Daily River Flow Forecasting with Hybrid Support Vector Machine – Particle Swarm Optimization

Daily River Flow Forecasting with Hybrid Support Vector Machine – Particle Swarm Optimization

N Zaini[1,2], M A Malek[1,2], M Yusoff[3,4], N H Mardi[1,2] and S Norhisham[1]

[1]Department of Civil Engineering, Universiti Tenaga Nasional, Selangor, Malaysia
[2]Institute of Sustainable Energy, Universiti Tenaga Nasional, Selangor, Malaysia
[3]Institute of Infrastructure Engineering and Sustainable Management (IIESM), Universiti Teknologi MARA (UiTM), Selangor, Malaysia
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Selangor, Malaysia

E-mail: Nur_Atiah@uniten.edu.my

Abstract. The application of artificial intelligence techniques for river flow forecasting can further improve the management of water resources and flood prevention. This study concerns the development of support vector machine (SVM) based model and its hybridization with particle swarm optimization (PSO) to forecast short term daily river flow at Upper Bertam Catchment located in Cameron Highland, Malaysia. Ten years duration of historical rainfall, antecedent river flow data and various meteorology parameters data from 2003 to 2012 are used in this study. Four SVM based models are proposed which are SVM1, SVM2, SVM-PSO1 and SVM-PSO2 to forecast 1 to 7 day ahead of river flow. SVM1 and SVM-PSO1 are the models with historical rainfall and antecedent river flow as its input, while SVM2 and SVM-PSO2 are the models with historical rainfall, antecedent river flow data and additional meteorological parameters as input. The performances of the proposed model are measured in term of RMSE and R^2. It is found that, SVM2 outperformed SVM1 and SVM-PSO2 outperformed SVM-PSO1 which meant the additional meteorology parameters used as input to the proposed models significantly affect the model performances. Hybrid models SVM-PSO1 and SVM-PSO2 yield higher performances as compared to SVM1 and SVM2. It is found that hybrid models are more effective in forecasting river flow at 1 to 7 day ahead at the study area.

1. Introduction

Rainfall and river flow are the main factors that contribute to floods. In Malaysia, river flow resulted from rainfall is an important source of water. However, heavy continuous rainfall and river flow effected by additional meteorology parameters such as evaporation, sunshine hours, humidity, and temperature have frequently lead to disaster. The parameters are significantly affect the total river flow [1-3]. Besides that, forecasting of river flow is found to be very difficult to measure due to nonlinear, time varying, and indeterminate of river flow data [2, 4-5].

Various statistical forecast and artificial intelligence (AI) models are developed in river flow forecasting. Artificial neural network (ANN) is widely and successfully used in river flow forecasting as it has the ability of mapping the nonlinear data [6-9]. However, ANN do have some drawbacks such as overfitting, subject to local convergence and slow learning [10-11]. Support vector machine (SVM) are developed as an alternative to conventional statistical and ANN model in forecasting of river flow

[12-14]. In some cases, SVM provides better estimation as compared to ANN and other conventional methods in rainfall runoff modelling [15-18].

Having a growing number of applications in river flow forecasting, SVM is proved to be a reliable method to compute large nonlinear time series data [6, 11, 16-17, 19-20]. The performance of SVM mostly depends on the choices of kernel function and hyper parameter namely soft margin C, ε-insensitive loss function and kernel function γ [21, 14]. RBF kernel seems to be the most common kernel function used in development of SVM model [11, 22]. However, SVM has certain drawbacks where, in order to determine these parameters, large time consumption and sufferings from dimensionality during data analysis are experienced [4, 13]. Thus, to find the optimum SVM parameters, particle swarm optimization (PSO) techniques is implemented in SVM based forecasting model resulting in a hybrid model, SVM-PSO.

Based on the above statements, this study proposed the development of river flow forecasting model using SVM as an alternative technique to the conventional statistical and ANN models. It is estimated that SVM will overcome the drawbacks of ANN. Besides that, the impacts of additional meteorology parameters to forecasting of river flow are investigated.

2. Support Vector Machine for Regression

In the early 1990s, SVMs were developed for classification then it was extended for regression purpose [23]. There are several advantages of SVM as stated by Lin et al. [13] which are SVMs is constructed using structural risk minimization (SRM) induction principle. As SRM induction principle applied in SVM, both empirical risk and model complexity should be minimized simultaneously. The use of SRM induction principle produces better generalization ability of SVMs. Besides that, SVMs based on SRM basically involves quadratic programming problem which can produces optimum results and shorten the time consumed. The methodology of support vector regression (SVR) used in this study is briefly described [4, 13, 23-25].

