A Comparison of Hourly Typhoon Rainfall Forecasting Models Based on Support Vector Machines and Random Forests with Different Predictor Sets

Kun-Hsiang Lin, Hung-Wei Tseng, Chen-Min Kuo, Tao-Chang Yang, Pao-Shan Yu*

Department of Hydraulic and Ocean Engineering, National Cheng Kung University, Taiwan (R.O.C.)
No. 1, University Road, Tainan City 70101, Taiwan (R.O.C.)
Tel: +886-6-2757575 ext. 63248
Email: yups@mail.ncku.edu.tw

This study aims to develop rainfall forecasting models based on two machine learning methods, support vector machines (SVMs) and random forests (RFs), and investigate the performances of the models with different predictor sets for searching the optimal predictor set in forecasting. Four predictor sets were used: (1) antecedent rainfalls, (2) antecedent rainfalls and typhoon characteristics, (3) antecedent rainfalls and meteorological factors, and (4) antecedent rainfalls, typhoon characteristics and meteorological factors to construct for 1- to 6-hour ahead rainfall forecasting. An application to three rainfall stations in Yilan River basin, northeastern Taiwan, was conducted. Firstly, the performance of the SVMs-based forecasting model with predictor set #1 was analyzed. The results show that the accuracy of the models for 2- to 6-hour ahead forecasting decrease rapidly as compared to the accuracy of the model for 1-hour ahead forecasting which is acceptable. For improving the model performance, each predictor set was further examined in the SVMs-based forecasting model. The results reveal that the SVMs-based model using predictor set #4 as input variables performs better than the other sets and a significant improvement of model performance is found especially for the long lead time forecasting. Lastly, the performance of the SVMs-based model using predictor set #4 as input variables was compared with the performance of the RFs-based model using predictor set #4 as input variables. It is found that the RFs-based model is superior to the SVMs-based model in hourly typhoon rainfall forecasting.

Keywords: hourly typhoon rainfall forecasting, predictor selection, support vector machines, random forests

Introduction

Typhoons with heavy rainfall and strong wind often cause severe floods and disasters in Taiwan. Therefore, having a warning system with an accurate rainfall forecasting model is essential for flood mitigation during typhoon landfall. Moreover, typhoon rainfall is one of the most difficult elements of the hydrologic cycle to forecast due to its high spatiotemporal variability. 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 (Lin *et al*., 2009). With the rapid development of computer performance, a new type of methods called machine learning (ML) such as artificial neural networks (ANNs) or support vector machines (SVMs) has been widely applied on hydrological process (Lin and Chen, 2008; Lin *et al*., 2009). In the past, only antecedent rainfalls are considered for hourly rainfall forecasting (Hong and Pai, 2007), but the forecasting result shows that the accuracy of the models for 2- to 6-hour ahead forecasting decrease rapidly as compared to the accuracy of the model for 1-hour ahead forecasting which is acceptable. For obtaining more accurate forecasts of typhoon rainfall, typhoon characteristics and meteorological factors were introduced in this study. However, due to many variables are used as input to construct rainfall forecasting models, searching the optimal combination of input variables by manual methods (e.g. trial-and-error) is a tedious and time-consuming task. Hence, this study adopts non-dominated sorting genetic algorithm (NSGA-II) which has been widely applied in various disciplines for multi-objective optimization (Deb *et al*., 2002) to construct a predictor selection method. Moreover, another ML method called random forests (RFs) seldom applied in

hydrological process is used to construct hourly rainfall forecasting models in our work.

To sum up, there are two purposes in this study. One is to develop predictor selection method for constructing hourly rainfall forecasting. The second one is to compare the performance of two algorithms (SVMs and RFs) for hourly typhoon rainfall forecasting.

Study Area and Dataset

Yilan River, which encloses an area of 149.06 km^2, is located in Yilan County in the northeast of Taiwan (Fig. 1). The length of mainstream is about 25 km, the slope of upstream basin is around 0.24, and the slope of mainstream is around 0.038. Every year between April and October, typhoons created in the southwest Pacific Ocean tend to proceed in a westerly or northwesterly direction. Yilan County which faces the Pacific Ocean to the east is often directly hit by typhoons, which makes severe loss of life and property damage. An effective and efficient warning system is essential for flood mitigation in this area. Extending the lead-time of forecast is extremely important for catchments where the response time is short. Constructing rainfall forecasting model is helpful to extending the lead time of flood forecasting in the study area.

Fourteen typhoon events with hourly rainfall data, typhoon characteristics, and meteorological factors simultaneously available were used herein. Table I lists the date of occurrence and duration of these fourteen typhoon events. Hourly rainfall data (mm) from the three rainfall stations (Dajiaoxi, Zailian, and Shuanglianbi stations) operated by the Water Resources Agency of Taiwan were collected. The maximum 24-hour rainfalls for each typhoon event at the three rainfall stations are also listed in Table I. Typhoon characteristics at hourly time scale were collected from the Central Weather Bureau of Taiwan, which include eight characteristics, i.e. the position (latitude, longitude) of the typhoon center (°N, °E), the distance between the center and the catchment (km), the maximum wind speed near the center (m/s), the maximum instantaneous wind speed (i.e. maximum gust speed) near the center (m/s), the atmospheric pressure of the center (hPa), the radius of winds over 15 m/s (km), and the speed of the typhoon movement (m/s). Meteorological factors at hourly time scale were collected from the Yilang meteorological station operated by the Central Weather Bureau of Taiwan, which include eight factors, i.e. air pressure (hPa), air temperature (°C), dew point temperature (°C), relative humidity (%), wind velocity (m/s), wind direction (degree), rainfall depth (mm), and the rainfall duration (minute). The locations of the three rainfall stations and the meteorological station are shown in Fig. 1.

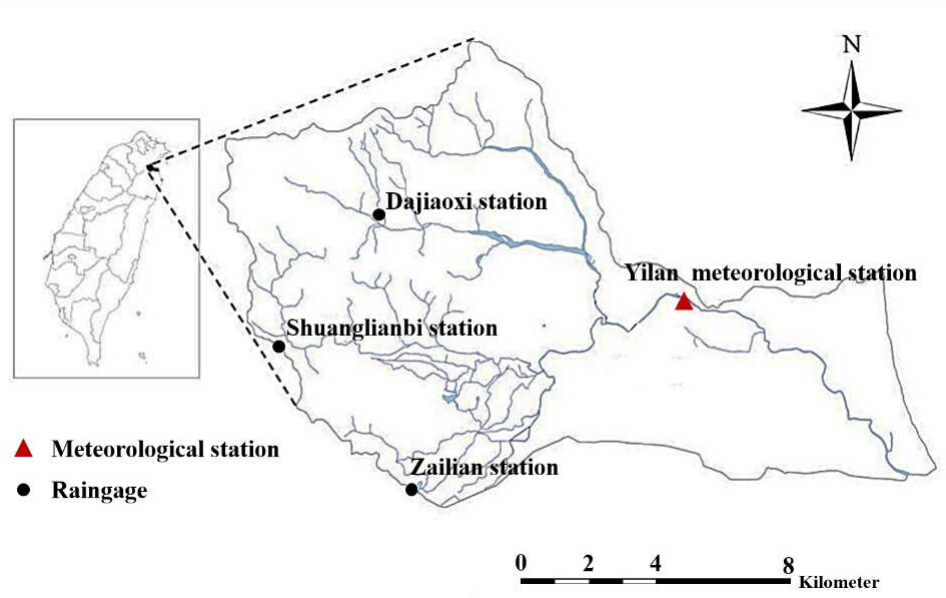

Fig. 1. Locations of rainfall and meteorological stations.

Table I. Description of typhoon events used in the study

No.	Name	Date (yyyy/mm/dd)	Duration (hour)	Maximum 24-hour rainfall (mm)		
				Daijiaoxi	Shunglianbi	Zailian
1	Herb	1996/07/30	65	514.5	369.0	249.5
2	Xangsane	2000/10/30	53	372.0	336.5	330.0
3	Nari	2001/09/15	101	528.5	475.0	469.5
4	Nakri	2002/07/09	47	267.5	205.5	133.0
5	Aere	2004/08/23	69	454.0	309.0	267.0
6	Nock-ten	2004/10/23	72	157.5	115.0	141.0
7	Haitang	2005/07/16	65	252.0	194.0	150.5
8	Talim	2005/08/30	44	175.5	122.5	91.0
9	Krosa	2007/10/04	49	421.0	285.5	134.5
10	Sinlaku	2008/09/11	69	468.5	314.5	342.5
11	Parma	2009/10/03	52	252.0	333.5	350.5
12	Megi	2010/10/21	72	269.0	240.5	320.5
13	Saola	2012/07/30	76	344.0	336.5	250.0
Test	Soudelour	2015/08/07	30	186.0	131.5	710.0

Methodology

1. Support Vector Machines (SVMs)

An SVMs that reduces the problem of over-fitting by adopting the theory of structural risk minimization has recently gained popularity in many disciplines. The SVMs is mainly utilized for classification and regression problems. Detailed principles and algorithms can be found in Vapnik (1995; 1998). In the study, the SVR was used for calculating the amounts of local precipitation. The description on the SVR is as follows (Yu et al., 2006; Chen and Yu, 2007).

