International Journal of Civil Engineering and Technology (IJCIET)
Volume 10, Issue 01, January 2019, pp. 1523-1534, Article ID: IJCIET_10_01_140
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=10&IType=01
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

© **IAEME** Publication **Scopus** Indexed

RAINFALL-RUNOFF FORECASTING UTILIZING GENETIC PROGRAMMING TECHNIQUE

Ali N. Ahmed

Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia

Gasim Hayder

Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia

Raihana Aliya Binti Abdul Rahman

Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia

Abdoulhdi A. Borhana

Department of Mechanical Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia.
Department of Mechanical Engineering, Faculty of Engineering Science and Technology, Sebha University, Libya

ABSTRACT

This paper reports how the rainfall-runoff is forecasted utilizing Genetic Programming (GP) technique. It is a program that was inspired by biological processes such as mutation, crossover, and inversion in order to create a new generation. It is a program that will learn and improve with each analysis done. It uses a trial an error method in order to forecast rainfall-runoff. GP uses Root Mean Squared Error (RMSE) as an indication of how accurate the results of the forecast. The lower and closer the RMSE to zero, the more accurate the rainfall-runoff forecasted. The study consists of running the data on the software until the lowest RMSE is obtained. This research contains three models which use a different number of input variables to see whether it will give an impact on the rainfall-runoff forecasting. The results are compared and a bar chart is plotted.

Key words: Forecasting, Genetic Programming, Rainfall, Root Mean Squared Error, Runoff

Cite this Article: Ali N. Ahmed, Gasim Hayder, Raihana Aliya Binti Abdul Rahman and Abdoulhdi A. Borhana, Rainfall-Runoff Forecasting Utilizing Genetic Programming Technique, International Journal of Civil Engineering and Technology,10(01),2019,pp.1523–1534
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=10&IType=01

1. INTRODUCTION

When rain keeps on falling, the rainwater which hits the ground surface will infiltrate into the soil where it will exceed the infiltration capacity of the soil. When the rainwater exceeds the infiltration (seepage) capacity of the soil, the water will flow over the soil and this is known as the surface runoff. Water will flow and fill the surface and it will cause puddles, filled ditches, and other depression storages. The capacity of the infiltration usually depends on the types of the surface such as its texture, structure, profile and its soil moisture content. The process of rainfall turning into a runoff is a very complicated and non-linear progression, almost an abstract process because alternative factors such as evaporation may have important consequences in hydrology, thus it may affect the accuracy of the results.

As for hydrology, specialists in water resources have come up with data based modeling approaches since the rainfall-runoff relationship has an unorganized behavior thus these lead researchers to pay more attention to soft-computing tools. Factors such as physiographic, biotic characteristics and climatic of the basin sometimes induce either a linear, non-linear or highly complex nature among the rainfall and runoff parameters [1].

Genetic Programming (GP) is a computing modeling and technique, created based on biological evolution and it is more capable when compared to other techniques available currently such as Artificial Neural Network (ANNs) because GP allows the user to obtain more data and information on the performance of the system. ANNs are inspired by the process and how the human brains work. Even though it is a much simpler model of the human brain, [2] has described that ANNs can be expressed as a simple processing unit that consists of the parallel distributed processor and has an inclination for storing experiential knowledge. They also consist of artificial neurons which represent as the processing elements for information learning. These neurons are able to perform frivolous functions but collectively in the pattern of a network, they are capable of solving difficult problems [3] and have been used in application of artificial intelligence that has shown quite a promise in many areas that required simulation of complex relationships for the purpose of modeling and prediction [4] and [5].

On the other hand, Fuzzy Logic (FL) was programmed by Ronald and Lotfi [6]. FL is a computational method where it provides insight into the area where inaccuracy and vague knowledge often prevail thus a more precise reasoning and decision-making ability are able to be created by the program to counter the inaccuracies and vagueness. FL is not a tool that can give exact reasoning but instead just only an approximation. A study by Tokar and Johnson shows that the characteristics and climatic conditions of a catchment area are important in order to obtain decent predictions [7]. Other than that, daily rainfall data from a small catchment area can be used to develop a rainfall-runoff modeling for a much bigger catchment area, and this method was used by Rajurkar [8]. It is not restricted to only use daily rainfall to predict runoff. Other parameters such as evaporation, precipitation, and discharge data can also be used to develop runoff modeling, and this idea was used by Wilby [9].