Based on n training data $[(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)]$, where $n = 1,2,\ldots,n$, x is the input, y is the output. The aim of SVR is to find a non-linear regression function to yield the output $\hat{y}$, which is the best approximation of the desired output $\boldsymbol{y}$ with an error tolerance of ε.

To learn non-linear relations with linear machine, a fixed non-linear, $\boldsymbol{\phi}(\boldsymbol{x})$ mapping of the data to a feature space is applied. Hence, the regression function can be written as

$$\hat{y} = f(x) = w\boldsymbol{\phi}(\boldsymbol{x}) + b \tag{1}$$

where $\boldsymbol{w}$ and b are weights and bias of the regression function, respectively. Parameter $\boldsymbol{w}$ and b are estimated by minimizing the following structural risk function based on SRM induction principle:

$$R = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n} L_\varepsilon(\hat{y}_i) \tag{2}$$

where the Vapnik's ε-insensitive loss function L_ε is defined as

$$L_\varepsilon(\hat{y}) = |y - \hat{y}|_\varepsilon = \begin{cases} 0 & for\ |y - \hat{y}| < \varepsilon \\ |y - \hat{y}| - \varepsilon & for\ |y - \hat{y}| \geq \varepsilon \end{cases} \tag{3}$$

The first and second terms in equation (3) represent the model complexity and the empirical error, respectively. The trade-off between the model complexity and the empirical error is specified by a user-defined parameter C. ε is called the tube size and equivalent to the approximation accuracy placed on the training data points. Both C and ε are user determined parameters [12].

Sequential minimal optimization (SMO) for regression algorithm is chose as learning algorithm in this study. Therefore, SVM adopted in this study will solve the optimization problem through SMOreg [4, 24 - 25]

$$\text{Minimize} \qquad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i') \tag{4}$$

$$\text{subject to} \quad \begin{cases} y_i - (w\phi(x_i) + b) \leq \varepsilon + \xi_i \\ (w\phi(x_i) + b) - y_i \leq \varepsilon + \xi_i' \\ \xi_i, \xi_i' \geq 0 \qquad i = 1,2,\dots,l \end{cases}$$

where ξ_i and ξ_i' are slack variables, representing the upper and lower training errors respectively. The optimization problem can be solved in its dual form using Lagrange multiplier and taking into account that $\xi_i \xi_i' = 0$. Therefore with the same relation $\alpha_i, \alpha_i' = 0$ where α_i, α_i' are Lagrange multiplier.

$$\begin{aligned} &\text{Maximize} \quad \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i') - \varepsilon \sum_{i=1}^{N}(\alpha_i + \alpha_i') \\ &\qquad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i')\left(\alpha_j - \alpha_j'\right)(\langle x_i \cdot x_j \rangle + \frac{1}{C}\delta_{ij}) \qquad (5) \\ &\text{subject to} \quad \sum_{i=1}^{N}(\alpha_i - \alpha_i') = 0 \\ &\qquad \alpha_i \geq 0, \alpha_i' \geq 0 \qquad i = 1,2,\dots,l \end{aligned}$$

In this study radial basis kernel (RBF) is used as kenel function. The equation for kernel function can be written as [13, 21, 26-27]

$$K(x_i, x) = exp(-\gamma |x_i - x|^2) \qquad (6)$$

where x_i denote the *i*th support vector and *x* is input vector. γ is RBF parameter which gives the width of the kernel.

3. Particle Swarm Optimization

Particle swarm optimization (PSO) is an optimization technique with a population-based search algorithm that is based on the metaphor of social behaviour [28]. PSO is implemented in this study to optimize the SVM parameter namely gamma (γ) [4, 12]. γ is a SVM user determined parameter thus, determination of the parameter selection technique is an important issue. SVM has certain drawbacks in determining the optimal value of γ. Therefore, this study proposed PSO technique in optimizing parameter γ.

