Civil Engineering and Architecture 7(6A): 19-32, 2019 http://www.hrpub.org
DOI: 10.13189/cea.2019.071403

Empirical Mode Decomposition Couple with Artificial Neural Network for Water Level Prediction

Eng Chuen Loh*, Shuhaida Binti Ismail, Azme Khamis

Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

Received August 4, 2019; Revised October 10, 2019; Accepted December 16, 2019

Abstract Natural disaster brings massive destruction towards properties and human being and flood is one of them. In order for the government to take earlier action to reduce the damages, an accurate flood prediction is necessary. In Malaysia, Kelantan is categorized as a high flood risk area, thus this study focuses on Kelantan flood prediction. This study is to investigate the effect of decomposition for water level prediction by applying Artificial Neural Network (ANN) forecasting model. In this study, Empirical Mode Decomposition (EMD) is used as the decomposition method. The best Intrinsic Mode Function (IMF) for each input variable is selected using correlation-based selection method. The results showed that the performance of hybrid EMD and ANN is superior compared to other models, especially classic ANN model. The reason for this outcome is that through decomposition methods, ANN is able to capture more in-depth information of the Kelantan hydrological time series data. The resulting model provides new insights for government and hydrologist in Kelantan to have better prediction towards flood occurrence.

Keywords Artificial Neural Network, Empirical Mode Decomposition, Intrinsic Mode Function, Flood Prediction, Water Level

1. Introduction

Flood is the most commonly happened natural disaster in the world. It also causes tremendous damages to economics, properties, besides threatening human life and safety. In order to reduce such damages, an early issued flood warning is essential. Thus, water level forecasting is essential to predict future flood occurrence. Prevention, protection and preparation plan can be made by the government while evacuating the affected citizens [1]. Water level prediction also benefits other sectors such as agriculture, plants, domestics and industrial and commercial [2].

There are several impactful factors that affect inconsistent flood occurrence. For example, temperature, humidity, dew point temperature, wind speed, streamflow volume, and rainfall volume. The streamflow volume indicates how much the volume of water the river can hold to sustain the rainfall volume. Higher temperature and wind speed result in faster water particles' moves thus easier to evaporate into the atmosphere. Humidity also affects the water particle in the air to be condensed out of the atmosphere.

Kelantan is chosen as the study area for water level prediction. According to reports, the occurrence of flood in Kelantan is twice as much compared to other states in Peninsular Malaysia. This is because the geological location of Kelantan with located near to the Northeast Sea which is affected by the Monsoon Season every November to March. Besides, Malaysia is located near the Equator which has higher average temperature compared to other countries.

Water level data are rather complex and consistent. Thus, classical forecasting methods which require the fulfillment of several assumptions such as stationary and linearity are not suitable. A more reliability and adaptability approach is essential to achieve a more accurate prediction. In the past decades, the number of studies applying machine learning approaches in hydrological time series prediction has increased. This is because machine learning approaches are more reliable and accurate compared to the classical methods which attract the attention of researchers [3].

Artificial Neural Network (ANN) is one of the commonly used methods in machine learning approaches. This method imitates the processes of a human brain which receives, processes and produces information through a series of nodes and connections. This enables ANN to provide higher flexibility and adaptability towards complex and chaotic data handling. Besides, these methods can also be reconstructed into different architecture in order to deal with different data such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN)

and even hybrid with other methods.

Due to the inconsistency property of the hydrological time series data, data preprocessing methods are needed to reduce the noise of the data [4]. Decomposition methods can be used to reduce the noises in the signal or time series thus increasing the accuracy of the prediction. There are several different decomposition methods existing such as Empirical Mode Decomposition (EMD), Singular Value Decomposition (SVD), Principle Component Analysis (PCA), Discrete Wavelet Transformation (DWT) and more.

Previous results show that the application of ANN in dealing with hydrological time series is significant. ANN has also been used in various fields, for example, groundwater [5, 6], flood magnitude [7], water level [8, 9], streamflow volume [10, 11], water quality [12, 13] and more. However, there are spaces for improvement and remodel However, there are still spaces for improvement and remodel. Besides, different location would have different outcome due to the different distribution of data and variables.

This study aimed to predict Kelantan water level using ANN model. Empirical Mode Decomposition (EMD) is used to decompose the Kelantan hydrological data. The model is evaluated using four performance accuracies which are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Arctangent Absolute Percentage Error (MAAPE). This study enables the government and hydrologist to have better insight in dealing with hydrological time series forecasting.

2. Methodology

EMD with the application of ANN in forecasting Kelantan water level for 30 days ahead forecast is compared with the single ANN model. The best model is determined by comparing the performance accuracy of each model.

2.1. Data

In this study, Kelantan is chosen as the study area for water level prediction. The hydrological data of Kelantan are retrieved from the Department of Irrigation and Drainage Malaysia (DID). The rainfall data ranging from 2005 to 2014 are taken from 33 stations whereas streamflow and water level data are taken from 3 stations. On the other hand, the temperature, humidity, dew point temperature, wind speed, and pressure are collected at Sultan Ismail Petra Airport which is located at Kota Bharu for the same period of study.

That this case study focused on these specific locations is because the locations are in high land and due to the completeness of the data. Moreover, these areas are near to farms and paddy fields. This case study will also benefit to the agriculture section as well as the government and hydrologists.

Figure 1, Figure 2, and Figure 3 show the water level time series for 3 different rivers in Kelantan which are Sungai Lanas, Sungai Kelantan, and Sungai Golok respectively.