Let data $[(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_l, y_l)]$ be a given training set, where x_i is an input vector, y_i is its corresponding output and l is the number of data pairs. SVR finds a regression function $f(x) = w^T \cdot \Phi(x) + b$ that best describes the observed output y with an error tolerance ε, where w and b are parameters, and $\Phi(x)$ is a nonlinear function. The penalized losses L_ε, when data are outside of the tube of error tolerance, are defined by the Vapnik's ε-insensitive loss function.

$$L_\varepsilon(y_i) = \begin{cases} 0, & \left| y_i - [w^T \cdot \Phi(x_i) + b] \right| < \varepsilon \\ \left| y_i - [w^T \cdot \Phi(x_i) + b] \right| - \varepsilon, & \left| y_i - [w^T \cdot \Phi(x_i) + b] \right| \geq \varepsilon \end{cases} \quad (1)$$

The SVR problem is then formulated as an optimization problem then can be solved by a dual set of Lagrange multipliers. Consequently, the approximate function can be expressed as:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i)^T \cdot \Phi(x) + b \quad (2)$$

The data with non-zero Lagrange multipliers ($\alpha_i - \alpha_i^*$) are actively in the regression function, and are the support vectors. With the use of a kernel function (herein, the radial basis function with a parameter γ), the regression function can be rewritten as:

$$f(x) = \sum_{k=1}^{m} (\alpha_k - \alpha_k^*) K(x_k, x) + b \quad (3)$$

where x_k is the support vector and m is the number of support vectors. The SVR has three parameters, C, ε and γ, that need to be calibrated. Given these three parameters, the Lagrange multipliers and parameter b in Equation (3) can be determined by the SVMs algorithm.

2. Random Forests (RFs)

RFs is an ensemble of decision trees where each tree is grown by using randomly selected samples and features. RFs has two major parts: (1) randomness and (2) ensemble learning. The details of each part are described as below (Breiman, 2001; Breiman, 2003; Liaw and Wiener, 2002).

1.1 Randomness

Given a dataset of size *N* with *M* features, the randomness in RF is to randomly sample from the entire dataset and features for building every single decision tree. Firstly, the dataset of N samples is randomly sampled with replacement for creating a subset of size N'. This process is also known as bootstrap aggregating (bagging). In RF, N' is equal to N, around two thirds of the entire dataset will be chosen as a subset (i.e., around one third of the subset are replicated in the subset). The left samples (the un-chosen samples) in the original dataset are called OOB data. The random selection from the entire features is also performed to select a subset of size m ($m < M$) but without replacement. This parameter needs to be tuned for optimal performance. However, the parameter is not very sensitive to the model performance. The optimal range of it is quite wide. Sometimes even $m = 1$ can provide a good performance. Breiman (2003) found the square root of M can generally give near optimum results for classification tasks and suggested trying a value twice as high and a value half as low for deciding the best one. As for regression tasks, one third of M is suggested (Liaw and Wiener, 2002).

1.2 Ensemble Learning

A subset of size N' with m features is drawn after random selection process. This subset is further used to construct a single decision tree. The method for tree construction is classification and regression tree (CART) but without pruning. The process is repeated K times to grow an ensemble of K trees. All those K trees vote for the final results upon the input data. This is so-called ensemble learning method. All individual decision trees inside the ensemble contribute for final prediction. In classification tasks, the final predicted class is decided based on majority rule. The class receives the most votes is the final result of prediction. In regression tasks, the final predicted value is derived through averaging the results from all individuals.

Since there is no problem of overfitting in RF (Breiman, 2001), a large enough number of decision trees is usually assigned for providing a stable estimate of variable importance (VI). Breiman (2003) suggested setting the number of trees (i.e., K) to 1000 or more and even 5000 for the cases where lots of features are involved (high dimension data). VI estimates are useful by-product obtained from RF which can be used for data exploration and interpretation. VI estimates measure the importance of a variable in both the classification tasks and regression tasks. Two different ways are given to judge VI: (1) error-based VI (EBVI) estimates and (2) impurity-based VI (IBVI) estimates. The EBVI is calculated upon the OOB data. For instance, firstly, the error rate of classification is drawn using OOB data over all the decision trees. Then, a permutation process is performed to permute all values of one selected variable and the error rate of classification is calculated again. If the error rate increases significantly after the permutation, it means the selected variable is relatively important. By comparing the increased error of each variable, the rank of VI for all variables can be decided. As for the regression tasks, the mean squared error (MSE) is used instead of the error rate.

The IBVI is also based on OOB data. For each classification tree, the impurity of a node is measured by Gini impurity which defines a misclassification rate. A higher value of Gini indicates a higher misclassification rate which means the data within a node are more impure. Then, the above permutation process is repeated again to detect the increase in Gini impurity for assessing the importance of a variable. Likewise, if Gini impurity increased notably after the permutation of a specific variable, the variable is more important. As for the regression tasks, the residual sum of squares (RSS) within a node is used as a measure of impurity. Following the above two ideas, RF can estimate the VI of each variable.

3. Rainfall forecasting model development

To investigate the model improvements by considering the typhoon characteristics and meteorological factors, four SVM-based forecasting models with different kinds of input variables are first constructed herein for each rainfall station. They are the model only considering the antecedent rainfalls (SVM-R), the model considering the antecedent rainfalls and typhoon characteristics (SVM-RT), antecedent rainfalls and meteorological factors (SVM-RW), and the model considering the antecedent rainfalls, typhoon characteristics, and meteorological factors (SVM-RTW). The general forms of SVM-R, SVM-RT, SVM-RW, and SVM-RTW, respectively, are as follows.

$$R_{t+\Delta t} = f(R_t, R_{t-1}, \ldots, R_{t-(L_R-1)}) \quad (6)$$

$$R_{t+\Delta t} = f(R_t, R_{t-1}, ..., R_{t-(L_R-1)}, T_t, T_{t-1}, ..., T_{t-(L_T-1)}) \quad (7)$$

$$R_{t+\Delta t} = f(R_t, R_{t-1}, ..., R_{t-(L_R-1)}, W_t, W_{t-1}, ..., W_{t-(L_W-1)}) \quad (8)$$

$$R_{t+\Delta t} = f(R_t, R_{t-1}, ..., R_{t-(L_R-1)}, T_t, T_{t-1}, ..., T_{t-(L_T-1)}, W_t, W_{t-1}, ..., W_{t-(L_W-1)}) \quad (9)$$

where t is the current time, Δt is the lead time period (from 1 to 6 hours), R_t is rainfall at time t, T_t is typhoon characteristics at time t, W_t is meteorological factors at time t, and L_R, L_T, and L_W denote the lag lengths of antecedent rainfall, typhoon characteristics, and meteorological factors, respectively. Typhoon characteristics, T_t, include the position (latitude, longitude) of the typhoon center (°N, °E), the distance between the center and the catchment (km), the maximum wind speed near the center (m/s), the maximum instantaneous wind speed near the center (m/s), the atmospheric pressure of the center (hPa), the radius of winds over 15 m/s (km), and the speed of the typhoon movement (m/s). Meteorological factors, W_t, include air pressure (hPa), air temperature (°C), dewpoint temperature (°C), relative humidity (%), wind velocity (m/s), wind direction (degree), rainfall depth (mm), and the rainfall duration (minute) at the meteorological station.

For each of the four SVM-based forecasting models (i.e. SVM-R, SVM-RT, SVM-RW and SVM-RTW), the input variables and their lag lengths are optimized by using the predictor selection method based on the NSGA-II. The predictor selection method is able to determine which input variable is useful, and the lag length of each useful input variable is optimized. Comparisons between the four SVM-based forecasting models with their optimal combinations of input variables are made and discussed in the section of results and discussions.

4. Predictor selection method by Non-dominated sorting genetic algorithm (NSGA-II)

The predictor selection method based on the NSGA-II is developed in the study. The NSGA-II developed by Deb *et al.* (2002) starts with a randomly generated population of size *s*. The members in this parent population are ranked based on the non-dominated level determined by using non-dominated sorting based on a crowded comparison operator. Through selection (reproduction), crossover, and mutation, an offspring population of equal size *s* is created. Combining the parent and offspring into a new population of size 2*s* is then formed. This procedure allows for elitism to be maintained in successive generations. The members in this new population of size 2*s* are ranked again according to the non-dominated level. The NSGA-II chooses new members of size *s* belonging to the lowest levels as the parent population of next generation. The procedure is repeated until a stable set of Pareto-optimal solutions is obtained. Details about NSGA-II can be found in the study by Deb *et al.* (2002).