Initially, the value of parameter γ is specified randomly. The value is then, is fed into SVM model for training and testing purposes. Next, the fitness function is evaluated. In this study, the fitness criterion used is relative mean square error (RMSE). The fitness evaluation of the particles are compared with particle's personal best (*pbest*) value. If the current value is better than *pbest*, *pbest* value is then set to current value and the position of *pbest* is equal to current position in dimensional space. Next, the fitness evaluation is compared with the population's previous overall best for global best (*gbest*) update. If the current value yield better than *gbest*, *gbest* is reset to current value of the particle. The particle file will change to new position by calculating the velocity according to equations (7) and (8) [4, 28].

Velocity,

$$v_{ij}(t) = \left[\omega v_{ij}(t-1) + c_1 r_1(pbest_{ij}(t-1) - x_{ij}(t-1) + c_2 r_2(gbest_j(t-1) - x_{ij}(t-1)\right] \qquad (7)$$

Position,

$$x_{ij}(t) = x_{ij}(t-1) + v_{ij}(t) \qquad (8)$$

where $v_{ij}(t)$ is the velocity of particles *i* in dimension $j = 1,\dots,n_x$ at time step *t*, $x_{ij}(t)$ is the position of particles *i* in dimension *j* at time step *t*, c_1 and c_2 are positive acceleration constant which used to determine how much the *pbest* and *gbest* influence its movement, r_1 and r_2 are random real numbers between 0 to 1. The random values introduce a stochastic element to the algorithm. Next, ω presents the inertia weight to control the impact of velocity's history on the current value. The RMSE values are repeatedly determined until the stopping criterion condition is met. The stopping condition used in this study is to terminate the iterative search process is when the number of iteration reached its maximum allowable [4, 28].

4. Study Area and Data Used

The study area is Bertam catchment located at Cameron Highlands in the state of Pahang, West Malaysia. The catchment has a total area of 108.15 km^2. Nevertheless, this study concerns only at the upstream of this catchment at an area of 33.55 km^2. Figure 1 shows the study area and the location of rainfall station and river flow station chosen.

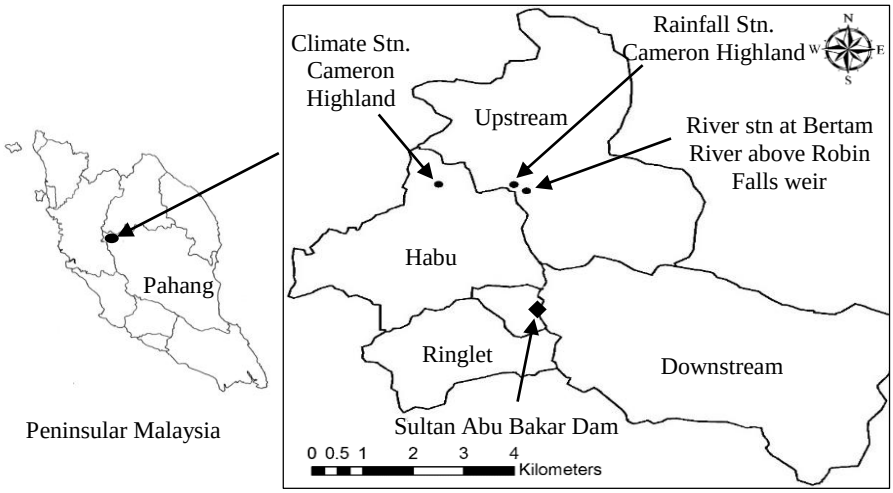

Figure 1. Study area

Total of 10 years data from 1 January 2003 to 31 December 2012 were used in the development of the proposed model. All data used namely rainfall, river flow and other meteorology data are at daily interval. Other than rainfall, various meteorology data being used as input to the model include minimum temperature (Celcius), maximum temperature (Celsius), mean relative humidity (percent), evaporation (mm) and mean wind speed (m/s). Table 1 summarizes the dataset, date and number of data for rainfall, river flow and additional meteorology parameters used.

Table 1. Description of dataset used

Variable	Dataset	
	Training (80%)	Testing (20%)
Rainfall River flow Additional Meteorology Parameters i. Min temperature ii. Max temperature iii. Mean relative humidity iv. Evaporation v. Mean wind speed	1 Jan 2003 – 31 Dec 2010 (2922 days)	1 Jan 2011 – 31 Dec 2012 (731 days)

5. Model Development and Input Design

In order to conduct comparison between the non-hybrid and hybrid model, two kinds of SVM based models which are SVM and SVM-PSO forecasting models are constructed to yield 1 to 7 day ahead of river flow forecasting. In addition, to evaluate the improvement in forecasting performance due to the addition of meteorology data used, two different types of model inputs are designed. First is with rainfall and river flow input data for SVM1 and SVM-PSO1 and second is with rainfall, river flow and other various meteorology data used as input to SVM2 and SVM-PSO2.