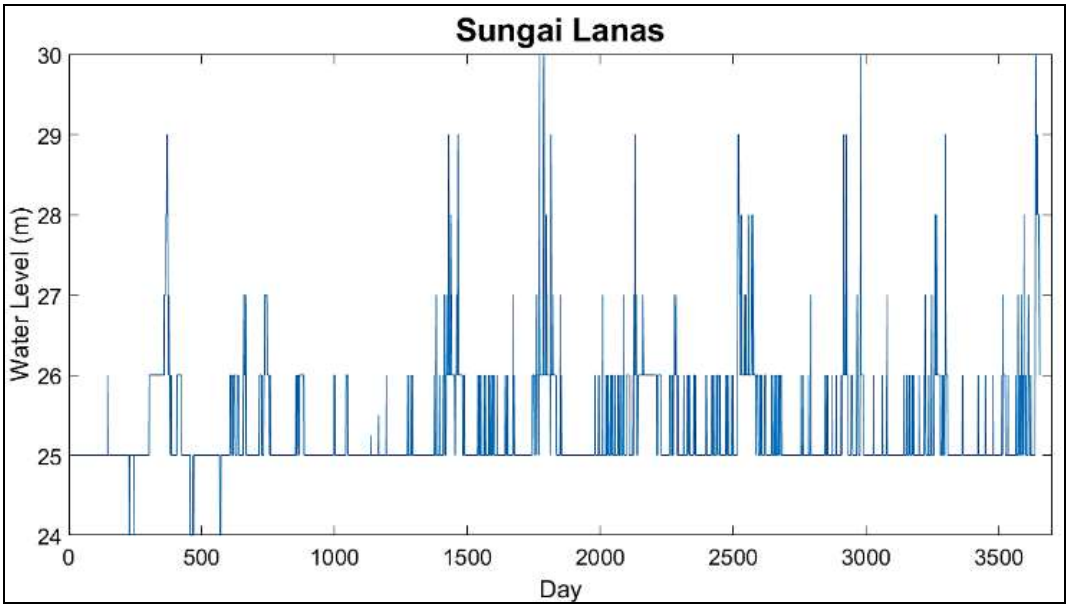

Figure 1. Water Level Time Series of Sungai Lanas

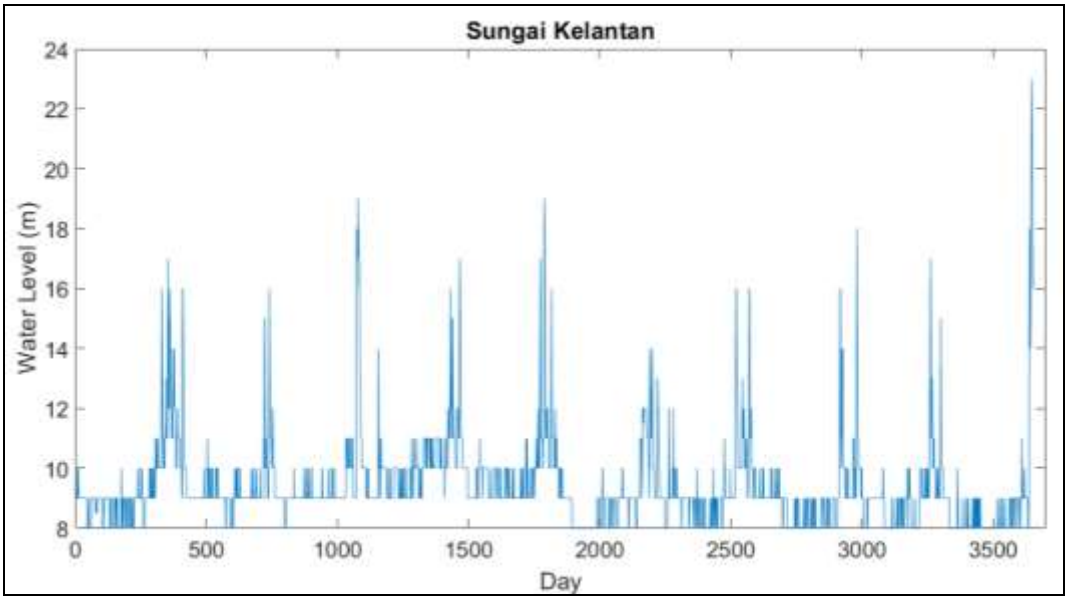

Figure 2. Water Level Time Series of Sungai Kelantan

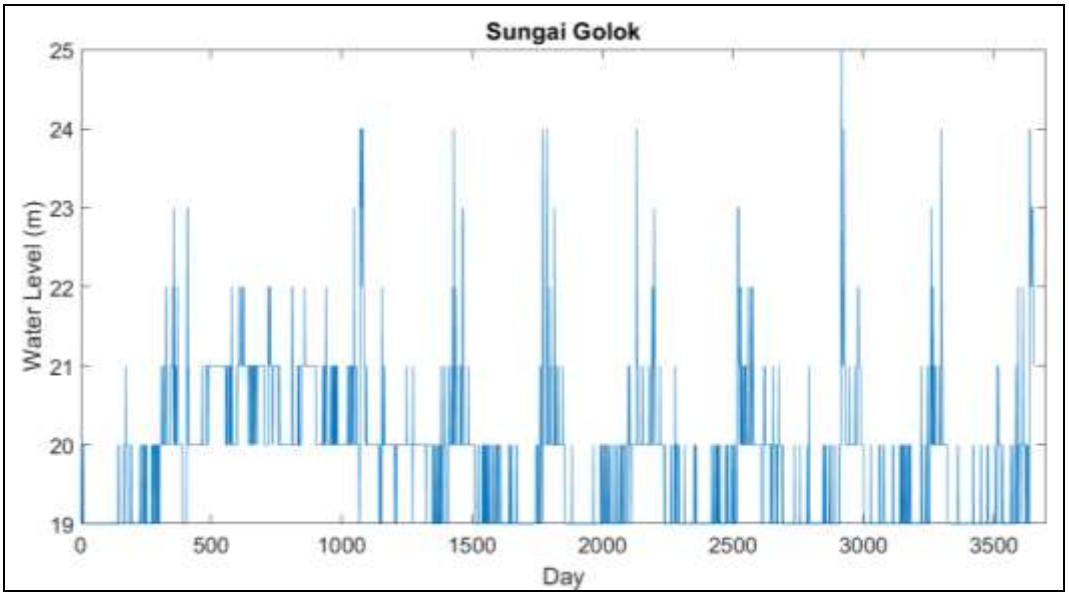

Figure 3. Water Level Time Series of Sungai Golok

2.2. Data Preprocessing

The tendency and accuracy of the ANN models would deteriorate if the input values' range is inconsistent or exceptionally large [14]. In order to overcome this issue, normalization method should be carried out to transform the input value into a predefined range. In most cases, the range is set within 0 and 1 [15]. In this study, min-max normalization method is used to normalize the Kelantan's hydrological time series data. The equation of min-max normalization is shown as follows:

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

where X_i is the normalized value, x_i is the actual value, $x_{\min}$ is the minimum actual value, and $x_{\max}$ is the maximum actual value.

2.3. Artificial Neural Network (ANN)

Artificial Neural Network is the commonly used method in machine learning approaches. It was introduced by McCulloch & Pitts in 1943 that was inspired by the architecture and the processing method of a human brain [16]. This method is able to learn the information through the input layer and processes it using one or several hidden layers then processes it through the output layer. Due to its superiority, it can be applied in various fields such as image data, signal data, time series data and more [17].

The basic architecture of ANN consists of one input

layer, hidden layer and output layer each. Each of the layers is interconnected with a series of weighted connections [18]. Each of the weighted connections is constantly changing in order to optimize the result by using a back-propagation training algorithm. The error is propagated back in order to adjust the weighted connection until the best performance is obtained. The weight of the connections is adjusted based on (2).

$$w_{t+1} = w_t + \eta (y_i - \hat{y}_i)\hat{y}_i \quad (2)$$

where w_{t+1} is the weight at time t, y_i is the target output, $\hat{y}_i$ is the predicted output and η is the learning rate.

In this study, a three-layer ANN with Levenberg -Marquardt backpropagation training algorithm is used. The Levenberg-Marquardt backpropagation training algorithm is modified from the Gauss-Newton method [19]. This architecture consists of α numbers of input nodes, β numbers of hidden nodes and 7 output nodes. The predicted outputs are obtained based on (3).

$$y_t = f\left[\sum_{j=1}^{\beta} w_j g\left(\sum_{i=1}^{\alpha} w_i x_i + w_i \theta\right) + w_j \theta\right] \quad (3)$$

where y_t is the output value at time t, x_t is the input value at time t, w_i is the weight connection between input and hidden nodes, w_j is the weight connection between hidden and output nodes, θ is the bias constant, α is the number of input nodes, β is the number of hidden nodes, $f(x)$ and $g(x)$ are the respective activation functions showed in (4) and (5).

$$f(x) = purelin(x) = x \quad (4)$$

$$f(x) = tan\ sig(x) = \frac{2}{1+e^{-2x}} - 1 \quad (5)$$

where x is the input/hidden node values.

