Prediction of Stream Flow in Humid Tropical Rivers by Support Vector Machines

Mohammed Seyam[1,], Faridah Othman[2] and Ahmed El-Shafie[2]*

[1*]*Civil Engineering Programme., University College of Technology Sarawak, Sibu, Sarawak. (a.seyam@ucts.edu.my)*

[2]*Civil Engineering Dept., Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia*

Abstract. Stream flow (SF) prediction is considered as a very complex due to the hydrological systems of surface water are complex and dynamic. The reliable prediction of stream flow (SF) can be performed by either conceptual or data-driven based models. In the modelling of hydrological processes, the support vector machine (SVM) is a novel, data-driven approach. Hence, six SVM-based models were generated in this study to predict real time hourly SF in the Selangor River Basin from the water level and rainfall of upstream stations. These models composed of six different combinations of input variables and were trained and tested under hourly records of SF, rainfall, and water level over one year (2011). Among the SVM-based models, SVM-M6, which has nine input variables, was the most effective. Under the training and testing data sets, its correlation coefficient and mean absolute error values were 0.992, 0.953, 0.061 and 0.253 respectively.

Keywords: stream flow, data-driven based models, support vector machine, prediction, hydrological modelling.

1 Introduction

To predict SF, various models have been established. These models can be classified into two main types: knowledge- driven and data-driven. Each type has specific advantages and disadvantages based on data availability and modelling condition [1, 2]. Knowledge-driven models are also known as physical or conceptual models. They are designed to simulate interior sub processes in prototypes, as well as physical mechanisms that dictate the natural process. These models use a mathematical structure that depends on basin features, such as the specific characteristics of rainfall (intensity and duration), the basin (area, shape, slope and land use, vegetation cover, and soil nature), and climate (temperature, humidity, and wind speed) to model and predict SF [3-5]. However, these models are too complex and demanding. In some cases, conceptual models cannot predict SF accurately and reliably given the lack of required data, especially in developing countries [6], furthermore, the physical process is complicated by the gathering of data on multiple model variables that vary spatially and temporally [7-10].

Data-driven models include those developed using artificial intelligence (AI) techniques, such as artificial neural networks, genetic algorithms, support vector machine (SVM), and fuzzy rule-based systems. These models are adequate alternatives in many hydrological applications, especially when data are inadequate to generate conceptual models [11-14].

This study mainly aims to develop SVM-based models to predict hourly SF of downstream area from the water level and rainfall records of upstream stations in the river basins of humid tropical regions. These models are generated based on hourly records of SF, rainfall, and water level throughout one year (2011). The performance of the models are assessed based on two criteria, namely, correlation coefficient (R) and mean absolute error (MAE).

2 Methodology

In developing SVM-based models for SF prediction, we primarily consider data collection and analyses, followed by the selection of adequate input and output variables for the model. In small basins, these variables depend completely on the estimated lag time between the upstream and downstream stations. Thereafter, we determine model structure. Finally, we assess the developed models according to the evaluation criteria to obtain the model that best predicts hourly SF.

2.1 Case study

In this study, we investigate the Selangor River Basin, which is one of the main rivers in Malaysia. It is located in the Selangor state and has an approximate area of 1960 km^2 [15]. From northeast to southwest, the Selangor River is approximately 110 km long [6, 16, 17]. Moreover, the Selangor River Basin provides approximately 50% of the water consumed in Selangor and Kuala Lumpur [18, 19]. Figure 1 presents the location map of the Selangor River Basin in Peninsular Malaysia, as well as its topography maps.

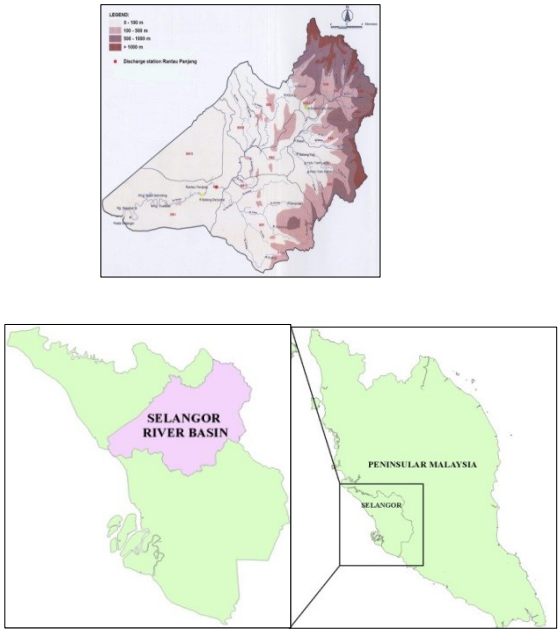

Fig. 1 Location and topography map of the Selangor River Basin [20].