The SVM1 and SVM-PSO1 models can be expressed in a general form as

$$RF_{t+\Delta t} = f(RF_t, RF_{t-1}, \dots, RF_{t-(L_{RF}-1)}, R_t, R_{t-1}, \dots, R_{t-(L_R-1)}) \quad (9)$$

where *t* is the current time, Δt is the lead-time period (from 1 to 7 day), RF_t and R_t are river flow and rainfall at time *t*, respectively. Besides, L_{RF} and L_R denote the lag length of river flow and rainfall, respectively.

Based on SVM1 and SVM-PSO1, additional meteorology parameters are added to develop SVM2 and SVM-PSO2 models. The form of SVM2 and SVM-PSO2 is

$$RF_{t+\Delta t} = f\left(RF_t, RF_{t-1}, \dots, RF_{t-(L_{RF}-1)}, R_t, R_{t-1}, \dots., R_{t-(L_R-1)}, MT_t, MT_{t-1}, \dots., MT_{t-(L_{MT}-1)}\right) \quad (10)$$

where MT_t is meteorology parameters at time *t* and L_{MT} denotes the lag length of meteorology parameters. All meteorology parameters namely maximum temperature, minimum temperature, evaporation, relative humidity and mean wind speed are added to the input. For all models, it should be noted that the lag length (L_{RF}, L_R, L_{MT}) are constant for lead time Δt.

6. Result and Discussion

In this study, SVM and SVM-PSO models are developed to evaluate the performance in river flow forecasting. The models are developed based on two scenarios, (i) historical rainfall and antecedent river flow data and (ii) historical rainfall and antecedent river flow data with other additional meteorology parameters. The forecasting models were developed to yield 1 to 7 day ahead of forecasting for river flow. The performance measurements used to evaluate the forecasting model which are root mean square error (RMSE) and coefficient of determination (R^2).

6.1. Comparison between SVM and SVM-PSO Models

To compare the forecasting performances between hybrid model and non-hybrid models, comparison between two forecasting models without additional meteorology parameters which are SVM1 and SVM-PSO1 is first being focused. The comparison of RMSE and R^2 between SVM1 and SVM-PSO1 is presented in figure 2. As shown in figure 2, RMSE for both SVM1 and SVM-PSO1 increased with increasing forecasting lead day. However, it is clear that SVM-PSO1 yields lower RMSE as compared to SVM1 at 1 to 7 day river flow forecasting. Besides that, R^2 for both models decreased with increasing lead day of forecasting. Figure 2 shows that R^2 values for SVM-PSO1 are relatively higher as compared to SVM1. Therefore, as the lead day of forecasting increased, the correlation between forecasted and actual river flow decreased. Also, SVM-PSO1 yields higher correlation as compared to SVM1 for 1 to 7 day ahead of river flow forecasting.In the scenario where the input to the model are only historical rainfall and antecedent river flow data, hybrid SVM-PSO model yielded better performances as compared to non-hybrid SVM model in river flow forecasting.

Similar results are also obtained by the comparison between two forecasting models with additional meteorology parameters which are SVM2 and SVM-PSO2 as presented in figure 3. It is found that SVM-PSO2 has better performances as compared to SVM2 in river flow forecasting. SVM-PSO2 yielded lower RMSE as compared to SVM2 and provides higher R^2 value which indicated better performances and high accuracy in forecasting the river flow.