2.4. Hybrid Decomposition with ANN

In this study, hybrid models with various decomposition methods are proposed to predict the water level at 3 rivers in Kelantan which are Sungai Lanas, Sungai Kelantan, and Sungai Golok

The input data undergo decomposition and then are connected to the hidden layer, simultaneously being processed with the raw input data. The architecture of the hybrid model is shown in Figure 4.

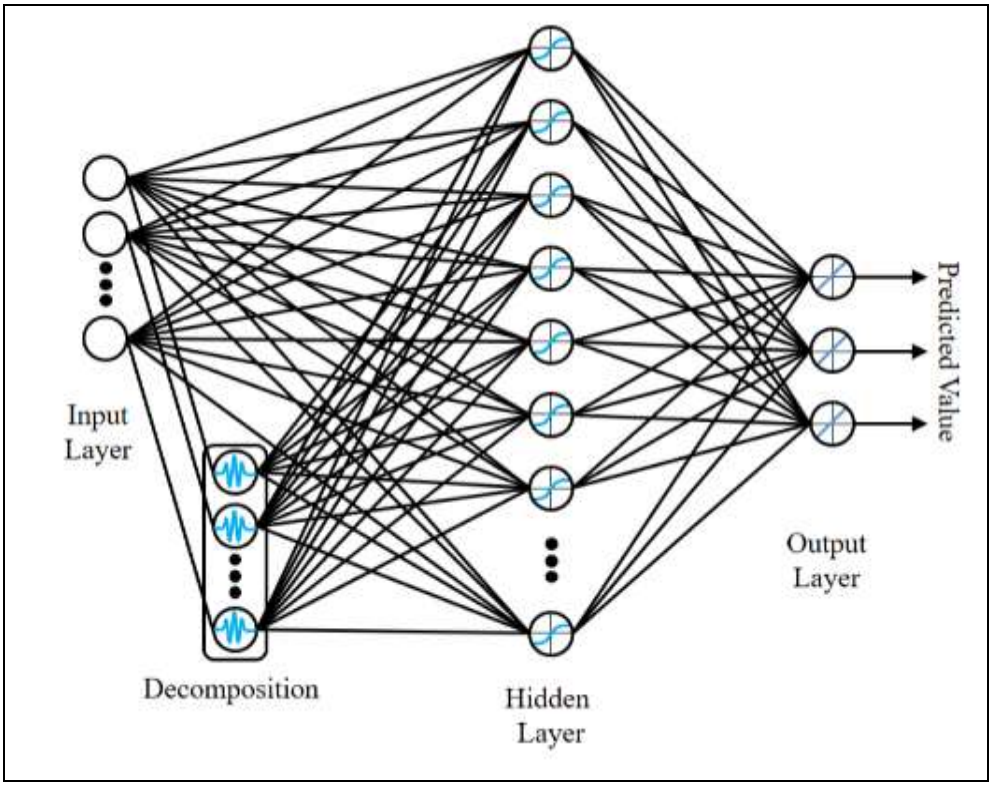

Figure 4. Hybrid ANN Architecture

2.5. Empirical Mode Decomposition (EMD)

EMD is the fundamental part of the Hilbert-Huang transform (HHT) that decomposes the signal into the terms of Intrinsic Mode Function (IMF) [20]. There are two main steps needed to perform EMD which are firstly obtaining the IMF through the EMD algorithm then obtaining the instantaneous frequency spectrum of the initial sequence using HHT.

In EMD algorithm, local maxima and minima are determined using the smooth envelope in order to produce upper and lower envelopes by connecting all local maxima and minima using cubic spline lines. The IMFs are determined through a series of subtraction with the local mean value. For every extraction of the IMF, a new set of maxima and minima are produced. These sifting processes stopped when the residual became a monotonic function where IMF extraction is no longer available [21]. Figure 5 illustrates the flow chart of the EMD algorithm. The original signal can be reconstructed as in (6).

$$S = \sum_{i=1}^{n} F_i + R_n \tag{6}$$

where F_i is the IMF, R_n is the final residue, and n is the number of IMF identified.

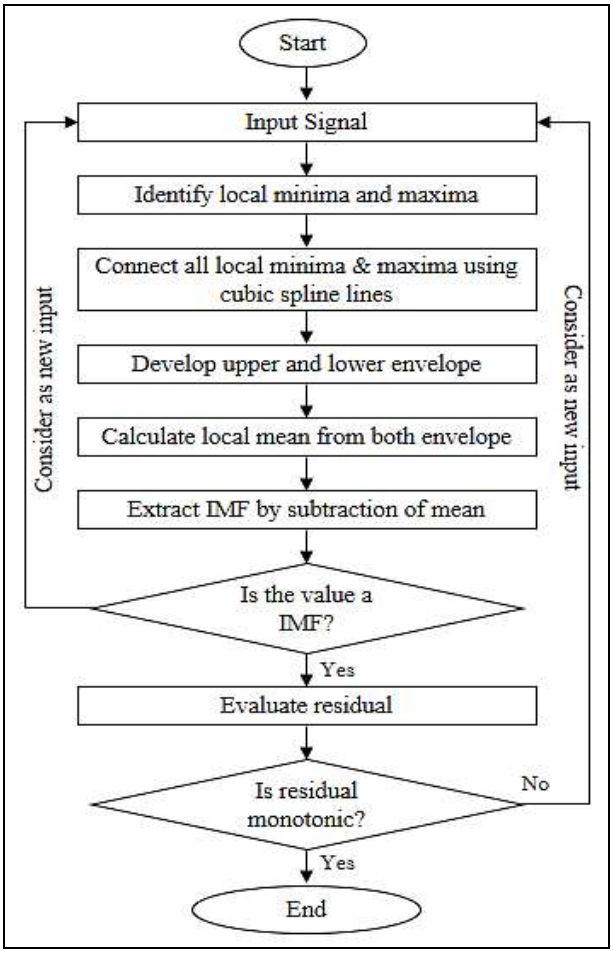

Figure 5. Flow of IMF Search using EMD Algorithm

There are several selection methods for reasonable IMF such as frequency-based method, Kurtosis-based method, energy-based method, correlation-based method and more [22]. In this study, the correlation-based method is used to determine the suitable IMF. IMF with higher correlation coefficient indicates that the IMF is significant. The formula of the correlation coefficient is as shown in (7).

$$r = \frac{1}{n-1}\sum\left(\frac{o-\bar{o}}{S_o}\right)\left(\frac{y-\bar{y}}{S_y}\right) \tag{7}$$

where n is the number of observation, o is the actual value, $\bar{o}$ is the mean actual value, S_o is the standard deviation of actual value, y is the IMF value, $\bar{y}$ is the mean IMF value, and S_y is the standard deviation of IMF value.

2.6. Training Parameters

The parameters shown in Table 1 were used for all the neural network models. All models are set with the same parameter in order to obtain fair results. Meanwhile, the architecture of each neural network model is shown in Table 2.