In the study, mean coefficient of efficiency (MCE) and mean absolute error (MAE), which are commonly used to evaluate the model performance, are considered as the separate objectives of the multi-objective optimization model (i.e. NSGA-II). The MAE(θ) and MCE(θ) of all typhoon events for a set of input variables (θ) are listed as follows:

$$MAE(\theta) = \frac{1}{N_f}\sum_{i=1}^{N_f}\left[\frac{1}{n_i}\sum_{t=1}^{n_i}\left|R_t^i - \overset{i}{R_t}(\theta)\right|\right] \quad (10)$$

$$MCE(\theta) = \frac{1}{N_f}\sum_{i=1}^{N_f}\left[1-\frac{\sum_{t=1}^{n_i}\left(R_t^i - R_t^i(\theta)\right)^2}{\sum_{t=1}^{n_i}\left(R_t^i - \overline{R}^i\right)^2}\right] \quad (11)$$

where N_f is the fold number of cross validation (N_f =10 herein); i means the i^{th} cross-validation; R_t^i and $R_t^i(\theta)$ denote the observed and forecasted rainfalls at time t, respectively; $\overline{R}^i$ is the average of the observed rainfalls; n_i is the number of forecasts. If the MAE value is closer to zero, the forecast performance is better; if the MCE value is closer to one, the forecast performance is better. The two objective functions, MAE(θ) and 1-MCE(θ), are minimized during the optimization process to obtain the Pareto-optimal solutions of decision variable, θ (i.e. the set of input variables).

Fig. 2 shows the flowchart of the proposed predictor selection method based on NSGA-II, which includes eleven steps described as follows. Some chromosomes of size s (i.e. different combinations of input variables) are first randomly generated in the form of binary code as the initial population (Step 1). Each variable (X) is encoded into four binary digits (i.e. 00, 01, 10, and 11) (Step 2). The four binary digits represent different lag lengths of variable. The binary value "00" represents no consideration of X as input variable; "01" represents X_{t-1}; "10" represents X_{t-1} and X_{t-2}; "11" represents X_{t-1}, X_{t-2}, and X_{t-3}. For example, a model has a set of input variables (θ) including three variables (θ^1, θ^2, and θ^3) and their binary values are 10, 00, and 01, respectively, which represents that the input variables are θ_{t-1}^1 θ_{t-2}^1, and θ_{t-1}^3. The variable θ^2 is not considered as the input variable due to its zero value. For a chromosome (i.e. a combination of input variables, θ), the SVM-based forecasting model for a specific lead time (e.g. 1-hour ahead) is constructed (Step 3), which is trained and tested based on k-fold cross-validation to calculate the MAE(θ) and MCE(θ) of this chromosome (Step 4).

In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as the calibration data for training the model. The cross-validation process is then repeated k times (k folds), with each of the k subsamples used exactly once as the validation data. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used (McLachlan *et al.*, 2004), which is adopted herein. Based on 10-fold cross-validation, 834 hourly data from the thirteen typhoon events (No. 1 to No 13. in Table I) are randomly selected to form 10 equal-length data set (or say, equal sized subsamples). Each data set is selected in turn for validation (testing) and the remaining 9 data set are used for training.

After calculating the values of the two objectives for all chromosomes of the initial population, the fast non-dominated sorting based on a crowded comparison operator (Deb *et al.*, 2002) is executed for further selection (reproduction), crossover, and mutation to obtain the non-dominated solutions (Step 6 and Step 8). The NSGA-II applies these solutions for generating the second offspring (Step 9). The aforementioned calculation is repeated for the second offspring to reach the second non-dominated solutions. Combining the parent and offspring into a new population of size 2s has to be executed after the 2nd iteration (Step 5). Calculations continue to reach the convergence criterion (Step 7). As the convergence criterion (i.e. no improvement in non-dominated solutions after some offspring) is satisfied, the Prato-optimal solutions can be obtained (Step 10). Details on the main steps of the NSGA-II can be found in the study of Deb *et al.* (2002) and Yandamuri *et al.* (2006).

One of the post optimization approaches (Deb, 2001) (i.e. the compromise programming approach) is used herein for selecting the best solution from Pareto-optimal solutions. Among all the Pareto-optimal solutions, one of them is nominated as the best compromise solution, being the solution closest to the ideal solution (the original point of objective space). The ideal solution herein has zero standardized MAE(θ) and the zero standardized IMCE(θ)=1-MCE(θ). The best compromise solution is defined as the solution with the

minimum weighted Euclidean distance (*WED*) to the ideal solution in the objective space (Step 11) (Deb, 2001; Yu *et al.*, 2015) as:

$$min(WED) = min\left(\sqrt{w_1(\mathrm{MAE}(\theta)^i - \mathrm{MAE}(\theta)^{\min})^2 + w_2(\mathrm{IMCE}(\theta)^i - \mathrm{IMCE}(\theta)^{\min})^2}\right) \quad (12)$$

$$w_1 = \left(\frac{1}{\mathrm{MAE}(\theta)^{\max} - \mathrm{MAE}(\theta)^{\min}}\right)^2; \quad w_2 = \left(\frac{1}{\mathrm{IMCE}(\theta)^{\max} - \mathrm{IMCE}(\theta)^{\min}}\right)^2 \quad (13)$$

where w_1 and w_2 are weights; $\mathrm{MAE}(\theta)^i$ and $\mathrm{IMCE}(\theta)^i$ are $\mathrm{MAE}(\theta)$ and $\mathrm{IMCE}(\theta)$, respectively, of the *i*-th Pareto-optimal solution (*i*=1, 2,…, *m*); *m* = the number of Pareto-optimal solutions; $\mathrm{MAE}(\theta)^{\max}$ and $\mathrm{MAE}(\theta)^{\min}$ are the maximum and minimum $\mathrm{MAE}(\theta)$, respectively, in the Pareto-optimal solutions; $\mathrm{IMCE}(\theta)^{\max}$ and $\mathrm{IMCE}(\theta)^{\min}$ are the maximum and minimum $\mathrm{IMCE}(\theta)$, respectively, in the Pareto-optimal solutions.

From the above procedure, the best compromise solution θ_{best} (i.e. the optimal combination of input variables) for a SVM-based forecasting model for a specific lead time (e.g. 1-hour ahead) can be determined. For a rainfall station, each of the SVM-based forecasting models (i.e. SVM-R, SVM-RT, SVM-RW, and SVM-RTW) includes six sub-models for 1- to 6-hour ahead forecasting, respectively. Therefore, there are totally 24 SVM-based forecasting models which have to be constructed and optimized in the study.

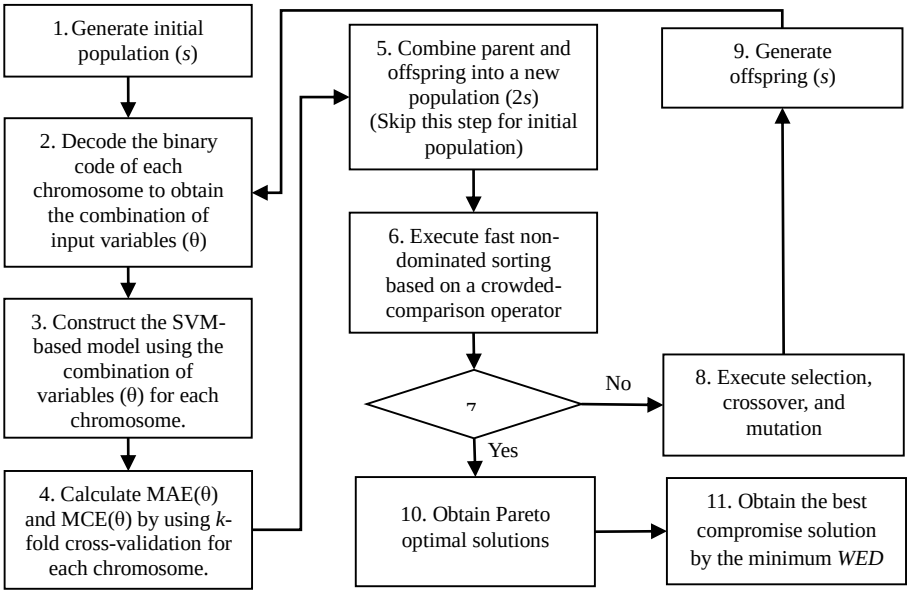

Fig. 2. Flowchart of the predictor selection method based on NSGA-II.