Unlike ANNs, GP is a more advanced method and it is used in rainfall-runoff forecasting, rainfall estimation, streamflow forecasting, water quality modeling, and groundwater modeling. GP allows the user to obtain more data or insight into the relationship between the input and output variables. The process of GP to produce accurate solutions are by picking random parse trees and identify the fittest which is the one that able to solve the given problem and chose the better parse tree. Other than that, GP also uses functions that use arithmetic operators, mathematical functions, Boolean operators, logical operators, iterative functions or any user-defined function. GP can also provide assistance to the problems where total insight for difficult and complicated physical process is required.

This paper focused on forecasting the rainfall runoff in Sungai Sayong, Johor by using GP technique and to investigate whether a different number of input variables used will give an impact on the rainfall-runoff forecasting. The advantages of using GP as rainfall-runoff modeling are due to its capability to choose input variables that are only profitable to the model and the ability to neglect variables that are not important to the model. The error messages in GP are presented in the form of RMSE, Mean Square Error (MSE) and etcetera. Sungai Sayong is neither as high nor as steep as a mountain, which makes it ideal to be a test area for this research. In justification, by having this research study, an improved technique for rainfall-runoff prediction can be developed for the future benefits for similar studies.

2. METHODOLOGY

The data used in this research was collected in Bukit Besar, Kulai, Johor. It is located basically 29 km from Johor Bahru city and 8 km from Skudai. It is surrounded by oil palm and rubber estates. Since it is located on a hilly area, it has some slopes to its geographical structure. It is neither as high nor steep as a mountain which makes it ideal to be a test area for this research since GP is not very effective for a steep catchment area just because of GP unable to predict data accurately when the area is too steep. During the test, the observation point was fixed at one point so that the condition in the surrounding area is not interrupted. In this case study, GP is used to develop a significant relationship between the future runoff at the catchment outlet, and rainfall and runoff data available up to the current time t. mathematically the relationship may be expressed as the formula below:

$$Qt + \delta\Delta t \; = f\,(Rt, Rt - \Delta t \;\ldots\; Rt - \omega\Delta t, Qt, Qt - \Delta t \;\ldots\; Qt - \omega\Delta t)$$

where Q is the runoff (m^3/s), R is the rainfall intensity (mm/day), δ (with δ = 1, 2, …) refers to how far into the future the runoff prediction is desired, ω (with ω = 1, 2, …) implies how far back the recorded data in the time series are affecting the runoff prediction while Δt stands for time step discretization such as time interval.

As stated by other researches, the most common used stopping criteria is stopping after a certain number of runs through all of the training data and stop the running when the target is achieved [10], [11], [12]. When the variables are tested many times, the error will gradually reduce. This is because the system will eliminate the error while trying to come up with the fittest solution to the problems.

3. RESULTS AND DISCUSSIONS

3.1. Input Variable Selection using GP

The number of input variables used for Sungai Sayong is 4 followed by 3 inputs and 2 inputs variable. Since Sungai Sayong is located in Johor Bahru which is on the west coast of Malaysia. The monsoon that affects the west coast of the peninsula is the southwest monsoon. West coast monsoon only occurs between May and October, thus the input variable used is the rainfall amount and the stage of the river between May and October. The stage is defined as the level of the river.

Table 1 shows the parameters used for the runs. The parameter “Max tree depth” was fixed to 5 because the higher the amount of the tree depth, it will increase the complexity of the model but at the same time will increase the possibility to produce an even more accurate model.

Table 1 Value of GP Control Parameters

Parameter	Value
Population size	100
Number of generations	100
Tournament size	3
Max tree depth	5
Number of inputs	4
Max genes	2
Population size	100
Number of generations	100

Figure 1 shows how the input variables were chosen for the first model. For example, to forecast runoff for the month of May, the input variables are the rainfall and the stage data for March and April. Two months prior to the forecasted month are taken into consideration. In the second model, as shown in Figure 2, the input selection is quite different as the input variables for the second model are stage data for March and April and also the rainfall data for that forecasted month which is May. While in the third model, as shown in Figure 3, the input variables are the stage data for April and the rainfall data for May.