Table 1. Training Parameters of Neural Network

Parameter	Value
Training Algorithm	Levenberg-Marquardt
Data Partition	70:15:15
Transfer function	tan-sigmoid + linear
Maximum fail	6
Maximum epochs	500
Learning rate, α	0.01
Performance goal	0
Minimum gradient	1.00×10^{-6}
μ	1.00×10^{-3}
Maximum μ	1.00×10^{10}

Table 2. Architecture for each Neural Network

Method	Model	Architecture
M_1	ANN	44-89-3
M_2	EMD-ANN	55-106-3

There are different approaches in deciding the number of hidden neurons in the neural network [23, 24]. In this study, the number of hidden neurons is set as $2n+1$ where n is the number of input neurons [25].

2.7. Performance Accuracy

In this study, four types of performance accuracy are applied to evaluate the accuracy of the predicted output for the forecasting models. The measurements that will be used in this study are MAE, MSE, RMSE, and MAAPE, whereby, the best model will be selected based on the

smallest values for all measurements.

MAAPE is used instead of Mean Absolute Percentage Error (MAPE) because MAPE will have difficulty dealing with actual value approaches zero. MAAPE is more robust and less biased compared to MAPE due to the bounded influence of outliers [26]. The equations for each of the performance accuracy are shown as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

$$MAAPE = \frac{1}{n}\sum_{i=1}^{n}\arctan\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (11)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted output, and n is the number of observation.

3. Result and Discussions

Figure 6 to Figure 16 show that the best IMF for each input variables using correlation-based method selection.

Based on results obtained, it is concluded that hybrid of EMD with ANN is able to predict the Kelantan's water level accurately. Moreover, the model is able to predict Sungai Lanas with the lowest error compared to other location where it obtained MAE of 0.3140, MAAPE of 0.0122, MSE of 0.2396, and RMSE of 0.4895.

The results also indicate that single ANN yielded the worst performance accuracy. This is due to the fact that single ANN is unable to perform well because of the incapability in dealing with an extremely large set of input variables and relatively complex data. Moreover, Table 6 shows the processing speed for training the ANN architecture. It shows that the single ANN's speed is as triple as lower compared to other models, however, its accuracy is insignificant.