Results and Discussions

1. Optimization of input variables for SVM-R, SVM-RT, SVM-RW, and SVM-RTW

For optimizing the combination of input variables, each variable (X) is encoded into 4 binary digits (i.e. 00, 01, 10, and 11) which represent (1) no consideration of X as input variable, (2) X_{t-1}, (3) X_{t-1} and X_{t-2}, and (4) X_{t-1}, X_{t-2}, and X_{t-3}, respectively. Therefore, for each of 1- to 6-hour ahead forecasting, the numbers of total combinations of input variables for SVM-R, SVM-RT, SVM-RW, and SVM-RTW are $[4^1-1]$, $[4^{(1+8)}-1]$, $[4^{(1+8)}-1]$, and $[4^{(1+8+8)}-1]$, respectively, due to the numbers of input variables of antecedent rainfall, typhoon characteristics, and meteorological factors are 1, 6, and 8, respectively. Herein, the initial population size of 100 and the crossover and mutation rates taken to be 0.9 and 0.1, respectively, are used in the NSGA-II to

derive the Pareto-optimal solutions. In this case study, the NSGA-II reaches the convergence criterion (i.e. there is no improvement in non-dominated solutions after some offspring) after the 20th generation of offspring. The 20th generation of offspring can be considered as the Pareto-optimal solutions due to the Pareto-optimal front is stable. The aforementioned "Pareto-optimal front" is the set of Pareto-optimal solutions which make up a front when viewed them together on the objective space. For 3-hour ahead forecasting at Dajiaoxi station as an example, Fig. 3 shows the offspring for the 1st to the 2nd generations for SVM-R and the offspring for the 1st to the last (20th) generations for SVM-RT, SVM-RW, and SVM-RTW, respectively. Due to only three (4^1-1) variable combinations for SVM-R, the 2nd generation early reaches the convergence criterion. In the Fig. 3, each point represents a combination of input variables (solution) for 3-hour ahead forecasting; the offspring of the last generation (Pareto-optimal solutions) have the smaller values of MAE and (1-MCE) than the values of the offspring of the former generations for each model. The figure also reveals that the offspring of the last generation for SVM-RTW have the smaller values of MAE and (1-MCE) than the values of the offspring of the last generations for SVM-R, SVM-RT and SVM-RW, which means the SVM-RTW with the input variables of the Pareto-optimal solutions performs the best among the four SVM-based forecasting models.

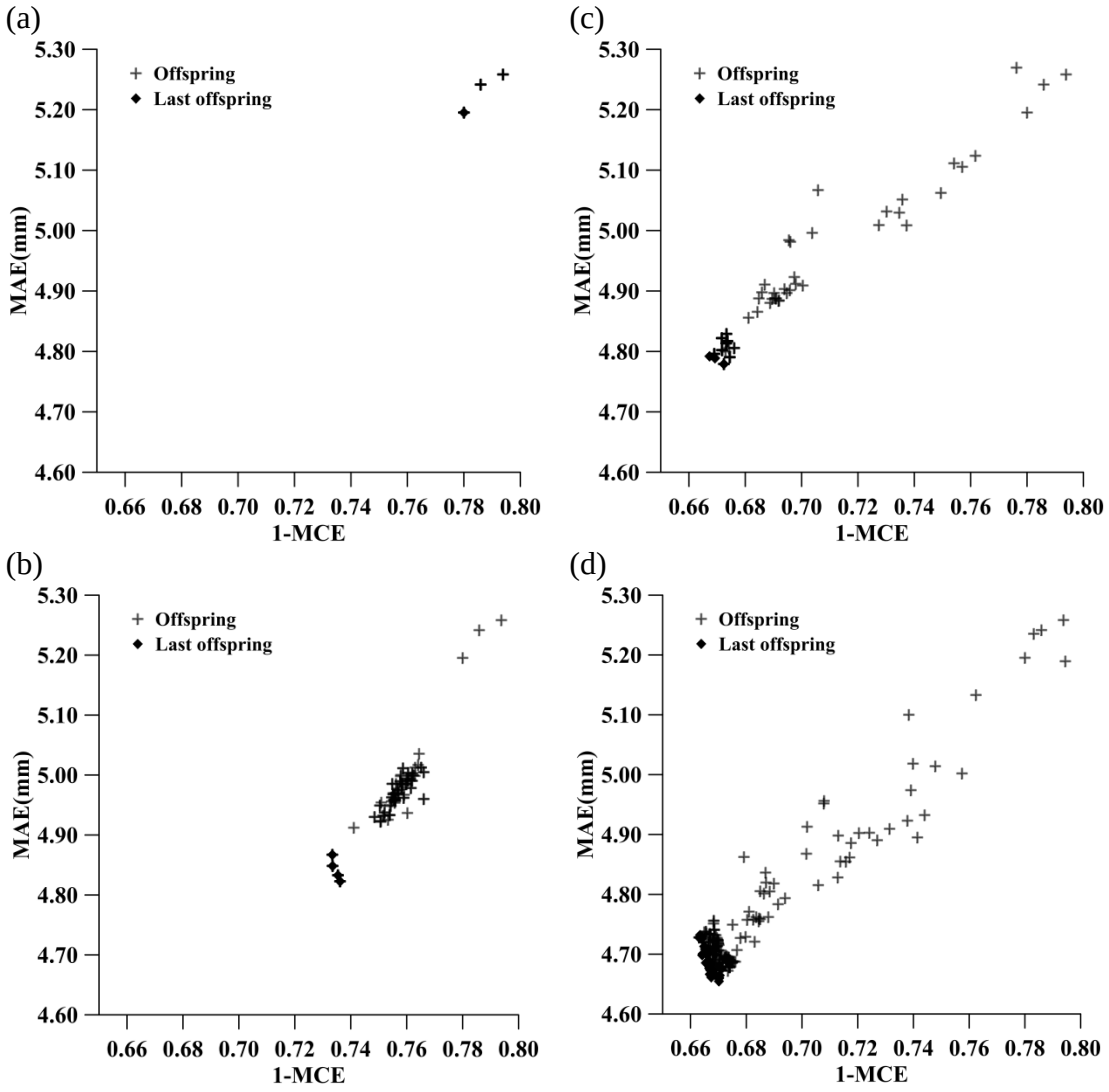

Fig. 3. The offspring for the 1st generation to the last (20th) generation for (a) SVM-R, (b) SVM-RT, (c) SVM-RW, and (d) SVM-RTW, respectively. (an example for 3-hour ahead forecasting at Dajiaoxi station)

Further, Equation (12) is used to calculate the minimum *WED* from the Pareto-optimal solutions to the

ideal solution in the objective space to obtain the best compromise solution (i.e. the optimal combination of input variables) for each of the four SVM-based forecasting models (i.e. SVM-R, SVM-RT, SVM-RW, and SVM-RTW). The best compromise solutions of the four SVM-based forecasting models for 1- to 6-hour lead time, respectively, can be determined at the three rainfall stations. For Dajiaoxi station as an example, the best compromise solutions (i.e. the input variables and their lag lengths of the optimal combinations) for SVM-R for 1 to 6 hours ahead forecasting, respectively, are 3 which means the input variables are R_t, R_{t-1}, and R_{t-2}; the lag lengths of input variables of the optimal combinations for SVM-RT, SVM-RW, and SVM-RTW, respectively, for 1- to 6-hour ahead forecasting are listed in Table II, Table III, and Table IV, respectively. The high correlation coefficient (the absolute value is larger than 0.7) between each pair of the four variables (P_{TY}, V_{MC}, r_{15}, and V_{IC}) is found in the data set; the positive correlations exist for the variable-pairs (V_{MC} and r_{15}), (V_{IC} and r_{15}), and (V_{MC} and V_{IC}); the negative correlations exist for the variable-pairs (P_{TY} and V_{MC}), (P_{TY} and r_{15}), and (P_{TY} and V_{IC}). The optimization results of input variables show that the proposed predictor selection method can reasonably choose one or two variables (not all the four variables) to avoid the repeated information of the high correlated variables used for model construction. For example, Table II shows that r_{15} is not adopted for all lead time forecasting and only one of the three variables (i.e. P_{TY}, V_{MC}, and V_{IC}) is selected for a lead time forecasting. In Table IV, P_{TY} and V_{IC} are not adopted for all lead time forecasting; V_{MC} and r_{15} are not selected for 1-hour ahead forecasting; V_{MC} and r_{15} are simultaneously selected for 2-, 4-, 5- and 6-hour ahead forecasting; V_{MC} is selected for 3-hour ahead forecasting. The above results conclude that the proposed predictor selection method can adequately deal with the problem of high correlated input variables.