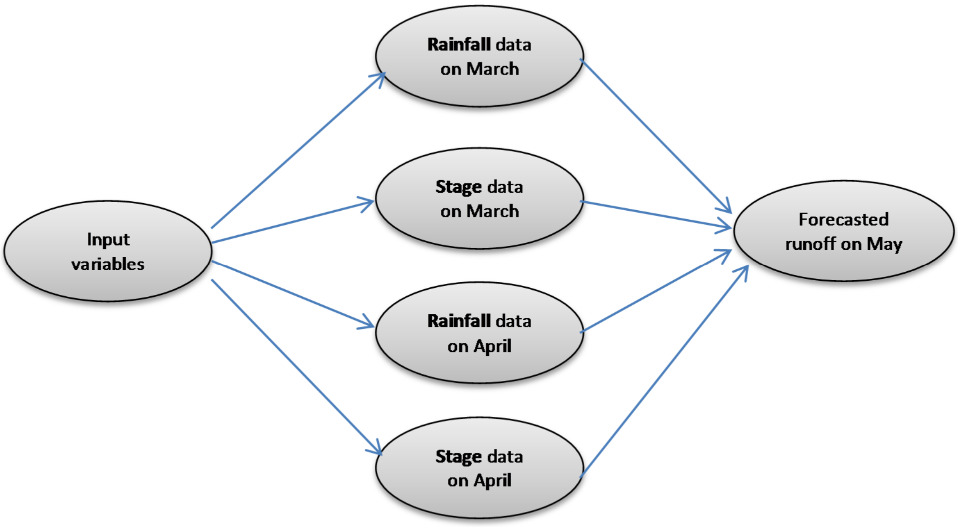


Figure 1 Input data for the GP based rainfall-runoff model (first model)

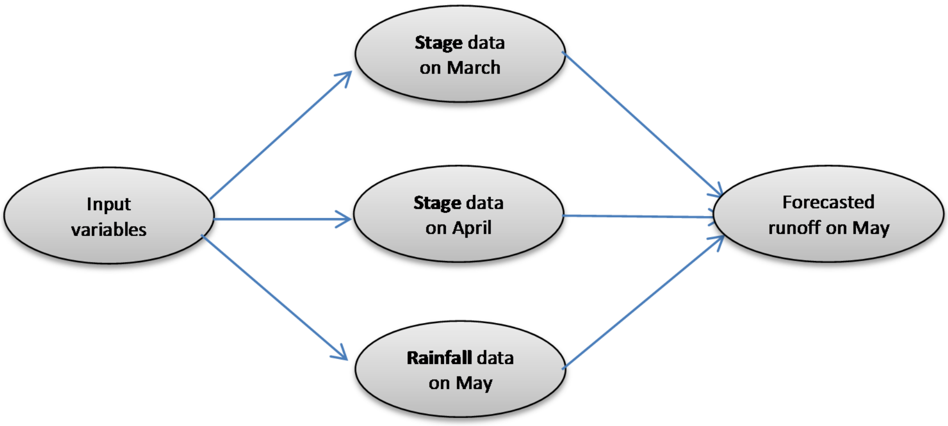


Figure 2 Input data for the GP based rainfall-runoff model (second model)

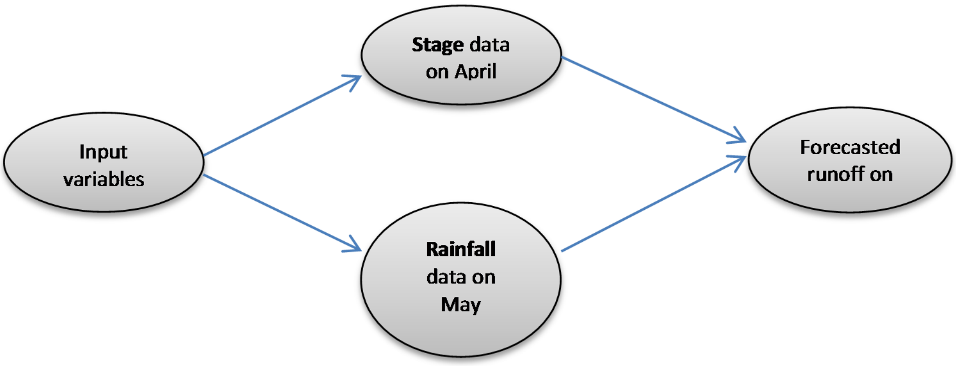


Figure 3 Input data for the GP based rainfall-runoff model (third model)

3.2. GP Runs

The inputs used for the 3 models are different from one another. The first model, for example, the first run is to test the output for the stage level (L) for the month of May. The inputs used are the rainfall (R) data for March, L for March, R for April and L for April. This is concluded that 2 months prior to the chosen output for the run is considered as the inputs for the rest of the runs on the first model.

For the second model, the inputs used for the output run for the month of May are L for March and April, and R for May. This model considers the rainfall data of the chosen output in the model. Lastly, the third model used the input of L for April and R for May as its inputs for the output run for the month of May. The amounts of input are decreasing throughout the models.

