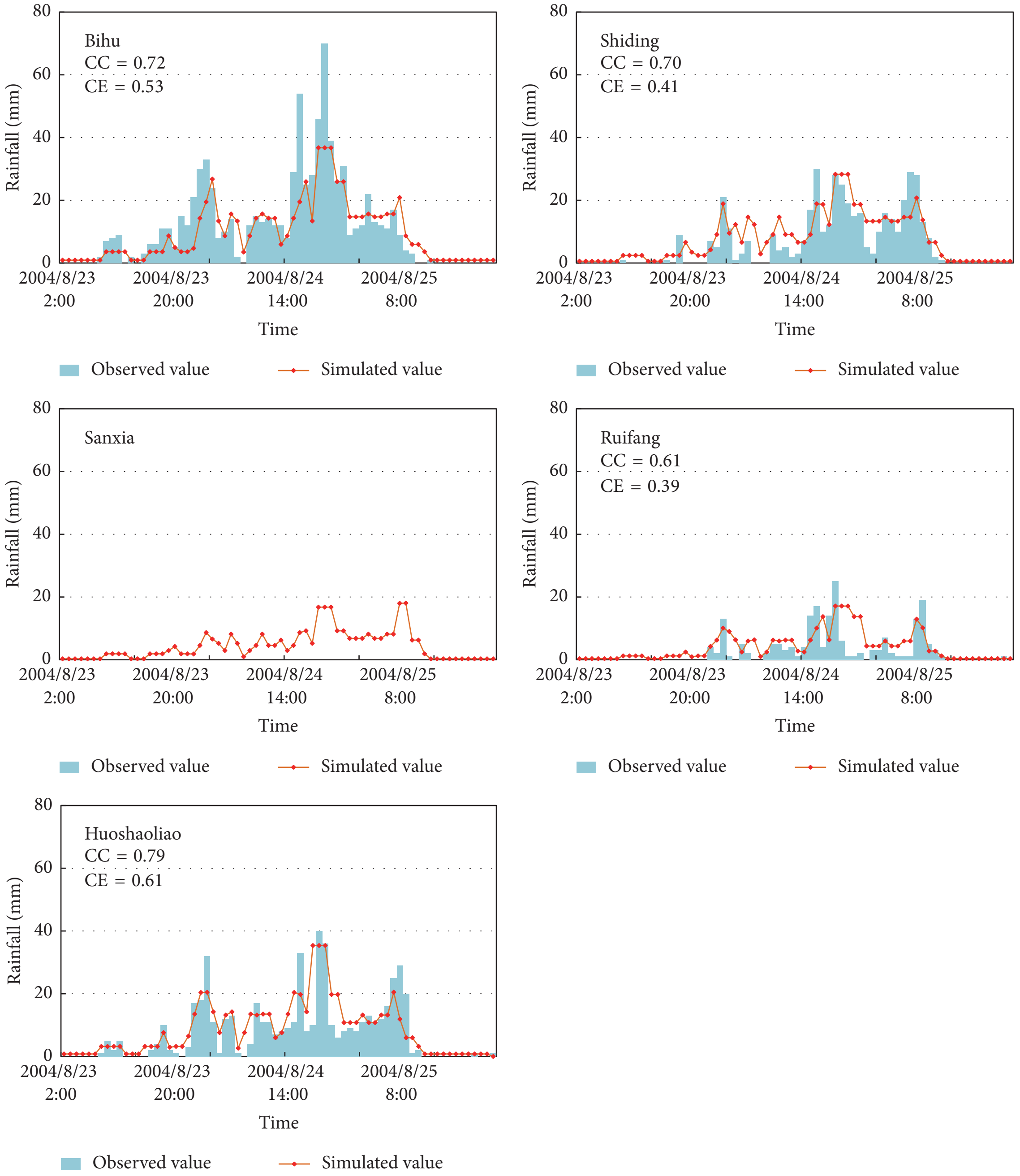


Figure 6: Results of hourly rainfall forecast during Typhoon Aere (2004) (data at Sanxia gage was missing).

coefficient of efficiency is 0.64 and reasonably good results have been observed. We thus think that the improvement is effective.

Note that the dimensions consist of two factors: the number of previous time steps and rain gages. They are also related to cluster sizes while sizes and number of clusters depend on how many samples we have. As we have limited number of samples, the cluster sizes would be too small if we use too many dimensions. That would cause inaccuracy by overfitting. Because the rainfall forecasting model is developed for the use of flood mitigation around the capital city of Taiwan, we should use as many rain gages in the entire watershed as possible. That is also the reason we need to make the system keep on working even if data missing occurs. When trying to develop a similar rainfall forecast system in other area with a larger sample size, one could have more options to test the effect of number of dimensions and cluster size. In this paper rainfall records of last and present hour are used as the input data to forecast gaged rainfall in the next hour. Totally 13 rain gages in space were used since data in remaining 3 rain gages are not enough. So we have totally 26 components in the input data.