Table II. The lag lengths of input variables of the optimal combinations for SVM-RT for 1- to 6-hour lead time forecasting at Dajiaoxi station

Lead time	Variables								
(hour)	R	P_{TY}	N_{TY}	E_{TY}	D	V_{MC}	r_{15}	V_{TY}	V_{IC}
1	3	1	1	2	0	0	0	0	0
2	2	0	1	2	0	0	0	0	1
3	2	0	0	2	2	1	0	0	0
4	3	0	0	3	3	1	0	1	0
5	3	0	0	3	3	1	0	1	0
6	3	0	0	3	3	0	0	1	1

Note: (1) A lag length of 0 means no consideration of the variable "X" as input; 1 means X_{t-1} as input; 2 means X_{t-1} and X_{t-2} as input; and 3 means X_{t-1}, X_{t-2}, and X_{t-3} as input.

(2) R: rainfall depth at rainfall station; P_{TY}: atmospheric pressure of typhoon center; N_{TY}: latitude of typhoon center; E_{TY}: longitude of typhoon center; D: distance between typhoon center and catchment; V_{MC}: maximum wind speed near typhoon center; r_{15}: radius of winds over 15 m/s; V_{TY}: Speed of typhoon movement; V_{IC}: Maximum instantaneous wind speed near typhoon center.

Table III. The lag lengths of input variables of the optimal combinations for SVM-RW for 1- to 6-hour lead time forecasting at Dajiaoxi station

Lead time	Variables								
(hour)	R	P_L	T	T_d	H	V_W	D_W	R_L	t_d
1	1	1	0	0	0	1	1	1	1
2	2	0	3	0	0	1	2	1	1
3	2	0	2	0	2	1	1	0	0
4	3	3	2	0	2	1	3	0	3
5	3	3	3	0	3	1	3	0	3
6	3	2	3	0	3	1	2	0	2

Note: (1) A lag length of 0 means no consideration of the variable "X" as input; 1 means X_{t-1} as input; 2 means X_{t-1} and X_{t-2} as input; and 3 means X_{t-1}, X_{t-2}, and X_{t-3} as input.

(2) R: rainfall depth at rainfall station; the following meteorological variables observed at meteorological station, including P_L: air pressure, T: air temperature, T_d: dew point, H: relative humidity, V_W: wind velocity, D_W: wind direction, R_L: rainfall depth, and t_d : rainfall duration.

Table IV. The lag lengths of input variables of the optimal combinations for SVM-RTW for 1- to 6-hour lead time forecasting for Dajiaoxi station

Lead time (hour)	Variables																
	R	P_L	T	T_d	H	V_W	D_W	R_L	t_d	P_{TY}	N_{TY}	E_{TY}	D	V_{MC}	r_{15}	V_{TY}	V_{IC}
1	1	1	1	1	0	1	3	1	3	0	0	3	0	0	0	0	0
2	1	0	1	0	0	1	2	0	1	0	0	0	0	3	1	2	0
3	1	3	2	1	0	1	2	0	1	0	0	0	2	2	0	3	0
4	1	3	0	0	1	1	3	0	3	0	0	2	3	1	1	3	0
5	1	3	0	0	2	1	3	0	3	0	0	2	3	1	2	2	0
6	3	3	3	0	2	2	3	0	3	0	0	2	3	1	2	1	0

Note: A lag length of 0 means no consideration of the variable "X" as input; 1 means X_{t-1} as input; 2 means X_{t-1} and X_{t-2} as input; and 3 means X_{t-1}, X_{t-2}, and X_{t-3} as input. The abbreviations of variables are shown in the notations of Table II and Table III.

Moreover, the optimal combination of input variables for SVM-RTW can be efficiently obtained by using the fast elitist NSGA-II. Due to 17 input variables and 4 lag lengths are used in SVM-RTW, the number of total input-variable combinations for SVM-RTW is huge, which approximately equals to 4^{17} (i.e. 1.718×10^{10}). Trying all the input-variable combinations to construct the forecasting models for finding the optimal one is such a tedious and time-consuming task. Using the NSGA-II in which the initial population size=100 and iteration number=20, the optimal combination of input variables can be efficiently searched by less than 100×20 (i.e. 2000) times of model construction.

2. Comparisons of performance among SVM-R, SVM-RT, SVM-RW, and SVM-RTW

Using the best compromise solutions of the four SVM-based forecasting models for 1- to 6-hour lead time, respectively, the values of MAE and MCE for SVM-R, SVM-RT, SVM-RW, and SVM-RTW for 1- to 6-hour lead time can be estimated. The results for the three rainfall stations (Dajiaoxi, Zailian, and Shuanglianbi stations) are shown in Fig. 4. For the Dajiaoxi station, Fig. 4(a) shows that the MAE values of all the four models (SVM-R, SVM-RT, SVM-RW, and SVM-RTW) increase with increasing forecast lead time. Results also show that the SVM- RTW yields the lower MAE than the other three SVM-based forecasting models (SVM-R, SVM-RT, and SVM-RW) for each lead time forecasting. Figure 4(d) shows that the MCE values of all the four SVM-based forecasting models decrease with increasing forecast lead time, but the SVM-RTW yields the higher MCE than the other three methods (SVM-R, SVM-RT, and SVM-RW) for each lead time forecasting. Overall, the two performance measures (MAE and MCE) for each lead time forecasting show that the SVM-RTW performs the best, which clearly indicates that the simultaneous addition of typhoon characteristics and meteorological factors as input variables actually improves the hourly typhoon rainfall forecasting. Similar results are found for the other two rainfall stations (i.e., Zailian station in Fig. 4(b) and Fig. 4(e) and Shuanglianbi station in Fig. 4(c) and Fig. 4(f)). From the comparisons of performances among the four SVM-based forecasting models, it reveals that the simultaneously considering typhoon characteristics and meteorological factors as input variables is positively helpful for hourly typhoon rainfall forecasting.

To investigate the improvement of forecasting performance due to the addition of typhoon characteristics or/and meteorological factors, the improvements of SVM-RT, SVM-RW, and SVM-RTW as compared to SVM-R are discussed herein. The improvement results in MAE and MCE for the three rainfall stations (Dajiaoxi, Zailian, and Shuanglianbi stations) are shown in Fig. 5. From the figure, the improvements in MAE and MCE increase with increasing forecast lead time for each station. For Dajiaoxi station, Fig. 5(a) shows the SVM-RTW has the greatest improvements of MAE for each lead time among all the three SVM-based forecasting models except for 1-hour ahead forecasting (SVM-RW and SVM-RTW have a close improvement); SVM-RW has better improvements than SVM-RT for 1- to 4-hour lead time; SVM-RT has better improvements than SVM-RW for 5- to 6-hour lead time. Figure 5(d) shows that the SVM-RTW has the greatest improvements of MCE for each lead time; SVM-RW has close improvements to SVM-RTW for 1- to 3-hour lead time and has better improvements than SVM-RT for all lead time.

For Zailian and Shuanglianbi stations, respectively, Fig. 5(b) and Fig. 5(c) show the SVM-RTW has the greatest improvements of MAE for each lead time among all the three SVM-based forecasting models at the two rainfall station. In Fig. 5(e), except for 1- and 5-hour ahead forecasting, the SVM-RTW has the greatest

improvements of MCE at Zailian station for the other lead time. In Fig. 5(f), except for 1-hour ahead forecasting (SVM-RW and SVM-RTW have a close improvement), the SVM-RTW has the greatest improvements of MCE at Shuanglianbi station for the other lead time.