Based on the GP software, gpdemo2 is able to perform the runs for multigene symbolic regression by using the formula consists of 4 inputs and 1 output as defined below. The formula is adapted from Cherkassky [13]. Both the testing & training data is free of noise. The multigene regression runs on 310 data points generated by the non-linear function with 4 inputs:

$$y = exp\,(2x1sin\,(\pi\; x4)) + sin(x2x3)$$

The configuration file used in this run is gpdemo2_config.m and the raw data is in demo2data.mat. At the end of the run, the predictions of the evolved model will be plotted for the training data as well as for an independent test data set generated from the same function. Here, 2 genes are used (plus a bias term) so the form of the model will be:

$$ypred = c0 + c1 * tree1 + c2 * tree2$$

Where:

c_0 = bias

c_1 and c_2 are the weights

The bias and weights (i.e. regression coefficients) are automatically determined by a least squares procedure for each multigene individual.

Table 2, 3 and 4 shows that the grey areas are the lowest RMSE on all third models of the GP run. The lowest value observed is in the month of July in the third model. The lower the amount of the RMSE and the closer the RMSE to 0 indicate that the rainfall-runoff forecasted is more accurate compared to the other runs. While the more and closer the Response Variable (R^2) is to 1 is better. The inputs used in this research vary from each other as evidenced by the different amount of inputs used in every model. The first model used 4 input variables, the second model used 3 input variables and the third and last model used 2 input variables.

Table 2 Value of GP Control Parameters

Input Variable (4 inputs)	RMSE		Response variable (R^2)
	Training	Testing	
L_{may} (R_{mar}, L_{mar}, R_{apr}, L_{apr})	0.049596	0.087022	0.12537
L_{jun} (R_{apr}, L_{apr}, R_{may}, L_{may})	0.054942	0.044921	0.1987
L_{july} (R_{may}, L_{may}, R_{jun}, L_{jun})	0.054748	0.045602	0.20437
L_{aug} (R_{jun}, L_{jun}, R_{july}, L_{july})	0.054866	0.043168	0.20093
L_{sept} (R_{july}, L_{july}, R_{aug}, L_{aug})	0.054963	0.043389	0.19804
L_{oct} (R_{aug}, L_{aug}, R_{sept}, L_{sept})	0.054752	0.047143	0.20424

Table 3 Second model of the GP run in Sungai Sayong

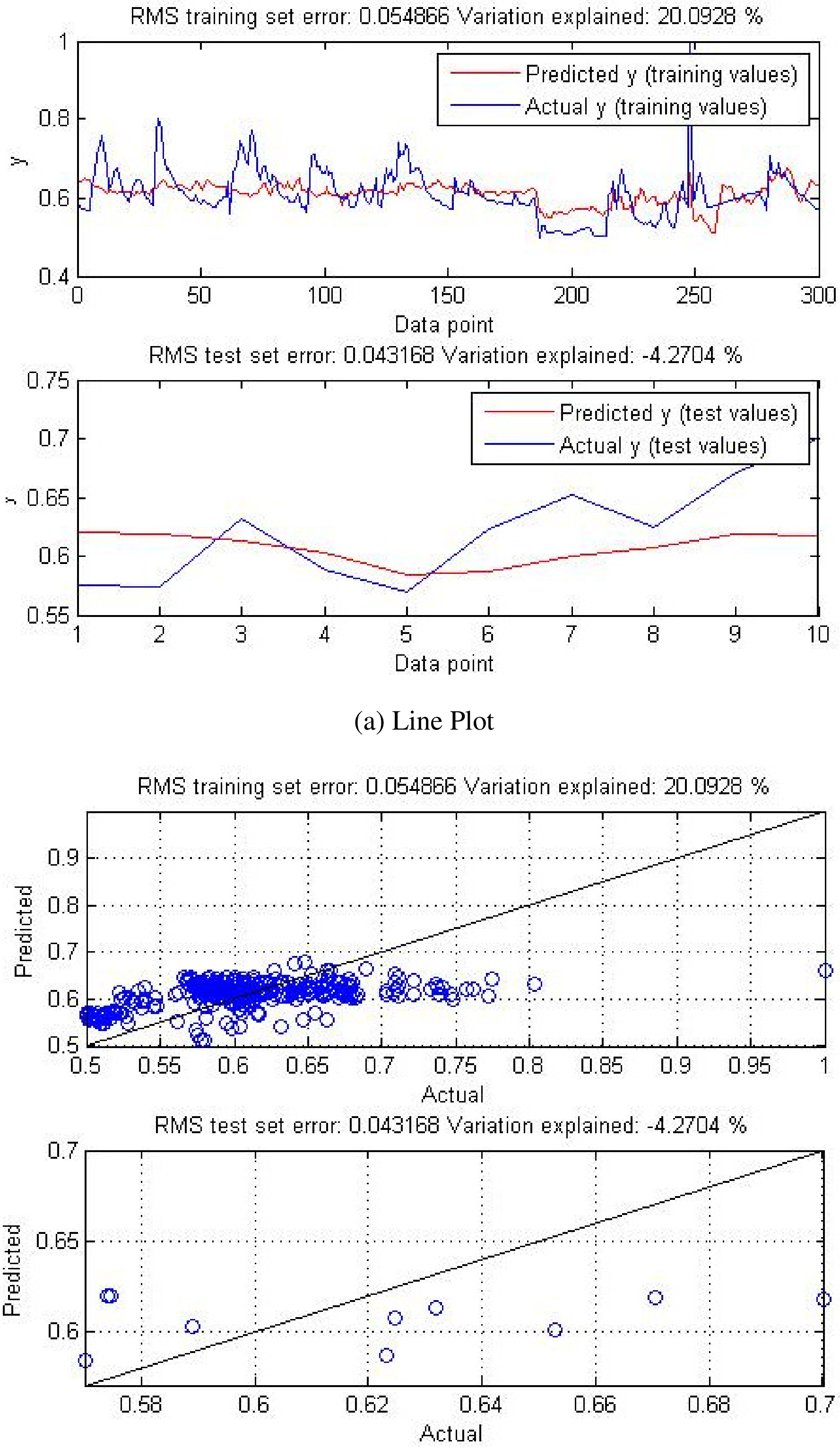
Input Variable (3 inputs)	RMSE		Response variable (R^2)
	Training	Testing	
L_{may} (L_{mar}, L_{apr}, R_{may})	0.054768	0.044261	0.20379
L_{jun} (L_{apr}, L_{may}, R_{jun})	0.054737	0.046861	0.20468
L_{july} (L_{apr}, L_{jun}, R_{july})	0.054753	0.047074	0.20421
L_{aug}	0.056271	0.054241	0.15949