Table 6: Coefficients of correlation and efficiency during Typhoon Aere (2004) (data at Sanxia gage was missing).

	Shimen	Dabao	Fushan	Datongshan	Pinglin	Jhongjheng Bridge	Wudu	Zhuzihu	Bihu	Shiding	Sanxia	Ruifang	Huoshaoliao	Average
CC	0.81	0.72	0.89	0.75	0.83	0.8	0.72	0.74	0.72	0.70	—	0.61	0.79	0.75
CE	0.61	0.52	0.76	0.56	0.68	0.63	0.39	0.54	0.53	0.41	—	0.39	0.61	0.55

5. Conclusions

By integrating technique of cluster analysis and pattern recognition, an unsupervised method is adopted in this paper to provide reasonably accurate and effective typhoon hourly rainfall forecast. Not only can the missing data be supplied but also rainfall data of space and time in the previous time steps are needed to forecast the hourly intensity of rainfall for next time steps. Present proposed forecast model is tested using historical rainfall data in Tamsui River Basin. Among 32 typhoon events from which complete rainfall records can be obtained, 30 of them are used to calibrate. The data are clustered into 22 patterns for the network construction. After the framework is built, the rest two of the typhoon events, Soudelor (2015) and Dujuan (2015), are used to validate the model. Additionally, another typhoon event, Aere (2004), during which the rainfall data was lost at one of the 13 gages, is used to illustrate how this model works when the input of the model is incomplete. The performance is testified by coefficient of correlation and coefficient of efficiency. Reasonably good results have been observed in these cases. It shows that present proposed forecast model is well suited for predicting the hourly rainfall during typhoon in Northern Taiwan.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors are grateful for the financial support from the Ministry of Science and Technology, Taiwan (Grant no. 105-2221-E-415-008). The authors would also like to thank the 10th River Management Office, Water Resources Agency, Taiwan, for providing the rainfall data.

References

[1] E. Toth, A. Brath, and A. Montanari, "Comparison of short-term rainfall prediction models for real-time flood forecasting," *Journal of Hydrology*, vol. 239, no. 1-4, pp. 132–147, 2000.

[2] C. Lai, T.-K. Tsay, C.-H. Chien, and I.-L. Wu, "Real-time flood forecasting," *American Scientist*, vol. 97, no. 2, pp. 119–125, 2009.

[3] I. L. Wu, T. K. Tsay, and C. Lai, "Numerical modelling of basin-wide flood flow using variable flow-resistance coefficients," *Journal of Flood Engineering*, vol. 2, no. 1-2, pp. 99–120, 2011.

[4] I. L. Wu, P. H. Chen, and T. K. Tsay, "Using a basin-wide river numerical model to evaluate flood control improvement measures—a case study on zhonggang main drainage at Xinzhuang District, New Taipei City," *Journal of the Chinese Institute of Civil and Hydraulic Engineering*, vol. 28, no. 1, pp. 45–55, 2016.

[5] B. W. Golding, "Nimrod: a system for generating automated very short range forecasts," *Meteorological Applications*, vol. 5, no. 1, pp. 1–16, 1998.

[6] J. N. K. Liu and R. S. T. Lee, "Rainfall forecasting from multiple point sources using neural networks," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC '99)*, October 1999.

[7] J. Joss, A. Waldvogel, and C. G. Collier, "Precipitation measurement and hydrology," in *Radar in Meteorology*, pp. 577–606, Springer, 1990.

[8] D. Rosenfeld and C. W. Ulbrich, "Cloud microphysical properties, processes, and rainfall estimation opportunities," *Meteorological Monographs*, vol. 30, no. 52, pp. 237–258, 2003.

[9] R. Krzysztofowicz, "Recent advances associated with flood forecast and warning systems," *Reviews of Geophysics*, vol. 33, supplement 2, pp. 1139–1147, 1995.

[10] D.-J. Seo, J. P. Breidenbach, and E. R. Johnson, "Real-time estimation of mean field bias in radar rainfall data," *Journal of Hydrology*, vol. 223, no. 3-4, pp. 131–147, 1999.

[11] K. C. Luk, J. E. Ball, and A. Sharma, "A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting," *Journal of Hydrology*, vol. 227, no. 1-4, pp. 56–65, 2000.

[12] K. C. Luk, J. E. Ball, and A. Sharma, "An application of artificial neural networks for rainfall forecasting," *Mathematical and Computer Modelling*, vol. 33, no. 6-7, pp. 683–693, 2001.