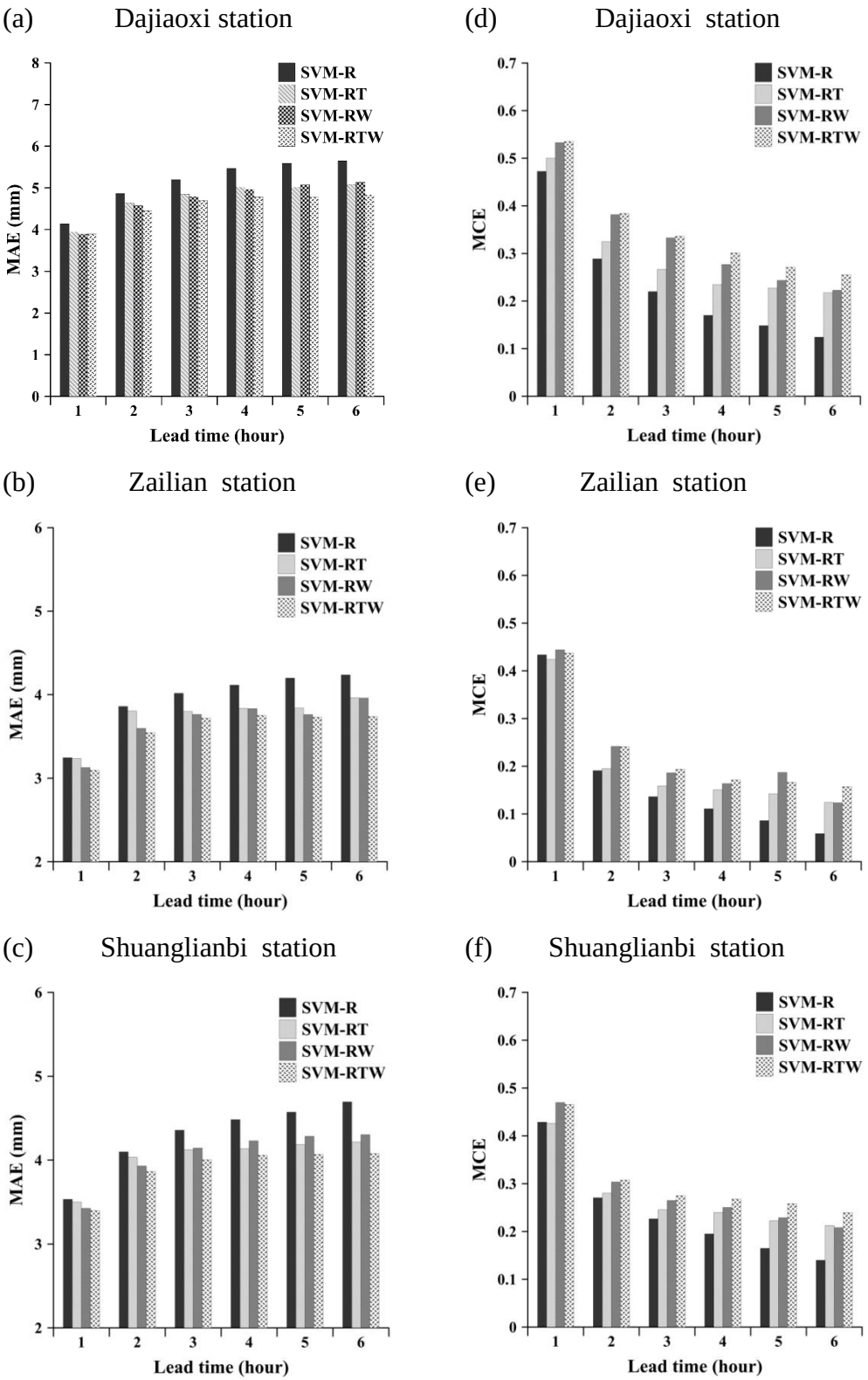

Fig 4. Values of MAE (a)(b)(c) and MCE (d)(e)(f) for the four optimal SVM-based forecasting models for 1- to 6-hour ahead forecasting at Dajiaoxi, Zailian, and Shuanglianbi stations.

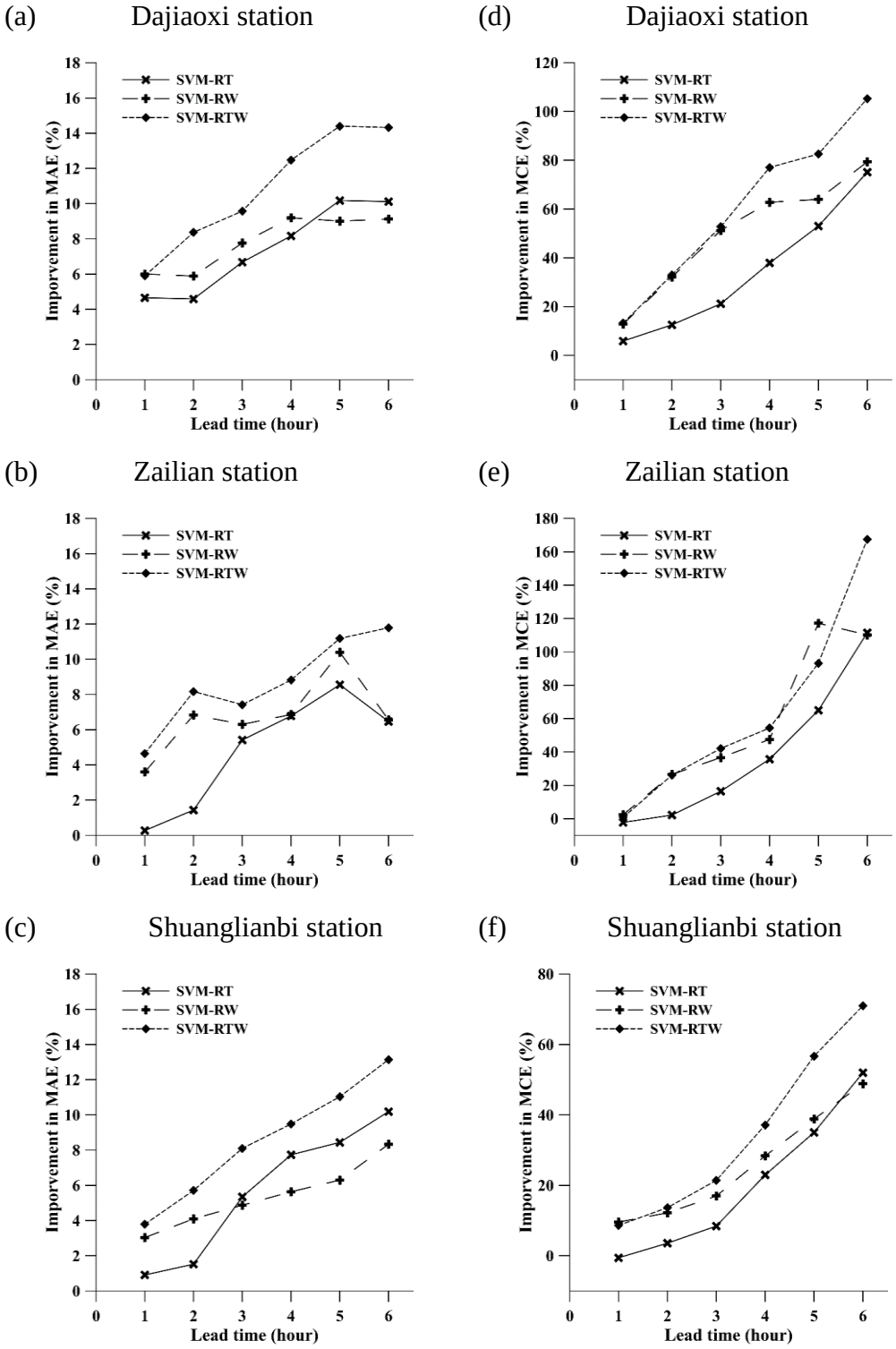

Fig 5. Improvements in MAE (a)(b)(c) and MCE (d)(e)(f) for the optimal SVM-RT, SVM-RW, and SVM-RTW models for 1- to 6-hour lead time at Dajiaoxi, Zailian, and Shuanglianbi stations.

Overall, the proposed three SVM-based forecasting models (i.e., SVM-RT, SVM-RW, and SVM-RTW) significantly improves the forecasting performance more than SVM-R, especially for the long lead time forecasting; the improvements of SVM-RTW are the greatest for all lead time forecasting and the improvements of SVM-RW are better than SVM-RT for most of lead time forecasting; the improvements of SVM-RT in either MAE or MCE are the lowest for most of lead time forecasting. The above findings reveal that simultaneous considering typhoon characteristics and meteorological factors as input variables for model construction (i.e., SVM-RTW) is more helpful for typhoon rainfall forecasting than only considering typhoon

characteristics (i.e., SVM-RT) or only considering meteorological factors (i.e., SVM-RW). Additionally, Typhoon Soudelour in 2015 is used for illustrating the hourly rainfall forecasting by the four optimal SVM-based forecasting models. Figure 6 shows their 1- to 6-hour lead time forecasts at Dajiaoxi station. It is found that the forecasting accuracy decreases as the lead time increases. For 1- and 2-hour lead time forecasting, SVM-RTW and SVM-RW have a close forecasting performance and perform better than SVM-R and SVM-RT. For 3- to 6-hour lead time forecasting, SVM-R seems to underestimate the rainfalls the most before the 18th hour. For 3- to 5-hour lead time forecasting, SVM-RT overestimates the rainfalls the most before 10 hours and after the 18th hour. For 4- to 6-hour lead time forecasting, SVM-RTW generally performs better than SVM-RW during the heavier rainfall period (from the 10th hour to the 20th hour). This test results are satisfactory and correspond with the aforementioned higher improvements of SVM-RTW for long lead time forecasting.