2.2 Data collection and analyses

The SF data of the downstream station were obtained from the Rantau Panjang gauging station, which is located in the downstream of the Selangor River. Before this station, all of the major tributaries of this river converge. Thus, the SF at the Rantau Panjang Station is the best indicator of the stream flow at the study area. Water level and rainfall data were obtained from four upstream stations. The study stations were selected based on data availability and modelling requirement. Moreover, the stations that gauge rainfall and water level are very close to one another. Figure 2 displays the location of the hydrological stations and the flow paths among them in the Selangor River Basin.

Table 1. Hydrological stations and the statistical characteristics of the data used.

Station	Function	Mean	Min.	Max.	Std. Dev.
Rantau	SF (m^3/s)	60.35	23.94	294.64	39.0
Ulu Yam	RF (mm/h)	32.24	30.56	35.49	0.49
Batang Kali	RF (mm/h)	32.42	27.03	34.71	0.78
Kerling	RF (mm/h)	44.18	43.93	45.61	0.12
Ampang Pecah	RF (mm/h)	50.16	49.61	50.89	0.15
Ulu Yam	WL (m)	0.16	0.00	19.33	0.73
Batang Kali	WL (m)	0.24	0.00	22.67	0.91
Kerling	WL (m)	0.25	0.00	25.33	1.06
Ampang Pecah	WL (m)	0.24	0.00	28.00	1.08

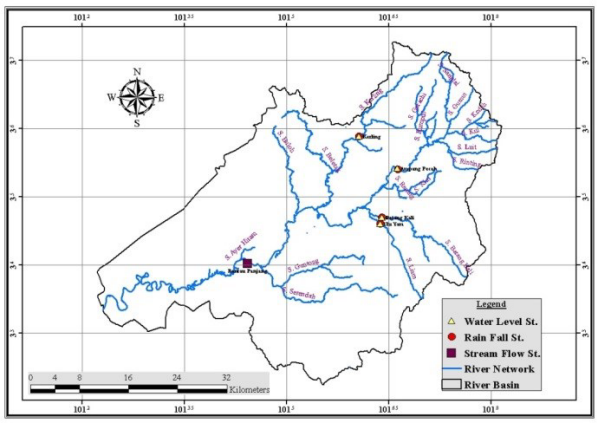

Fig. 2 Location of the hydrological stations in the Selangor River Basin.

2.3 Models development

2.3.1 Determination of model variables

In the development of AI-based models, determining the adequate input and output variables is a key issue. In models of SF prediction, model variables are commonly selected based on a priori knowledge of river basin hydrology, which provides initial indications of potential inputs and outputs [21]. The SF in tropical rivers can be characterized as the function of several influential variables, including rainfall, water level, and the physical characteristics of the river [22]. This study mainly aims to predict hourly SF of downstream area from the water level and rainfall records of upstream stations. Thus, we use the hourly

records of water level and rainfall at the upstream stations as input variables and those of SF data in the downstream station as the output variable. Eq. 2 describe the relationship between SF and the influential variables:

$$Sf_t = f(X_{(t)}) + e \quad (2)$$

where $Sf_{(t)}$ represents the SF; $X(t)$ is the input vector that includes the input variables (i.e., rainfall and/or water level); and e is the random error.

We consider three scenarios in selecting the input and output variables of the models. First, we apply the rainfall data of upstream stations as input variables. Second, we regard the water level data of these stations as inputs. Third, we utilize both water level and rainfall data from these stations as inputs.

In these three scenarios, we apply two input vectors. In the first, we use the single antecedent record of upstream stations. In the second, we obtain the average of these antecedent records. Given six input vectors, the single antecedent record of SF in the downstream station is considered another input variable that predicts the SF for a head period equal to the lag time between the upstream and downstream stations. The estimated lag time between these stations determines the final input variables for the six input vectors. Using these vectors, we generated six SVM-based models to predict hourly SF.