(L_{jun}, L_{july}, R_{aug})			
L_{sept}	0.054824	0.046506	0.20215
(L_{july}, L_{aug}, R_{sept})			
L_{oct}	0.054962	0.042499	0.19811
(L_{aug}, L_{sept}, R_{oct})			

Table 4 Third model of the GP run in Sungai Sayong

Input Variable (2 inputs)	RMSE		Response variable (R^2)
	Training	Testing	
L_{may}	0.054837	0.045202	0.20177
(L_{apr}, R_{may})			
L_{jun}	0.054703	0.046237	0.20566
(L_{may}, R_{jun})			
L_{july}	0.054366	0.043511	0.21541
(L_{jun}, R_{july})			
L_{aug}	0.056139	0.055845	0.16342
(L_{july}, R_{aug})			
L_{sept}	0.054781	0.043642	0.20341
(L_{aug}, R_{sept})			
L_{oct}	0.055144	0.050446	0.1928
(L_{sept}, R_{oct})			

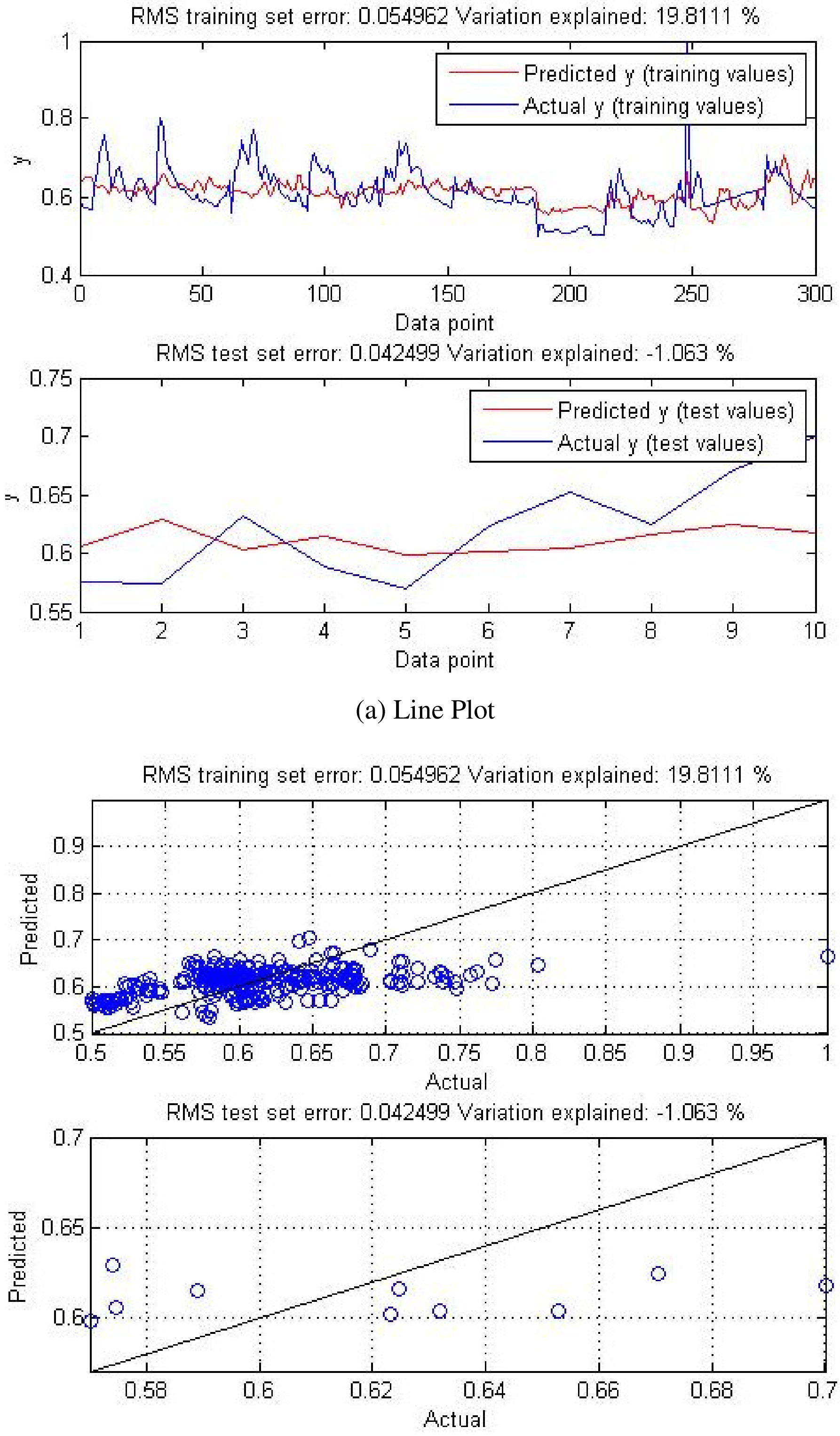
In the GP software manual, it shows that the required input variables are 4 inputs. However, based on the study and analysis conducted above, other multiple input variables produced verifiable results too. Thus this supports the assumption that GP also works if the variable inputs are varied and different.