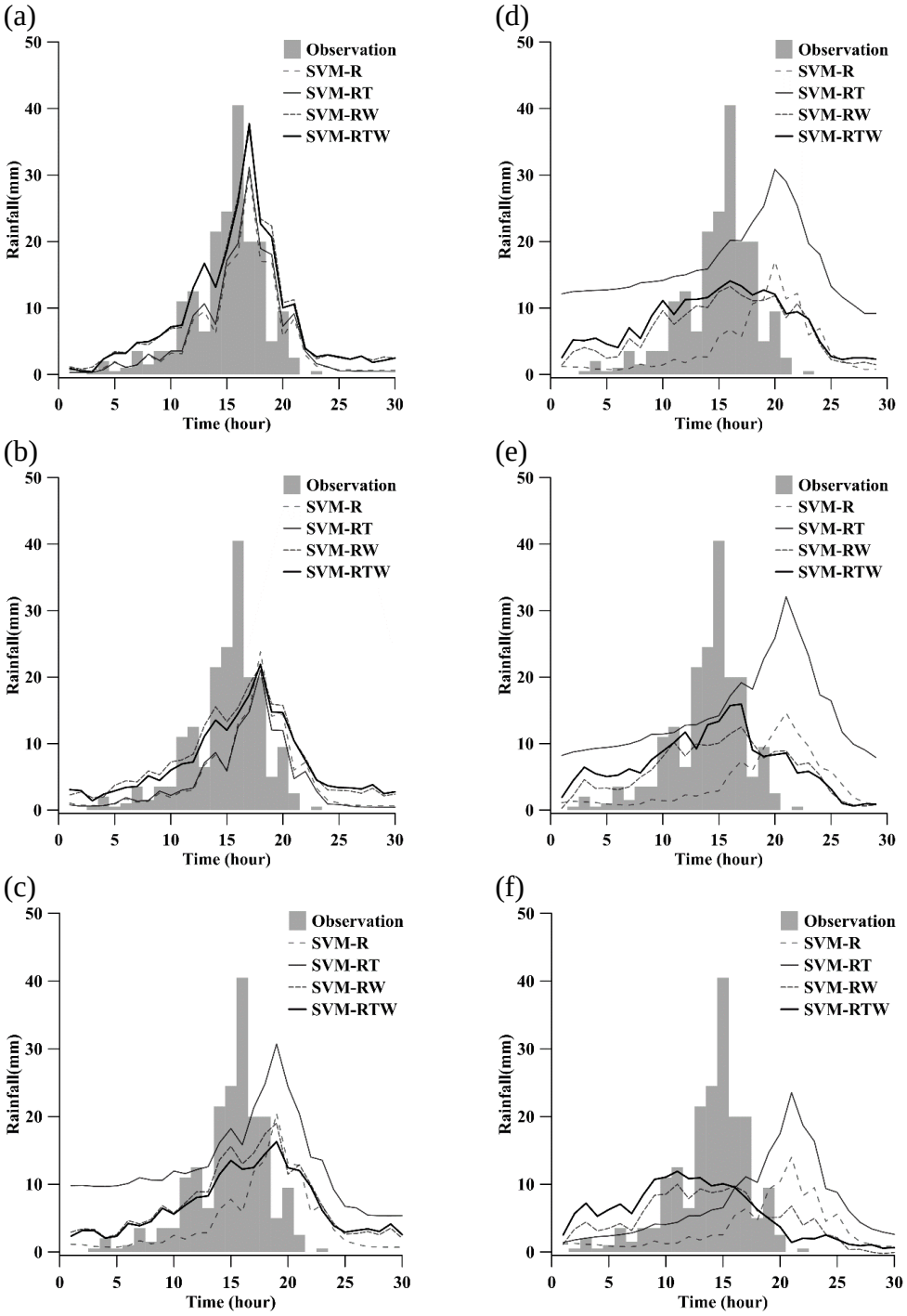

Fig. 6. Comparisons between observed and forecasted rainfalls by SVM-R and SVM-RTW for (a) 1-, (b) 2-, (c) 3-, (d) 4-, (e) 5-, and (f) 6-hour lead time forecasting at Dajiaoxi station during Typhoon Soudelour (2015).

3. Comparisons of performance between SVM-RTW and RF-RTW

Obviously, the predictor selection method shows efficient and acceptable performance in deciding predictors. Figure 6 reveals the SVM-based model considering the antecedent rainfalls, typhoon characteristics, and meteorological factors is the best predictor combination into the predictor selection method. Further, we replace the ML by RFs to combine predictor selection method and construct a RF-based model called RF-RTW. Figure 7 shows the improvements in MAE and MCE for the optimal RF-RTW and SVM-RTW models compared with SVM-R for 1- to 6-hour lead time at Dajiaoxi (a)(d), Shuanglianbi (b)(e), and Zailian (c)(f) stations. The result stands a strong proof the performance of RF-RTW for 2- to 6-hour lead time is better than the performance of SVM for 2- to 6-hour lead time.

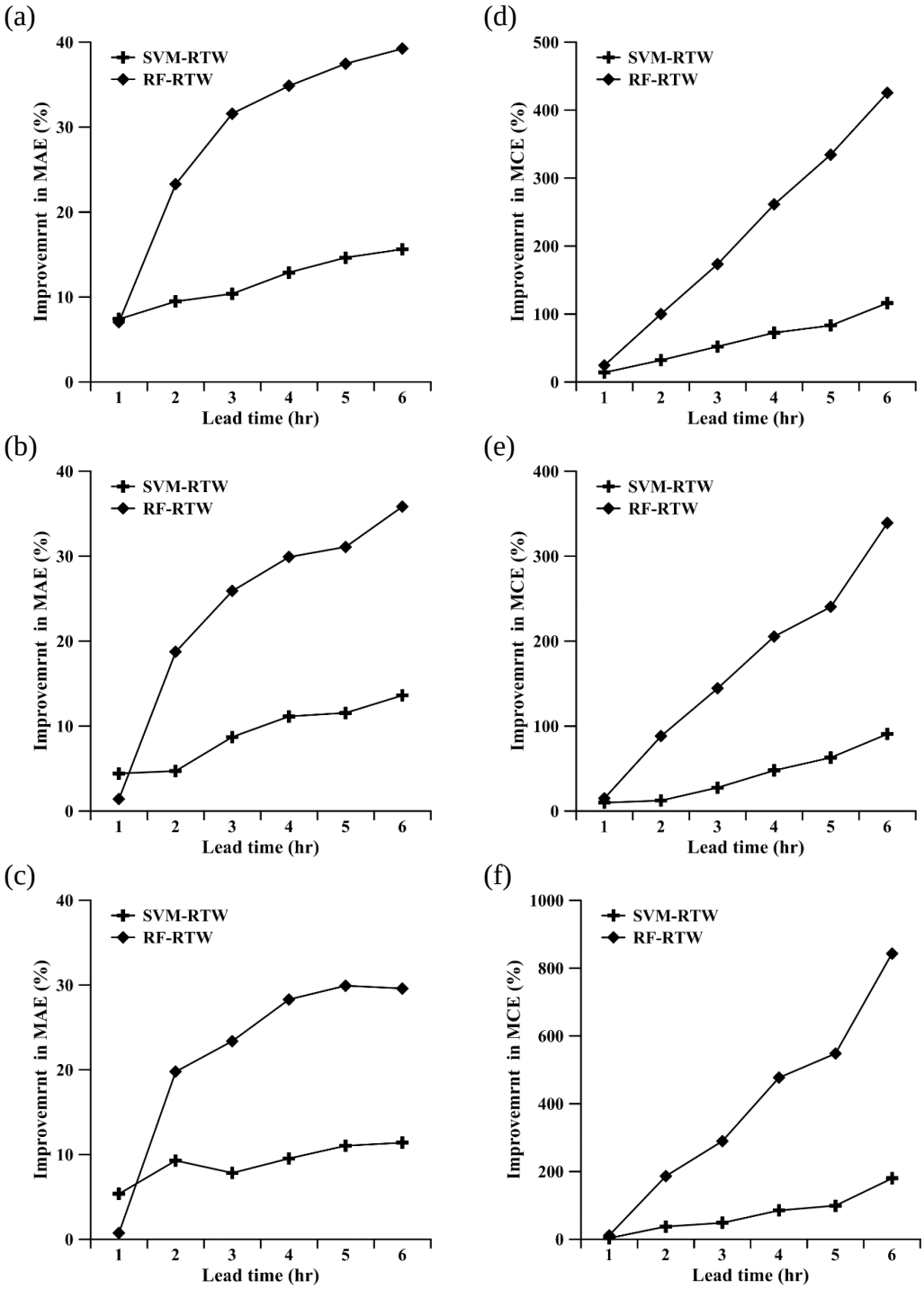

Fig. 7. Improvements in MAE and MCE for the optimal RF-RTW and SVM-RTW models compared with SVM-R for 1- to 6-hour lead time at Dajiaoxi (a)(d), Shuanglianbi (b)(e), and Zailian (c)(f) stations.