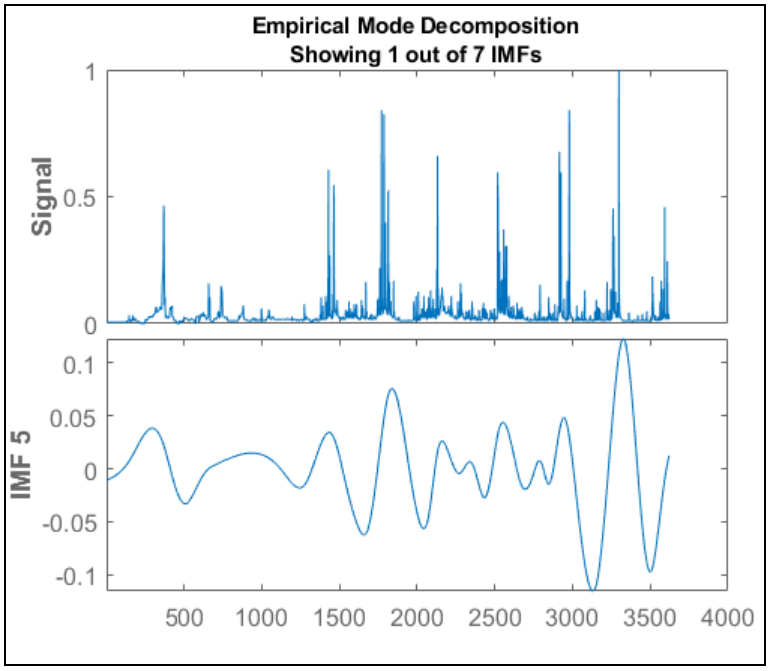

Figure 6. Best IMF of Streamflow Volume of Sungai Lanas

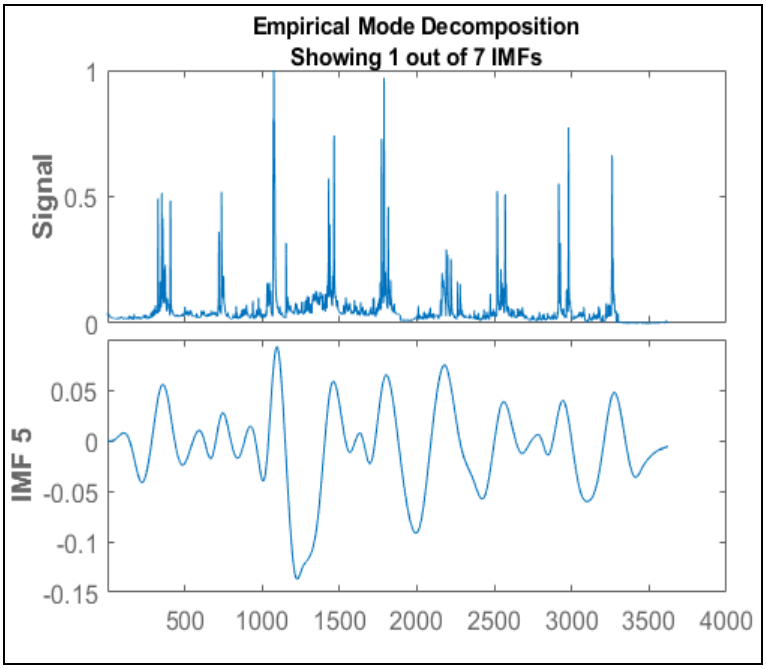

Figure 7. Best IMF of Streamflow Volume of Sungai Kelantan

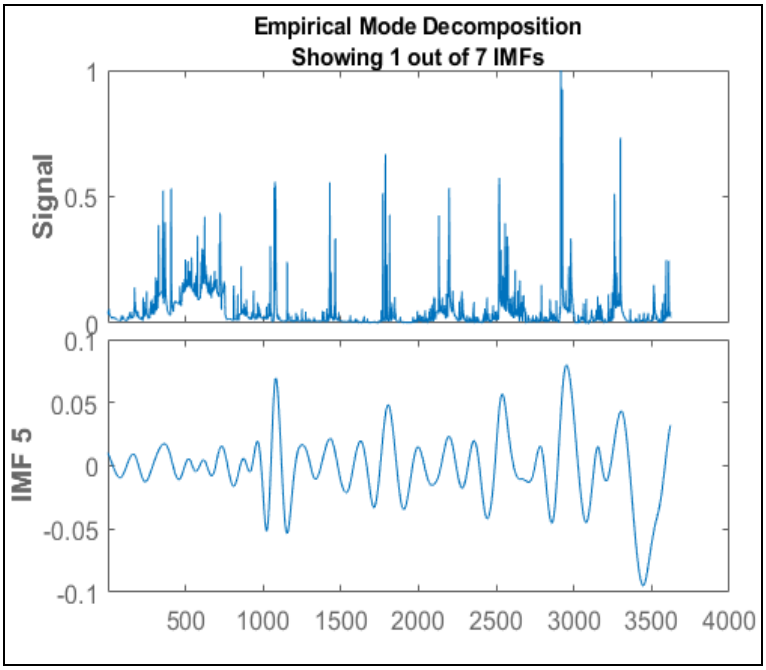

Figure 8. Best IMF of Streamflow Volume of Sungai Golok

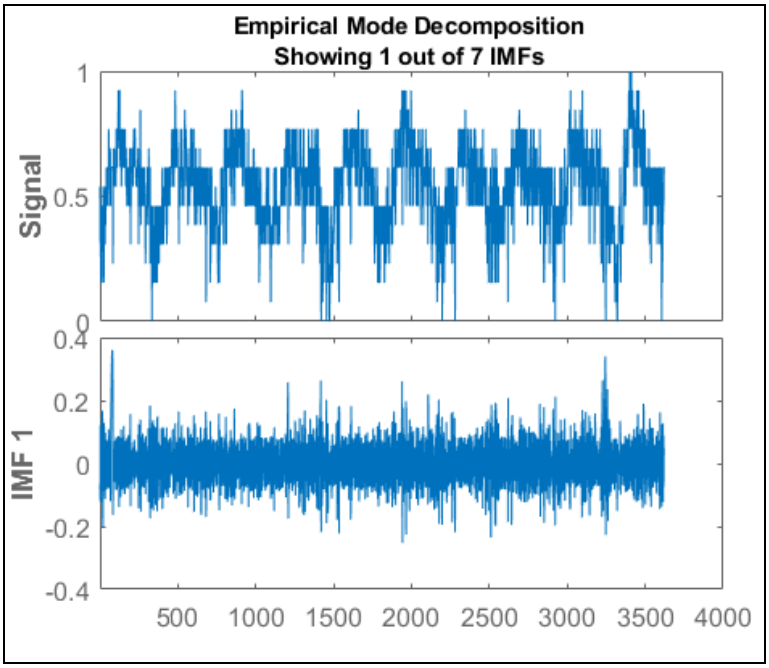

Figure 9. Best IMF of Temperature

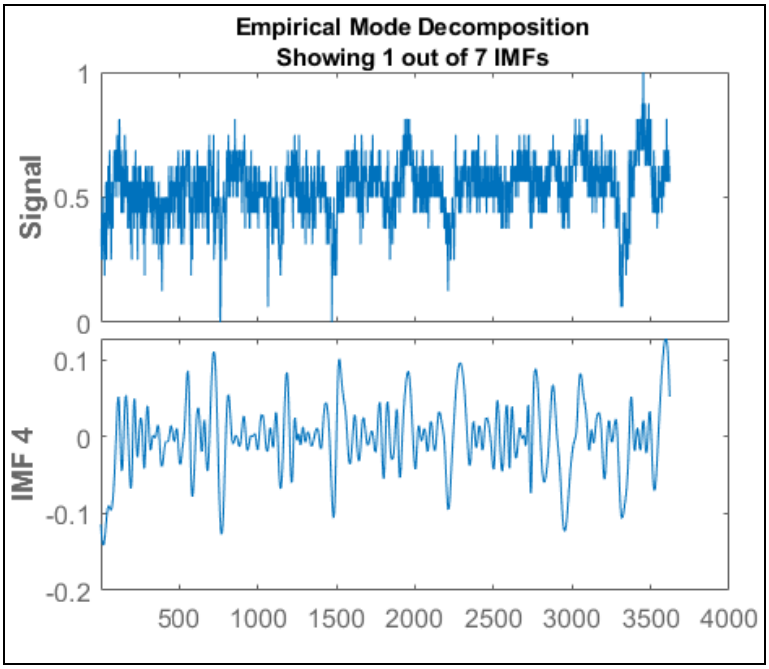

Figure 10. Best IMF of Dew Point Temperature

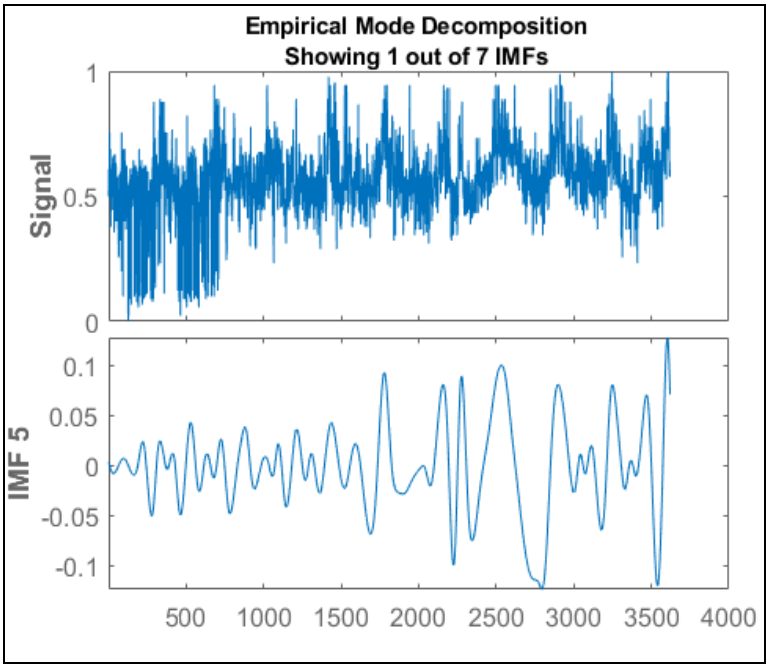

Figure 11. Best IMF of Humidity

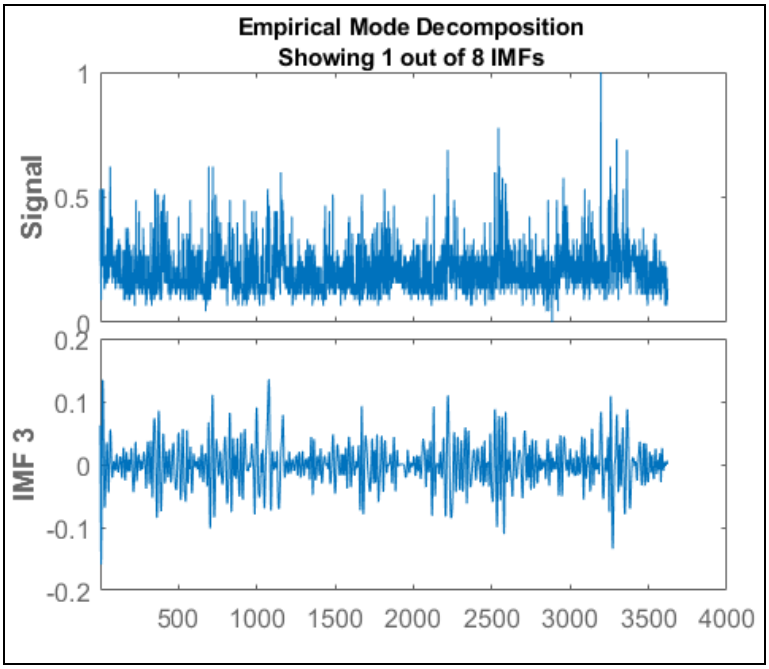

Figure 12. Best IMF of Wind Speed

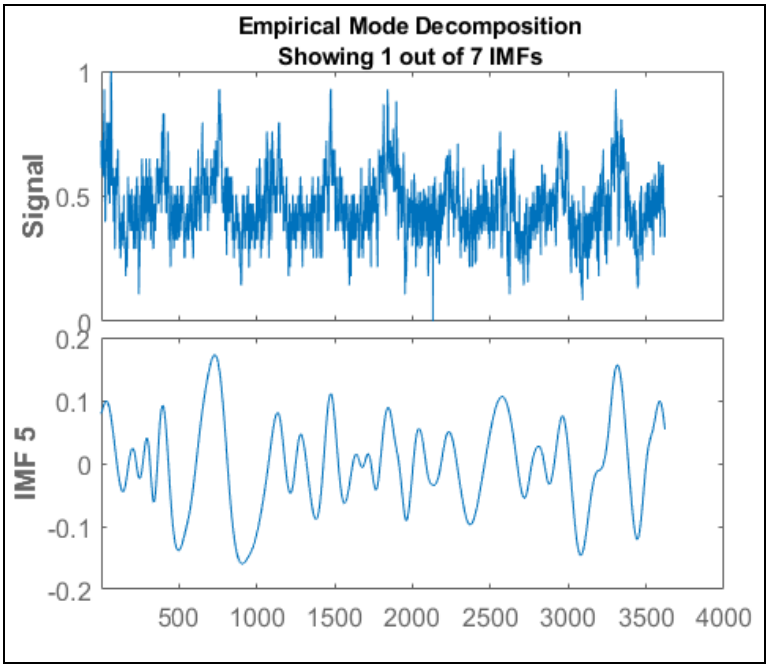

Figure 13. Best IMF of Atmospheric Pressure

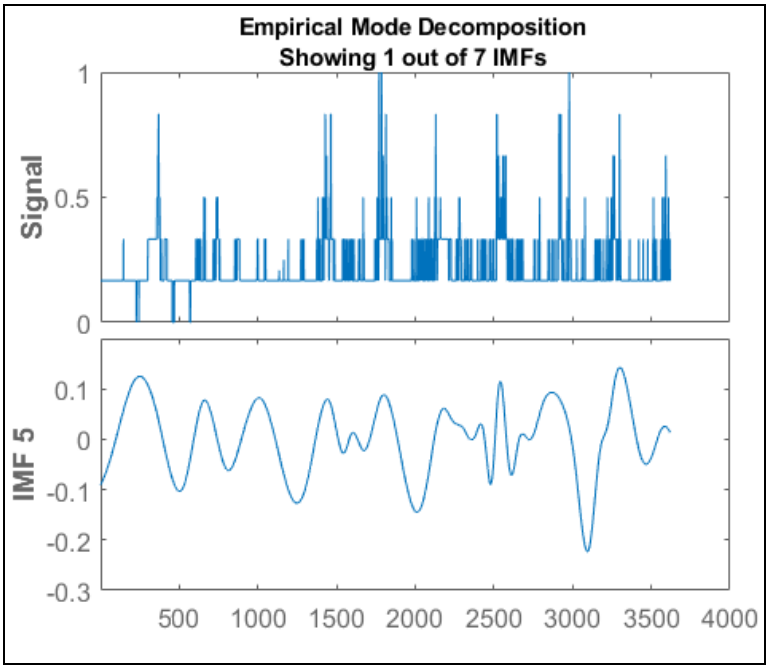

Figure 14. Best IMF of Water Level at Sungai Lanas

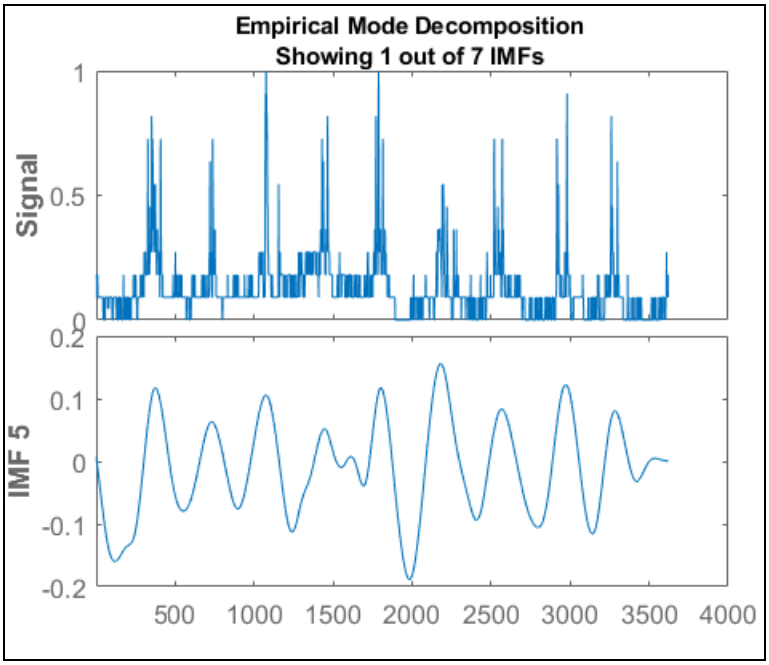

Figure 15. Best IMF of Water Level at Sungai Kelantan

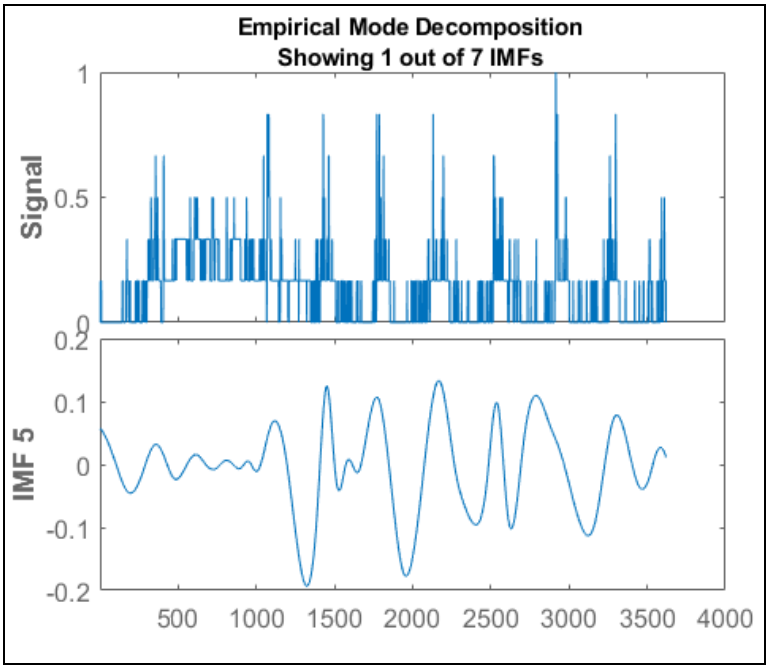

Figure 16. Best IMF of Water Level at Sungai Golok

Table 3. Performance Accuracy for Sungai Lanas

Model	MAE	MAAPE	MSE	RMSE
M_1	0.3911	0.0151	0.3526	0.5938
M_2	**0.3140***	**0.0122***	**0.2396***	**0.4895***

***** Indicate the best result among all of the models

Table 4. Performance Accuracy for Sungai Kelantan

Model	MAE	MAAPE	MSE	RMSE
M_1	0.7945	0.0784	1.6873	1.2990
M_2	**0.6207***	**0.0623***	**1.0053***	**1.0027***

***** Indicate the best result among all of the models

Table 5. Performance Accuracy for Sungai Golok

Model	MAE	MAAPE	MSE	RMSE
M_1	0.5094	0.0253	0.5121	0.7156
M_2	**0.3745***	**0.0187***	**0.2837***	**0.5327***