(a) Line Plot

(b) Scatter plot

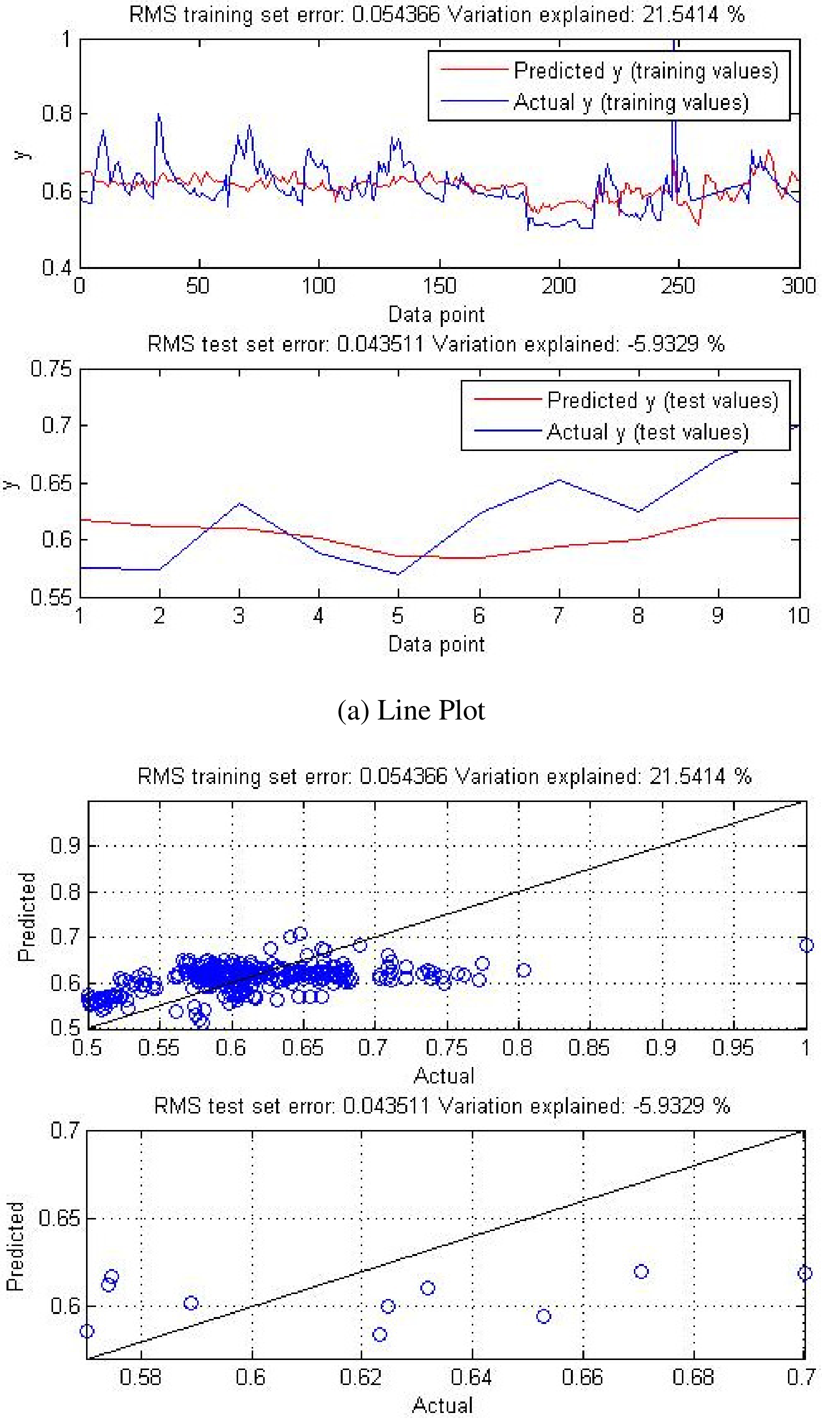
Figure 4 Predicted runoff versus actual runoff for training and test value for first model



(a) Line Plot

(b) Scatter plot

Figure 5 Predicted runoff versus actual runoff for training and test value for second model



(a) Line Plot

(b) Scatter plot

Figure 6 Predicted runoff versus actual runoff for training and test value for second model

Figure 4, 5 and 6 shows the graph plotted for the best RMSE and R^2 in both line plot and scatter plot for the training and test set data. The red line indicates the predicted runoff and the blue is the actual runoff on that month. The closer the red and blue lines tracked each other, it shows the higher is the accuracy of the forecasted runoff. For visualization of the proposed model performance, a demonstration of the actual versus the predicted runoff during the phase of training and testing is shown in the graphs.

3.3. RMSE Comparison

Figure 7 shows the second model has the lowest RMSE value compared to the first and the third model. This indicates that the second model is the most accurate model for rainfall-runoff forecasting. As the lowest the RMSE value the more accurate a model is. As a refresher, the RMSE is essentially the distance of a data point from the fitted line, measured along a vertical line, thus it is a measure of goodness of fit. The RMSE compares the observable variation in measurements of a typical point, wherein for a reasonable fit there should be little variation.