2.3.2 Model description

SVM is a new learning system that has been developed based on the statistical learning theory aiming at minimizing the generalized model error rather than just minimizing the training error, which consequently increases SVM generalization ability [23, 24].

SVM is a comparatively new AI modelling technique based on statistical learning theory introduced by Vapnik in the 1970s. SVM has been developed as a classification tool and it was applied successfully in a wide range of classification and clustering applications in. Recently, SVM have been successfully extended to apply in regression and prediction applications [11, 25, 26].

Figure 3 presents the Schematic diagram of SVM, where the K(xi,x) is the output of the ith hidden node for input vector x, it is a mapping of the input x and the support vector xi by selecting the kernel function (Chen & Yu, 2007).

SVM has been applied in the time-series prediction of river flow by Samsudin, Saad [6]; in SF prediction under multiple time scales by Asefa, Kemblowski [27]; in the real-time forecasting of flood stage by Yu, Chen [25]; in flood forecasting by [28]; in long-term discharge prediction by Lin, Cheng [29]; in the long-range forecast of SF by [30]; and in the monthly forecasting of SF by Guo, Zhou [31], Noori, Karbassi [32], Shabri and Suhartono [33], and Ch, Anand [34].

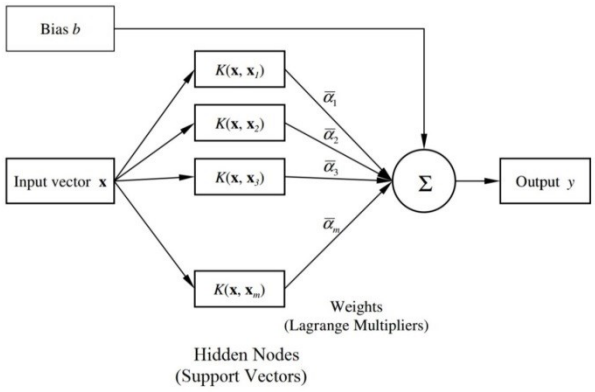

Fig. 3: Schematic diagram of SVM architecture

2.3.3 Model training

Once the structure of the SVM-based model has been determined, we set the conditions that halt the training processes. These conditions should be set prior to model training. Training is controlled by conditions such as the maximum number of iterations, maximum training time, the target performance that specifies the tolerance between the observed and predicted SF, and the minimum learning rate.

To generate the models, we determined approximately 8,753 patterns of hourly SF, water level, and rainfall records throughout a single year (2011). Table 1 lists the basic statistical characteristics of the hourly records obtained from the stations, such as minimum, maximum, mean, standard deviation, and skewness. The modelling data were divided into two data sets: 75% for training with 6,580 patterns and 25% for model testing with 2,193 patterns. The training data set is used to train the models, and the testing data set assesses the performance of the SVM-based models [35].

In this study, we are applying the SVM as modelling tool. The training process was performed internally using close-source programing which is available in AI toolbox in Statistica software. Hence, the best training algorithm was selected by the software using built-in optimization technique, where

Levenberg-Marquardt technique is adopted as the training algorithm because it provides the best performance over other algorithms. The main attention of the this paper is maximize the correlation and minimize the error, regardless the details of training algorithms and techniques.

2.3.4 *Performance evaluation criteria*

The performance of the models was assessed based on two criteria: R and MAE. R is a statistical technique that indicates the strength and direction of a linear relationship between two variables [36, 37]. In this study, R was used to validate the agreement between the observed and predicted hourly SFs. R^2 describes the variance between two variables as determined by the linear fit. R can be calculated under different modes, but the most popular one is the Pearson R. This value is computed by dividing the covariance of the two variables by the product of their standard deviations, as described in the following equation:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

where n is the number of data pairs and x and y are the variables.

In a perfectly increasing linear relationship, R is +1. By contrast R is −1 in a perfectly decreasing linear relationship. The R values between +1 and −1 indicate the strength of the linear relationship between the variables. R = 0 signifies that the variables are not linearly related.