Furthermore, Typhoon Soudelour in 2015 is also used for demonstrating the hourly rainfall forecasting by the SVM-based (SVM-RTW) and RF-based (RF-RTW) forecasting models. Figure. 8 shows their 1- to 6-hour lead time forecasts at Dajiaoxi station. It is found that the forecasting accuracy keeps an acceptable performance as the lead time increases on RF-RTW. For 1-hour lead time forecasting, SVM-RTW and RF-RTW have a similar forecasting performance. For 2- to 3-hour lead time forecasting, it appears the lag problem on SVM-RTW but doesn't on RF-RTW. For 4- to 6-hour lead time forecasting, lag problem of SVM-RTW disappears but SVM-RTW underestimates the peak rainfall. Generally, RF-RTW does a better performance than SVM-RTW does. Also, this test results are satisfactory and correspond with the aforementioned higher improvements of RF-RTW for long lead time forecasting.

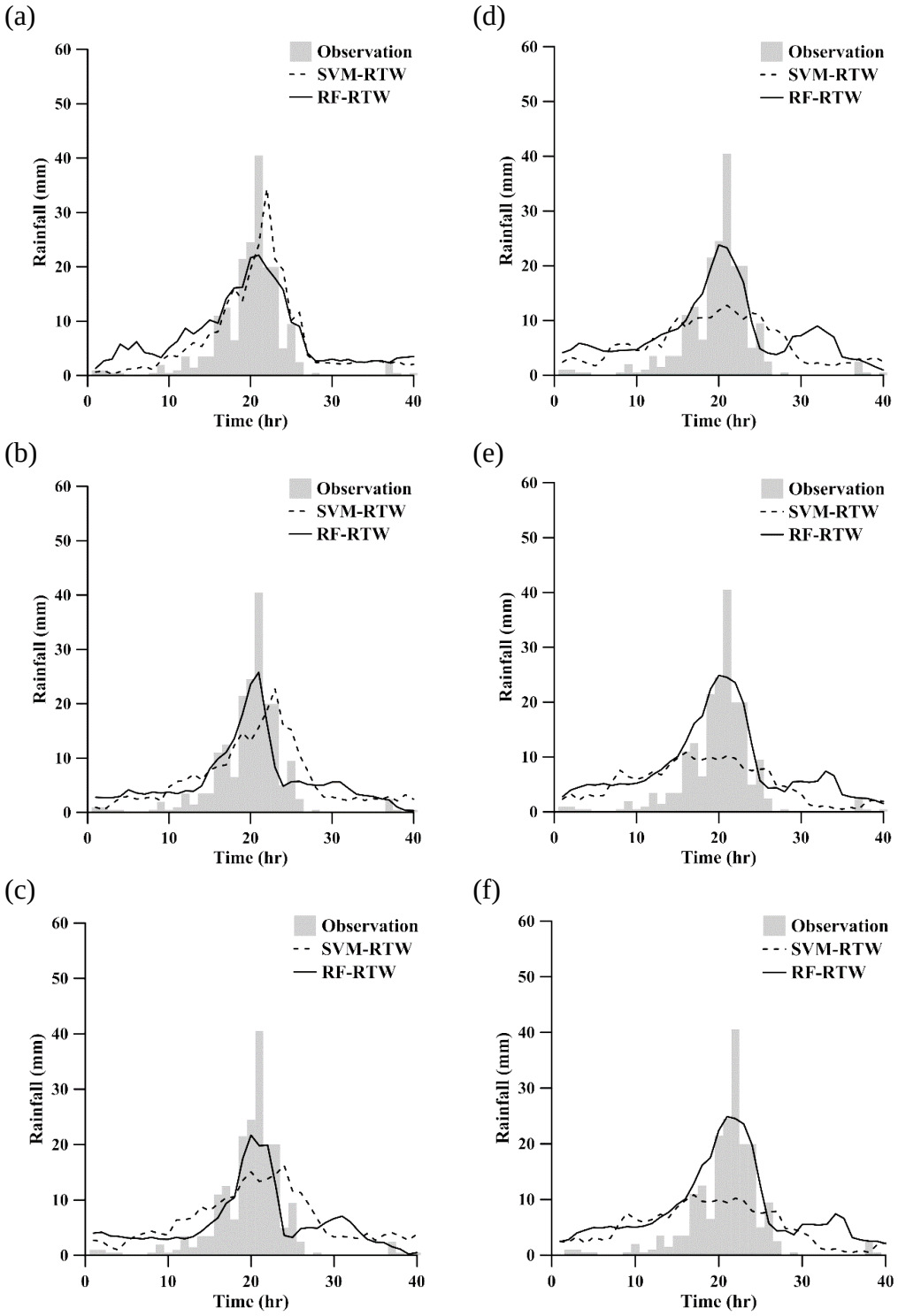

Fig. 8. Comparisons between observed and forecasted rainfalls by SVM-RTW and RF-RTW for (a) 1-, (b) 2-, (c) 3-, (d) 4-, (e) 5-, and (f) 6-hour lead time forecasting at Dajiaoxi station during Typhoon Soudelour (2015).

Conclusions

This study proposed a predictor selection method for effective and efficient construction of the hourly typhoon rainfall forecasting model using a fast elitist non-dominated sorting genetic algorithm (i.e. NSGA-II). Four SVM-based forecasting models with four different sets of input variables: (1) antecedent rainfalls, (2) antecedent rainfalls and typhoon characteristics, (3) antecedent rainfalls and meteorological factors, and (4) antecedent rainfalls, typhoon characteristics and meteorological factors, respectively, are constructed to yield 1- to 6-h ahead forecasts at three rainfall stations in Yilan River basin, northeastern Taiwan. For each station, using the NSGA-II, the optimal combinations of predictors (input variables) for the four SVM-based forecasting models can be effectively and efficiently determined for 1- to 6-hor ahead forecasts. The results show that the model (i.e., SVM-RTW) using all the three kinds of variables (antecedent rainfalls, typhoon characteristics, and meteorological factors) as input variables performs best among four SVM-based forecasting models (i.e., SVM-R, SVM-RT, SVM-RW, and SVM-RTW). The proposed forecasting model, SVM-RTW, can significantly improve hourly typhoon rainfall forecasting, especially for the long lead time forecasting. The above findings reveal that simultaneously considering typhoon characteristics and meteorological factors as input variables for model construction is more helpful for typhoon rainfall forecasting than considering only typhoon characteristics or only meteorological factors. Moreover, the optimal combination of input variables for SVM-RTW can be efficiently obtained by using the fast elitist NSGA-II. The proposed predictor selection method based on NAGA-II is expected to be useful for constructing an effective typhoon rainfall forecasting model with numerous input variables for disaster warning systems. Furthermore, this study applied another ML method called random forests (RFs) and combined the predictor selection method to construct better performance of typhoon rainfall forecasting models (RF-RTW). The test results reveal the performances of RF-RTW in predicting peak rainfall and total volume of rainfall are better than SVM-RTW. Also, the lag problem in SVM does not appear in RF. Generally, RF with the predictor selection method showed the best performance in this study.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. DOI: 10.1023/A:1010933404324
2. Breiman, L. (2003). Manual: setting up, using and understanding Random Forests V4.0. The manual is available online at http://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf
3. Deb, K. (2001). *Multi-objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons: Chichester, England.
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6**(2): 182–197.
5. Chen, S.T. and Yu, P.S. (2007). Pruning of support vector networks on flood forecasting. Journal of Hydrology 347(1–2): 67–78.
6. Hong, W.C., Pai, P.F. (2007). Potential assessment of the support vector regression technique in rainfall forecasting, Water Resour. Manage., 21, 495–513, doi:10.1007/s11269-006-9026-2.
7. Lin, G.F., Chen, G.R. (2008). A systematic approach to the input determination for neural network rainfall-runoff models. *Hydrological Processes* **22**(14): 2524–2530.
8. Lin, G.F., Chen, G.R., Wu, M.C., Chou, Y.C. (2009). Effective forecasting of hourly typhoon rainfall using support vector machines. *Water Resources Research* **45**: W08440
9. Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. R News, 2/3, 18-22.
10. McLachlan, G., Do, K.A., Ambroise, C. 2004. *Analyzing Microarray Gene Expression Data*. John Wiley & Sons: New York.
11. Vapnik, V.N. (1995). The Nature of Statistical Learning Theory, Springer-Verlag: New York.

12. Vapnik., V.N. (1998). Statistical Learning Theory, Wiley: New York.
13. Yandamuri, S.R.M., Srinivasan, K., Bhallamudi, S.M. (2006). Multiobjective optimal waste load allocation models for rivers using non-dominated sorting genetic algorithm-II. *Journal of Water Resources Planning and Management* **132**(3): 133–143.
14. Yu, P.S., Chen, S.T. and Chang, I.F. (2006). Support vector regression for real-time flood stage forecasting. Journal of Hydrology 328(3–4): 704–716.
15. Yu, P.S., Yang, T.C., Kuo, C.M., Tai, C.W. (2015). Integration of physiographic drainage-inundation model and nondominated sorting genetic algorithm for detention-pond optimization. *Journal of Water Resources Planning and Management* **141**(11): 04015028.