[13] F.-J. Chang, Y.-M. Chiang, and L.-C. Chang, "Multi-step-ahead neural networks for flood forecasting," *Hydrological Sciences Journal*, vol. 52, no. 1, pp. 114–130, 2007.

[14] G. Lin, G. Chen, M. Wu, and Y. Chou, "Effective forecasting of hourly typhoon rainfall using support vector machines," *Water Resources Research*, vol. 45, no. 8, 2009.

[15] R. Tryon and D. Bailey, *Cluster Analysis*, McGraw-Hill Book, New York, NY, USA, 1970.

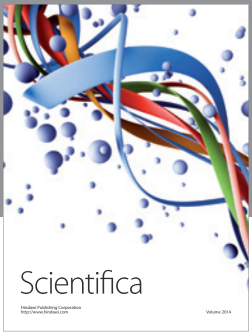
[16] R. J. Jang, "Data Clustering and Pattern Recognition," http://mirlab.org/jang/.

[17] G. Punj and D. W. Stewart, "Cluster analysis in marketing research: review and suggestions for application," *Journal of Marketing Research*, vol. 20, no. 2, pp. 134–148, 1983.

[18] SPSS, "The SPSS TwoStep cluster component," Tech. Rep., 2001, http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf.

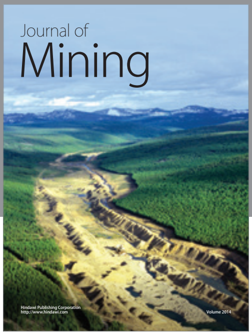
[19] IBM Knowledge Center, https://www.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/alg_twostep.htm.

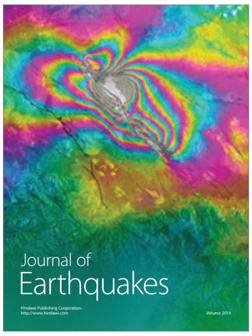
Journal of Geochemistry


Scientifica

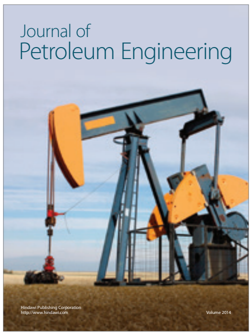
International Journal of Ecology

The Scientific World Journal

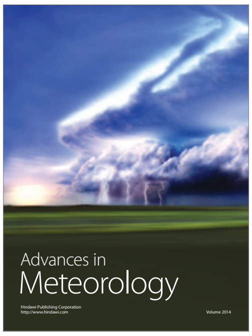

Journal of Mining

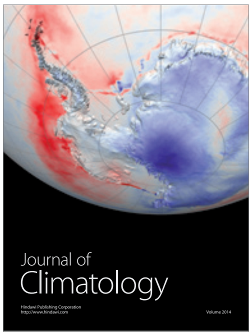

Journal of Earthquakes

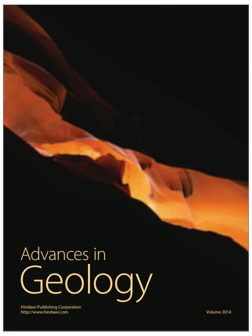
Paleontology Journal

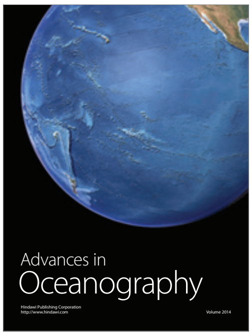

Journal of Petroleum Engineering

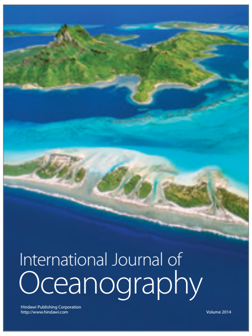
International Journal of Geophysics


Advances in Meteorology


Journal of Climatology

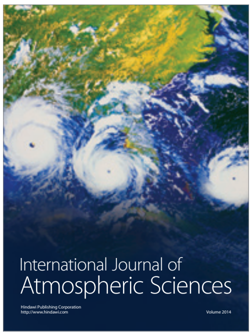

Advances in Geology

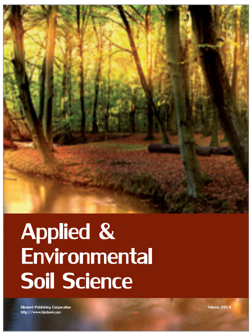

Advances in Oceanography


International Journal of Oceanography

International Journal of Mineralogy

Journal of Geological Research


International Journal of Atmospheric Sciences


Applied & Environmental Soil Science

Journal of Computational Environmental Sciences