MAE evaluates the residual or the differences between the observed and predicted SF. Theoretically, its minimum value should be zero to indicate the perfect fit of the model. However, this value is difficult to obtain. Moreover, MAE has no maximum value and is calculated using the following equation:

$$MAE = \frac{\sum_{i=1}^{n}\left|X_{o,i} - X_{m,i}\right|}{n} \quad (4)$$

where X_m represents the data predicted by a model and X_o is the observed data.

3 Results and discussion

To predict hourly SF in the Rantau Panjang Station, we selected six AI-based models under different combinations of input variables. Table 2 presents the model structures. The six models were trained and developed by SVM to predict hourly SF. The performances of the models were assessed based on the training and testing data sets, as well as the overall performance of the data sets. The best fit model to predict hourly SF is thus determined according to the performance of the testing data sets.

SVM-M6 model displays the highest R values (0.992 and 0.953) and the lowest MAE (0.061 and 0.253) in both the training and testing data sets, respectively. Figure 4 shows the correlation between the observed and predicted hourly SF in the SVM-M6 model given training and testing data sets. The observed and predicted hourly stream flow of the training and testing data sets, seem to be in good accord with R^2 0.986 and 0.909 respectively. Figure 5 compares the observed and predicted hourly SF in SVM-M6 for the period of September 2013. These flows are highly consistent.

Table 2. Input and output variables of the AI models.

Model	Inputs	Output	No. of input Variables
M1	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Sf_{(t)}$	$Sf_{(t+17)}$	5
M2	$Rf_{u(t)}$, $Rf_{b(t)}$, $Rf_{k(t)}$, $Rf_{a(t)}$, $Sf_{(t)}$	$Sf_{(t+17)}$	5
M3	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Sf_{(t)}$	$Sf_{(t+12)}$	5
M4	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Sf_{(t)}$	$Sf_{(t+12)}$	5
M5	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-5)}$, $Rf_{b(t-5)}$, $Rf_{k(t-5)}$, $Rf_{a(t-5)}$, $Sf_{(t)}$	$Sf_{(t+12)}$	9
M6	$Wl_{u(t)}$, $Wl_{b(t)}$, $Wl_{k(t)}$, $Wl_{a(t)}$, $Rf_{u(t-5)}$, $Rf_{b(t-5)}$, $Rf_{k(t-5)}$, $Rf_{a(t-5)}$, $Sf_{(t)}$	$Sf_{(t+12)}$	9

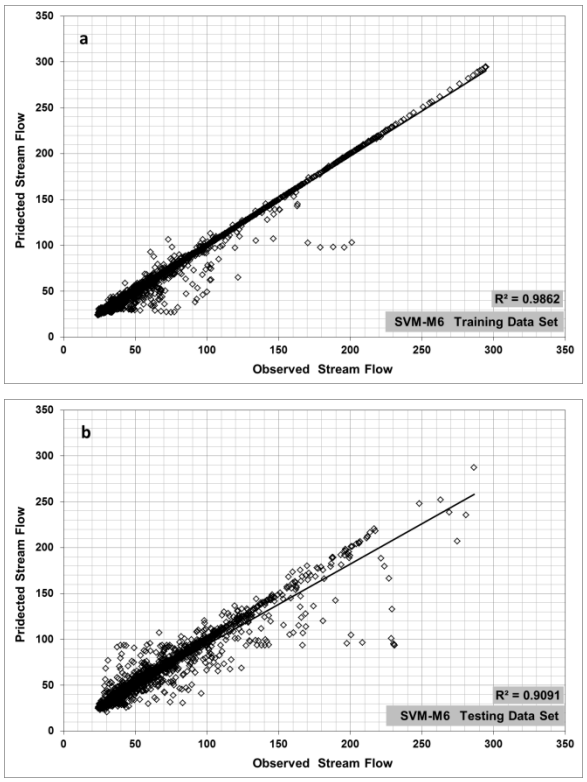

Fig. 4 Correlation between the observed and predicted hourly stream flow in the SVM-M6 model: (a) training data set and (b) testing data set

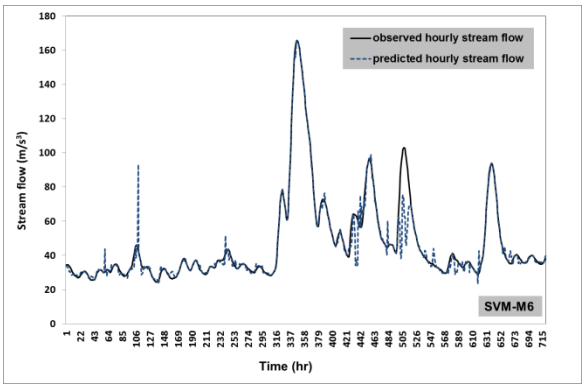

Fig. 5 Comparison between the observed and predicted hourly stream flows in the SVM-M6 model