***** Indicate the best result among all of the models

Table 6. Processing Speed for each Model

Model	Iteration	Time Taken (s)
M_1	13*	224*
M_2	15	765

***** Indicate the best result among all of the models

From the visual inspection of Figure 17, the model is able to capture the peaks within the period of 1st day to 500th day, 1500th day to 2000th day, and 3000th to 3621st day as shown in red rounded rectangle. However, the model has failed to predict several peaks during period of 500th day to 1500th day, and 2000th day to 3000th day as shown in purple rounded rectangle.

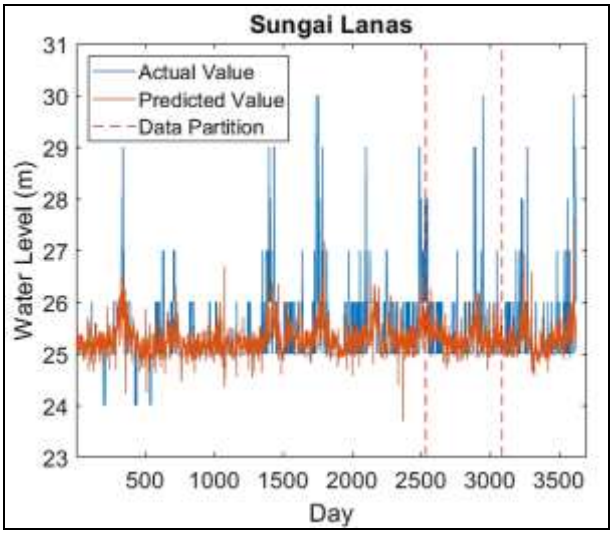

Figure 17. Actual vs Predicted Value of Lanas

From the visual inspection of Figure 18, the model is only able to capture the peaks within period of 1st day to 500th day, 1500th day to 2000th day, and 3000th day to 3500th day as shown in red rounded rectangle. But, the model failed to predict most of the peaks during period 500th day to 1500th day, 2500th day to 3000th day, and 3500th day to 3700th day as shown in purple rounded rectangle. Also, the model predicted a dried out situation where the water level is much lower compared to the average water level. This situation occurred during the period of 2000th day to 2500th day as shown in green rounded rectangle.

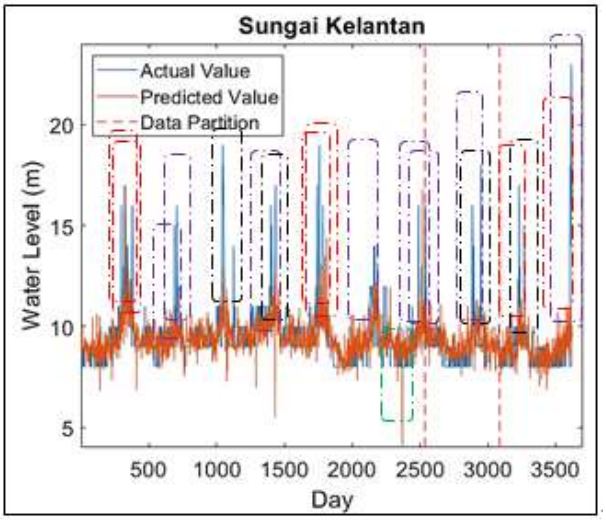

Figure 18. Actual vs Predicted Value of Kelantan