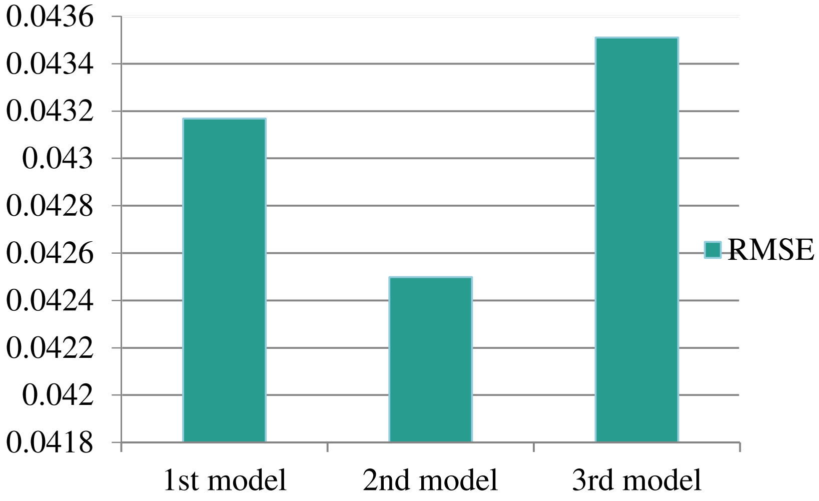


Figure 7 RMSE comparison between all proposed models

4. CONCLUSION

GP was used to solve the problem of real-time runoff forecasting for the Sungai Sayong, Johor. The rainfall data and stage of the river for ten years intensity were used to forecast the runoff since rainfall and runoff are deeply related matters and the team wants to ensure suitably severe conditions were utilized for the analysis for the safety, health and environment aspect of the study are covered. Based on the study and analysis, the RMSE of the data runs in those multiple input variables cases are in a range of 0.04 – 0.08 which is relatively small and within an acceptable range. Being the latest technique developed for rainfall-runoff modeling and runoff forecasting, it is concluded that GP is the better technique compared to the other rainfall-runoff modeling methods available.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Ministry of Higher Education of Malaysia for the financial support under Fundamental Research Grant Scheme (FRGS) with reference number FRGS/1/2015/TK01/UNITEN/02/2(20160109FRGS) and FRGS/1/2016/TK03/UNITEN/02/1 and the Universiti Tenaga Nasional (UNITEN) for the financial support under Bold Grant 10289176/B/9/2017/14.

REFERENCES

[1] Chandwani, V., Vyas, S. K., Agrawal, V., & Sharma, G. (2015). Soft computing approach for rainfall-runoff modelling: A review. Aquatic Procedia, 4, 1054-1061.

[2] Haykin, S. (2009). Neural Networks and Learning Machines (3rd Ed.). New York Boston San Francisco London Toronto Sydney Tokyo Singapore Madrid Mexico City Munich Paris Cape Town Hong Kong Montreal: Pearson Prentice Hall.

[3] Flood, I., & Kartam, N. (1994). Neural Networks in Civil Engineering. I: Principles and Understanding. Journal of Computing In Civil Engineering, 8(2), 131-148. doi: 10.1061/(ASCE)0887-3801(1994)8:2(131).

[4] Gasim, H. A., Kutty, S. R. M., Isa, M. H., & Alemu, L. T. (2013). Optimization of anaerobic treatment of petroleum refinery wastewater using artificial neural networks. Research Journal of Applied Sciences, Engineering and Technology, 6(11), 2077-2082.

[5] Hayder, G., Ramli, M. Z., Malek, M. A., Khamis, A., & Hilmin, N. M. (2014). Prediction model development for petroleum refinery wastewater treatment. Journal of Water Process Engineering, 4, 1-5.

[6] [6] Ronald R. Yager, Lotfi A. Zadeh, 1992. An Introduction to Fuzzy Logic Applications in Intelligent Systems.

[7] Tokar, A., & Johnson, P. (1999). Rainfall-Runoff Modeling Using Artificial Neural Networks. Journal of Hydrologic Engineering, 4(3), 232-239. doi: 10.1061/(asce)1084-0699(1999)4:3(232).

[8] M.P. Rajurkara, U.C. Kothyarib, U.C. Chaubec, 2004. Modeling of the daily rainfall-runoff relationship with artificialneural network. Journal of Hydrology 285 (2004) 96–113.

[9] Leung, L.R., Mearns, L.O., Giorgi, F. and Wilby, R.L. 2003. Regional climate research: needs and opportunities. Bulletin of the American Meteorological Society, 84, 89-95.

[10] Palani S, Liong SY, Tkalich P (2008). An ANN application for water quality forecasting. Mar Pollut Bull 56:1586–1597.

[11] Singh KP, Basant A, Malik A, Jain G (2009). Artificial neural network modeling of the river water quality, a case study. Ecol Model 220:888–895.

[12] Basant N, Gupta S, Malik A, Singh KP (2010). Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water—a case study. Chemometr Intell Lab Syst 104:172–180.

[13] Cherkassky V, Gehring D, Mulier F. Comparison of adaptive methods for function estimation from samples, IEEE Transactions on Neural Networks, 7 (4): 969-984, 1996.

[14] Eshanthini P, P. Vijayalakshmi, P.K. Raji, Rainfall Runoff Estimation Using SCS Model and Arc Gis for Micro Watershed in Cuddalore District. International Journal of Civil Engineering and Technology, 9(9), 2018, pp. 990-996.

[15] Mrs. Anie John. S and Dr. J. Brema, Rainfall Trend Analysis by Mann-Kendall Test for Vamanapuram River Basin, Kerala, International Journal of Civil Engineering and Technology, 9(13), 2018, pp. 1549-1556