4 Conclusions

In this study, the ability of SVM to predict hourly SF in downstream area from the upstream water level and rainfall records in a humid tropical area was explored. In the process, we developed six SVM-based models with different combinations of input variables. The hourly records of SF, rainfall, and water level throughout one year (2011) were utilized to train and test the SVM models. The hourly water level and rainfall data in the upstream stations were considered input variables, and the hourly SF data in the downstream station were regarded as the output variable. The performance levels of the models were assessed based on two evaluation criteria, namely, *R* and MAE. The best fit model to predict hourly stream flow was determined based on the performance of the testing data sets. Among the SVM-based models, SVM-M6 model displays the highest *R* values (0.992 and 0.953) and the lowest MAE (0.061 and 0.253) in both the training and testing data sets, respectively. The ability of the three techniques of SVM SVM for SF prediction in the downstream area from the upstream WL and RF records in the Selangor River basin was explored and successfully achieved with high performance throughout two modelling phases. Although SVM performed well in real-time SF prediction, higher R in the Selangor River basin can be investigated by employing other AI techniques, such as FRBSs and GAs.

References

1. Aqil, M., et al., *Neural Networks for Real Time Catchment Flow Modeling and Prediction.* Water Resources Management, 2007. **21**(10): p. 1781-1796.
2. Kentel, E., *Estimation of river flow by artificial neural networks and identification of input vectors susceptible to producing unreliable flow estimates.* Journal of Hydrology, 2009. **375**(3–4): p. 481-488.
3. Jain, A. and A.M. Kumar, *Hybrid neural network models for hydrologic time series forecasting.* Applied Soft Computing, 2007. **7**(2): p. 585-592.
4. Nilsson, P., C.B. Uvo, and R. Berndtsson, *Monthly runoff simulation: Comparing and combining conceptual and neural network models.* Journal of Hydrology, 2006. **321**(1–4): p. 344-363.
5. Kneis, D., *A lightweight framework for rapid development of object-based hydrological model engines.* Environmental Modelling & Software, 2015. **68**(0): p. 110-121.
6. Samsudin, R., P. Saad, and A. Shabri, *River flow time series using least squares support vector machines.* Hydrol. Earth Syst. Sci., 2011. **15**(6): p. 1835-1852.
7. Firat, M. and M.E. Turan, *Monthly river flow forecasting by an adaptive neuro-fuzzy inference system.* Water and Environment Journal, 2010. **24**(2): p. 116-125.
8. Akhtar, M.K., et al., *River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin.* Hydrol. Earth Syst. Sci., 2009. **13**(9): p. 1607-1618.
9. Nativi, S., et al., *Big Data challenges in building the Global Earth Observation System of Systems.* Environmental Modelling & Software, 2015. **68**(0): p. 1-26.
10. Seyam, M. and Y. Mogheir, *Application of artificial neural networks model as analytical tool for groundwater salinity.* Journal of Environmental Protection, 2011. **2**(01): p. 56.
11. Solomatine, D., L.M. See, and R.J. Abrahart, *Data-Driven Modelling: Concepts, Approaches and Experiences*, in *Practical Hydroinformatics*, R. Abrahart, L. See, and D. Solomatine, Editors. 2008, Springer Berlin Heidelberg. p. 17-30.
12. Kisi, O., et al., *Intermittent Streamflow Forecasting by Using Several Data Driven Techniques.* Water Resources Management, 2012. **26**(2): p. 457-474.
13. Daniel, E.B., et al., *Watershed Modeling and its Applications: A State-of-the-Art Review.* The Open Hydrology Journal, 2011. **5**: p. 26-50.
14. Kanevski, M., et al., *Environmental data mining and modeling based on machine learning algorithms and geostatistics.* Environmental Modelling & Software, 2004. **19**(9): p. 845-855.
15. Seyam, M. and F. Othman, *Long-term variation analysis of a tropical river's annual streamflow regime over a 50-year period.* Theoretical and Applied Climatology, 2015. **121**(1): p. 71-85.
16. Lee, C.M., *Master Plan Study on Flood Mitigation and River Management for Sg. Selangor River Basin.*, 2002, Drainage and Irrigation Department (DID) Malaysia.
17. Hassan, A.J., A.A. Ghani, and R. Abdullah, *Development Of Flood Risk Map Using GIS For Sg. Selangor Basin*, 2004, National Hydraulic Research Institute of Malaysia: Malaysia.
18. Subramaniam, V., *Managing Water Supply In Selangor And Kuala Lumpur*, in *BULETIN INGENIEUR*2004, THE BOARD OF ENGINEERS MALAYSIA: 50580 Kuala Lumpur, Malaysia. p. 12-20.
19. Seyam, M. and F. Othman, *The Influence of Accurate Lag Time Estimation on the Performance of Stream Flow Data-driven Based Models.* Water Resources Management, 2014. **28**(9): p. 2583-2597.
20. Seyam, M. and F. Othman, *Long-term variation analysis of a tropical river's annual streamflow regime over a 50-year period.* Theoretical and Applied Climatology, 2014.
21. Maier, H.R. and G.C. Dandy, *Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications.* Environmental Modelling & Software, 2000. **15**(1): p. 101-124.
22. Firat, M., *Artificial Intelligence Techniques for river flow forecasting in the Seyhan River Catchment, Turkey.* Hydrol. Earth Syst. Sci. Discuss., 2007. **4**(3): p. 1369-1406.