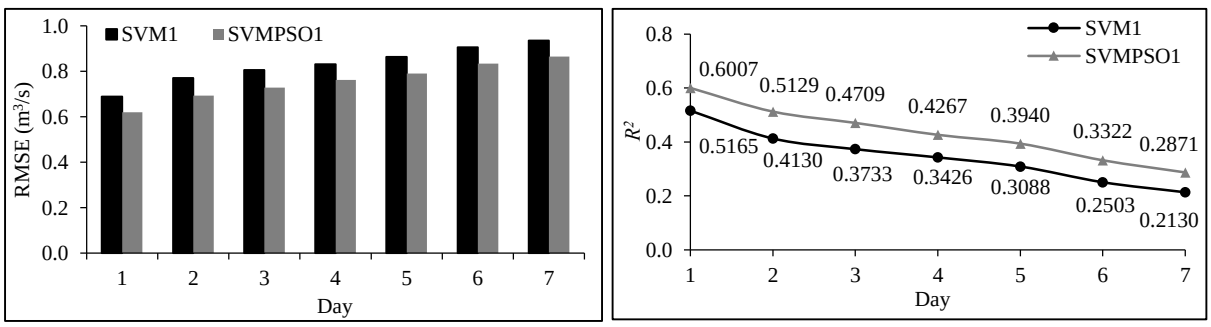

Figure 2. Comparison of RMSE and R^2 between SVM1 and SVM-PSO1

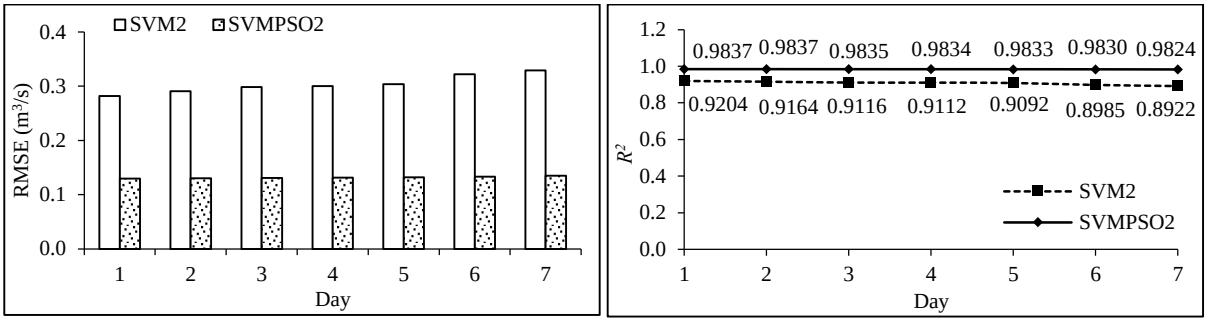

Figure 3. Comparison of RMSE and R^2 between SVM2 and SVM-PSO2

6.2. Improvement due to Additional Meteorology Parameters

To highlight the effect of additional meteorology parameters to river flow forecasting, comparison between SVM1 and SVM2 is then being focused. As shown in figure 4, RMSE values for both model SVM1 and SVM2 increased with increasing lead day forecasting at 1 to 7 day ahead of river flow forecasting. However, it is clear that SVM2 has lower RMSE as compared to SVM1. Besides, R^2 values for both models decreased with increasing lead day forecast. R^2 values for SVM2 are relatively higher than SVM1. The results show that SVM2 forecasted river flow more accurate as compared to SVM1.

Figure 5 illustrates RMSE and R^2 values for SVM-PSO1 and SVM-PSO2 in river flow forecasting at 1 to 7 day ahead of forecasting. It is found that hybrid model with additional meteorology parameters, SVM-PSO2, yielded lower RMSE value as compared to model without additional meteorology parameters, SVM-PSO1. SVM-PSO2 also produced R^2 higher than SVM-PSO1. Thus, SVM-PSO2 yielded higher performance in river flow forecasting at 1 to 7 day ahead as compared to SVM-PSO1.