From the visual inspection of Figure 19, the model successfully captured peaks during the period of 1000th day, 2000th day to 2500th day, and 3000th day to 3700th day as shown in red rounded rectangle. However, the model failed to capture a few of the peaks within the period of 1000th day to 2000th day, and 2500th day to 3000th day as shown in purple rounded rectangle. In general, these models successfully capture the seasonal trend of water level data.

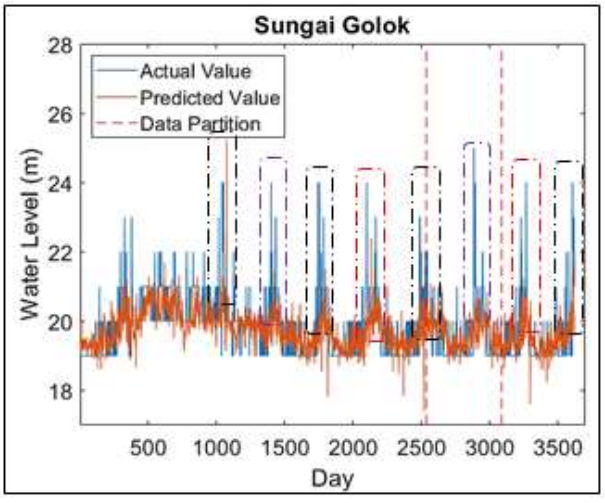

Figure 19. Actual vs Predicted Value of Sungai Golok

According to the figures of actual versus predicted value, they showed that the predicted values are relatively close to the actual value. In addition, the figures always proved that

the hybrid of EMD and ANN has the ability to capture the seasonal trend of the hydrological time series significantly.

From the figures, the model is also able to capture the peak of each location. This enables the government to foresight the future flood occurrence more accurately and how severe the flood will be. However, the predicted values contain more noise compared to the actual value. This is due to the noise carried forward from the 33 stations for rainfall volume data collection. Yet, the robustness of the model remains.

According to the visual inspection of Figure 20 to Fig 22, it is concluded that the model is significant in predicting Kelantan water level data. Figure 20 and Figure 22 showed that the model is able the capture the uptrend of the water level in Sungai Lanas and Sungai Golok during the time period from 17th day onwards. However, Figure 21 shows that the model unsuccessfully captures the uptrend in Sungai Kelantan.