23. Behzad, M., K. Asghari, and E.A. Coppola Jr, *Comparative Study of SVMs and ANNs in Aquifer Water Level Prediction.* Journal of Computing in Civil Engineering, 2010. **24**: p. 408.
24. Asefa, T., et al., *Multi-time scale stream flow predictions: The support vector machines approach.* Journal of Hydrology, 2006. **318**(1-4): p. 7-16.
25. Yu, P.-S., S.-T. Chen, and I.F. Chang, *Support vector regression for real-time flood stage forecasting.* Journal of Hydrology, 2006. **328**(3–4): p. 704-716.
26. Wu, C.L., K.W. Chau, and Y.S. Li, *River stage prediction based on a distributed support vector regression.* Journal of Hydrology, 2008. **358**(1–2): p. 96-111.
27. Asefa, T., et al., *Multi-time scale stream flow predictions: The support vector machines approach.* Journal of Hydrology, 2006. **318**(1–4): p. 7-16.
28. Chen, S.-T. and P.-S. Yu, *Pruning of support vector networks on flood forecasting.* Journal of Hydrology, 2007. **347**(1–2): p. 67-78.
29. Lin, J.Y., C.T. Cheng, and K.W. Chau, *Using support vector machines for long-term discharge prediction.* Hydrological Sciences Journal, 2006. **51**(4): p. 599-612.
30. Basketfield, D. and N. She, *Long Range Forecast of Streamflow Using Support Vector Machine*, in *Impacts of Global Climate Change*. 2005. p. 1-9.
31. Guo, J., et al., *Monthly streamflow forecasting based on improved support vector machine model.* Expert Systems with Applications, 2011. **38**(10): p. 13073-13081.
32. Noori, R., et al., *Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction.* Journal of Hydrology, 2011. **401**(3–4): p. 177-189.
33. Shabri, A. and Suhartono, *Streamflow forecasting using least-squares support vector machines.* Hydrological Sciences Journal, 2012. **57**(7): p. 1275-1293.
34. Ch, S., et al., *Streamflow forecasting by SVM with quantum behaved particle swarm optimization.* Neurocomputing, 2013. **101**(0): p. 18-23.
35. Tiwari, M.K. and C. Chatterjee, *Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach.* Journal of Hydrology, 2010. **394**(3–4): p. 458-470.
36. Perugu, M., A. Singam, and C. Kamasani, *Multiple Linear Correlation Analysis of Daily Reference Evapotranspiration.* Water Resources Management, 2013. **27**(5): p. 1489-1500.
37. Seyam, M., F. Othman, and A. El-Shafie, *RBFNN Versus Empirical Models for Lag Time Prediction in Tropical Humid Rivers.* Water Resources Management, 2016.