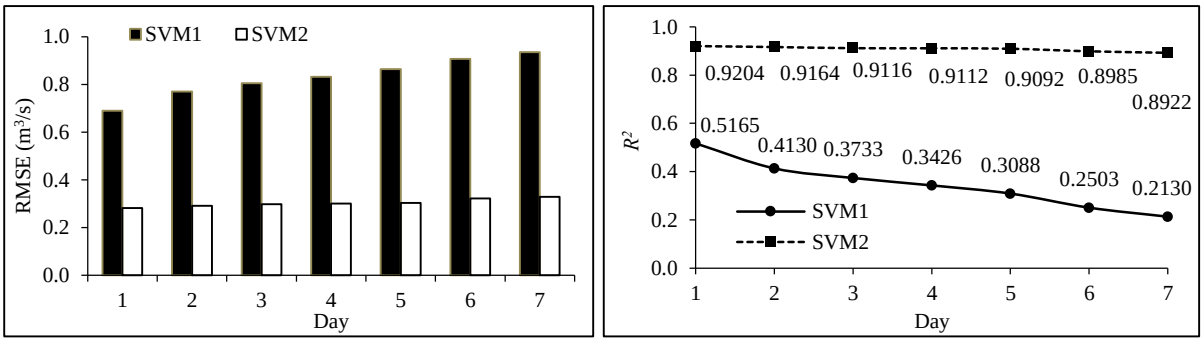

Figure 4. Comparison of RMSE and R^2 for SVM1 and SVM2

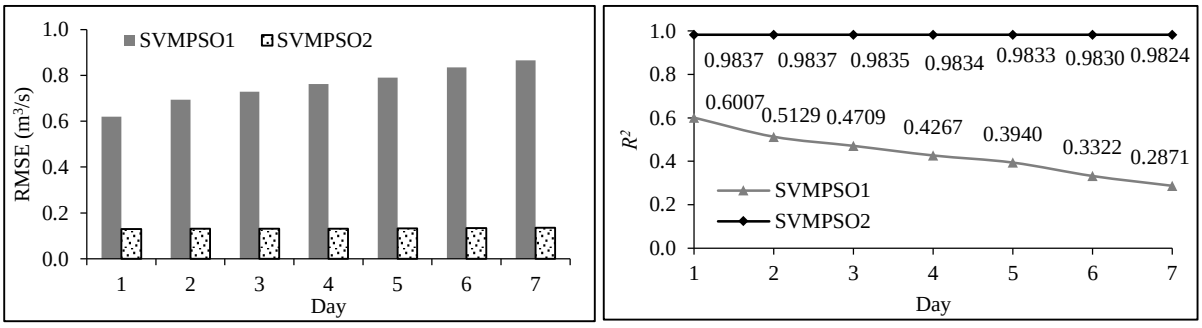

Figure 5. Comparison of RMSE and R^2 for SVM-PSO1 and SVM-PSO2

7. Conclusion

Total of four SVM based models were developed in this study, in order to forecast daily river flow at upstream of Bertam catchment. SVM and hybrid SVM-PSO models are developed based on two

different scenarios or input design which are; (i) only historical rainfall and antecedent river flow as input, (ii) historical rainfall, antecedent river flow and additional meteorology parameters used as input. Testing on the forecasting performance and accuracy of each model are conducted.

Hybrid models, SVM-PSO performed better than non-hybrid models, SVM. SVM-PSO models produced lower RMSE and higher R^2 as compared to SVM models. Besides that, it can be proved that the use of hybrid models, instead of non-hybrid models has effectively improved the forecasting performance in river flow. The models that incorporate additional meteorology parameters such as temperature, evaporation, relative humidity and wind speed as well as rainfall have effectively improve the long lead day forecasting. For non-hybrid model, SVM2 yields the highest performance and produced the most accurate forecasting at 1 to 7 day ahead river flow forecasting. For the hybrid models, SVM-PSO2 produced the highest performance and forecast the river flow most accurately. Thus, additional meteorology parameters including rainfall have influenced the performances and accuracy of river flow forecasting at 1 to 7 day ahead. In short, the hybrid SVM-PSO model with additional meteorology parameters including rainfall data is the most reliable and robust model in forecasting of daily river flow.

References

[1] Ling H, Xu H and Fu J 2013 Changes in Intra-Annual Runoff and its Response to Climate Change and Human Activities in the Headstream Areas of the Tarim River Basin, China *Quaternary International* 1

[2] Nazif S, Karamouz M and Zahmatkesh Z 2012 Climate Change Impacts on Runoff Evaluation: A Case Study in *World Environmental and Water Resources Congress 2012* United States

[3] Silberstein R, Aryal S, Durrant J, Pearcey M, Braccia M, Charles S, Boniecka L, Hodgson G, Bari M, Viney N and McFarlane D 2012 Climate Change and Runoff in South-Western Australia *J. of Hydrology* 441

[4] Sudheer C, Anand N, Panigrahi B and Mathur S 2013 Streamflow Forecasting by SVM with Quantum Behaved Particle Swarm Optimization *Neurocomputing* **101** 18.

[5] Guo J, Zhou J, Qin H, Zou Q and Li Q 2011 Monthly streamflow forecasting based on improved support vector machine *Expert System with Applications* 13073

[6] He Z, Wen X, Liu H and Du J 2014 A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region *J. of Hydrology* 379

[7] Chen S M, Wang Y M and Tsou I 2013 Using Artificial Neural Network Approach for Modeling Rainfall-Runoff due to Typhoon *J. of Earth System Science* **122** 399

[8] Asadi S, Shahrabi J, Abbaszadeh P and Tabanmehr S 2013 A New Hybrid Artificial Neural Networks for Rainfall-Runoff Process Modeling *Neurocomputing* **121** 470

[9] Kalteh A M 2008 Rainfall-Runoff Modelling Using Artificial Neural Networks (ANNs): Modelling and Understanding *Caspian Journal of Environmental Sciences* **6** 53

[10] Liu F, Zhou J-Z, Qiu F-P, Yang J-J and Liu L 2006 Nonlinear Hydrological Time Series Forecasting Based on the Relevance Vector Regression *Neural Information Processing* 880

[11] Bray M and Han D 2004 Identification of Support Vector Machine for Runoff Modelling *J. of Hydroinformatics* 265

[12] Wang W-c, Xu D-m, Chau K-w and Chen S 2013 Improved Annual Rainfall-Runoff Forecasting Using PSO-SVM Model Based on EEMD *J. of Hydroinfomatics* 1377

[13] Lin G-F, Chen G-P, Huang P-Y and Chou Y-C 2009 Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods *J. of Hydrology* 17

[14] Lin J-Y, Cheng C-T and Chau K.-W 2006 Using support vector machines for long-term discharge prediction *Hydrological Sciences Journal* 599

[15] Hu C-h, Wu Z-n, Wang J-j and Lina-Liu 2011 Application of the Support Vector Machine on Precipitation-Runoff Modeling in Fenhe River in *Water Resource and Environmental Protection (ISWREP), 2011 Intenational Symposium on*, Xi'an.

[16] Misra D, Oommen T, Agarwal A, Mishra S K and Thompson A M 2009 Application and Analysis of Support Vector Machine based Simulation for Runoff and Sediment Yield *Biosystem Engineeering* 527

[17] Xu J, Wei J and Liu Y 2010 Modeling Daily Runoff in a Large-scale Basin based on Support Vector Machines in *Computer and Communication Technologies in Agriculture Engineering (CCTAE), 2010 International Conference On*, Chengdu

[18] Behzad M, Asghari K, Eazi M and Palhang M 2009 Generalization Performance of Support Vector Machine and Neural Networks in Runoff Modeling *Expert System with Applications* 7624

[19] Kalra A, Ahmad S and Nayak A 2013 Increasing Streamflow Forecast Lead Time for Snowmelt-driven Catchment based on Large-scale Climate Patterns *Advance in Water Resources* 150

[20] Zakaria Z A and Shabri A 2012 Streamflow Forecasting at Ungaged Sites Using Support Vector Machine *Applied Mathematical Science* 3003

[21] Nieto P G, Garcia-Gonzalo E, Fernandez J A and Muniz C D 2014 Hybrid PSO-SVM-based method for long term forecasting of turbidity in the Nalon river basin: A case study in Nothern Spain *Ecology Engineering* 192

[22] Raghavendra N S and Deka P C 2014 Support Vector Machine Application in the Field Hydrology: Areview *Applied Soft Computing* 372

[23] Vapnik V N 1998 Statistical learning theory (New York: Wiley)

[24] Vapnik V N 1995 The nature of statistical learning theory (New Yok,USA: Springer)

[25] Cristianini N and Shawe-Taylor J 2000 An Introduction to Support Vector Machines and Other Kernel-based Learning Methods (United Kingdom: Cambridge University Press)

[26] Sudheer C, Maheswaran R, Panigrahi B K and Mathur S A 2013 Hybrid SVM-PSO Model for Forecasting Monthly Streamflow *Neural Computing and Application* **24** 1381

[27] Rubio G, Pomares H, Rojas I and Herrera L J 2011 A heuristic method for parameter selection in LS-SVM: Application to time series prediction *International journal of forecasting* 725

[28] Engelbrecht A P 2007 Computational Intelligence An Introduction (England: Wiley)

[29] Negnevitsky M 2002 Artificial Intelligence A guide to Intelligent Systems (England: Addison-Wesley)