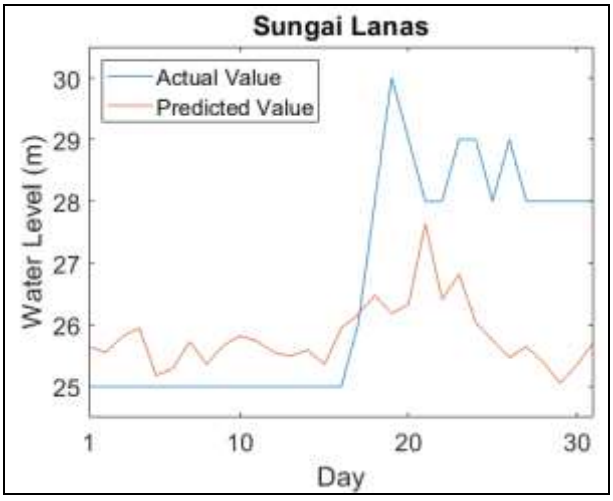

Figure 20. A Month Ahead Forecast of Sungai Lanas

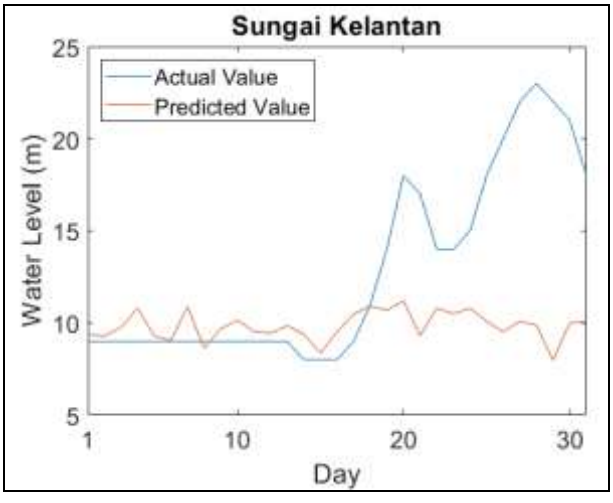

Figure 21. A Month Ahead Forecast of Sungai Kelantan

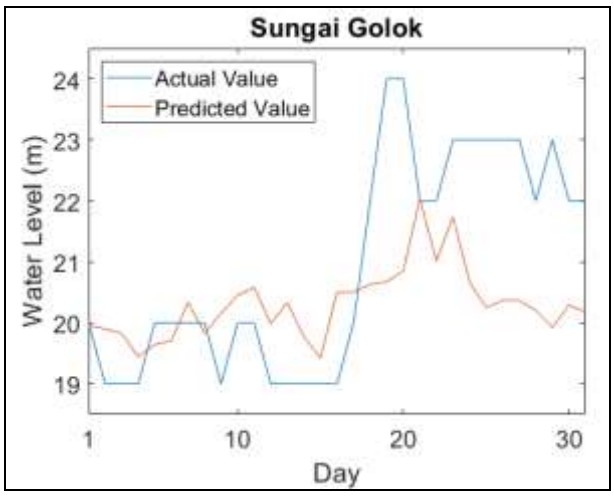

Figure 22. A Month Ahead Forecast of Sungai Golok

4. Conclusions

In conclusion, the hybrid of EMD and ANN model is superior compared to the other five models. Moreover, this model is able to predict Sungai Lanas with the lowest error among all three rivers.

However, there are some limitations in forecasting Kelantan water level data using the current model. The network structure of the model contributes significant influence toward its accuracy. For example, the number of the hidden layers, the number of the hidden neurons, the type of family member value, the level of decomposition, the training algorithm and more. Thus, in future studies, determining the suitable network structure is essential through trial and error method or optimization methods such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and more.

Acknowledgements

This study is supported by Ministry of Education and Universiti Tun Hussein Onn Malaysia (UTHM) via the Fundamental Research Grant Scheme (FRGS) Vot K082.

REFERENCES

[1] Mosavi A, Ozturk O & Chau KW (2018), Flood prediction using machine learning models: Literature review, *Water (Switzerland)*, Vol. 10, No. 11, pp. 1-40.

[2] Yadav V & Eliza K (2017), A hybrid wavelet-support vector machine model for prediction of lake water level fluctuations using hydro-meteorological data, *Measurement: Journal of the International Measurement Confederation*, Vol. 103, pp. 2655-2675.

[3] Alexander AA, Thampi SG & Chithra NR (2018), Development of hybrid wavelet-ANN model for hourly flood stage forecasting, *ISH Journal of Hydraulic Engineering*, Vol. 24, No. 2, pp. 266-274

[4] Oh S-K, Kim W-D & Pedrycz W (2016), Design of radial basis function neural network classifier realized with the aid of data preprocessing techniques: Design and analysis, *International Journal of General Systems*, Vol. 45, No. 4, pp. 434-454.

[5] Gong Y, Zhang Y, Lan S & Wang H (2016), A comparative study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for forecasting groundwater levels near Lake Okeechobee, Florida, *Water Resource Management*, Vol. 30, No. 1, pp. 375-391.

[6] Mohanty S, Jha M, Raul S, Panda RK & Sudheer KP (2015), Using Artificial Neural Network approach for simultaneous forecasting of weekly groundwater levels at multiple sites, *Water Resources Management*, Vol. 29, No. 15, pp. 5521-5532.

[7] Hitokoto M & Sakuraba M (2018), Application of the deep learning flood forecast model against the inexperienced magnitude of flood, *EPiC Series in Engineering*, Vol. 3, pp. 893-901

[8] Khan M & Coulibaly P (2006), Application of Support Vector Machine in lake water level prediction, *Journal of Hydrological Engineering*, Vol. 11, No. 3, pp. 199-205.

[9] Ang HTN, Dat NQ, Van NT, Doanh NN & An NL (2018), Wavelet-Artificial Neural Network model for water level forecasting, *Proceedings of 2018 3rd IEEE International Conference on Research in Intelligent and Computing in Engineering, RICE 2018*, pp 1-6.

[10] Deo RC & Şahin M (2016), An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland, *Environmental Monitoring and Assessment*, Vol. 188, No. 2, pp. 1-24.

[11] Adhikary S, Muttil N & Yilmaz A (2017), Improving streamflow forecast using optimal rain gauge network-based input to Artificial Neural Network models, *Hydrology Research*, pp. 1-20.

[12] Alizadeh MJ & Kavianpour MR (2015), Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean, *Marine Pollution Bulletin*, Vol. 98, No. 1, pp. 171- 178.

[13] Sarkar A & Pandey P (2015), River water quality modelling using Artificial Neural Network technique, *Aquatic Procedia*, Vol. 4, pp. 1070-1077.

[14] Zhang GP (2012), Neural network for time-series forecasting, eds Rozenberg G, Bäck T & Kok JN, in *Handbook of Natural Computing*, Berlin, Heidelberg: Springer, pp. 461-477

[15] Chou JS & Thedja JPP (2016), Metaheuristic optimization within machine learning-based classification system for early warnings related to geotechnical problems, *Automation in Construction*, Vol. 68, pp. 65-80.

[16] McCulloch WS & Pitts W (1943), A logical calculus of the ideas immanent in nervous activity, *Buletting of Mathematical Biophysics*, Vol. 5, pp. 115-116.

[17] Cocianu C-L & Grigoryan H (2015), An artificial neural network for data forecasting purposes. *Informatica Economica*, Vol. 20, No. 2, pp. 34-45.

[18] Rosenblatt F (1961), *Principle of Neurodynamics. Perceptrons and the Theory of Brain Mechanism*, Cornell Aeroneutical Lab Inc. Buffalo, New York.

[19] Reynaldi A, Lukas S & Margaretha H (2012). Backpropagation and levenberg-marquardt algorithm for training finite element neural network, *Proceedings – UKSim – AMSS 6th European Modelling Symposium, EMS 2012*, Vol. 2, pp. 89-94.

[20] Damaševičius R, Napoli C, Sidekerskienė T & Woźniak M (2017), IMF mode demixing in EMD for jitter analysis, *Journal of Computational Science*, Vol. 22, pp. 240-252.

[21] Malik H & Mishra S (2016), Artificial neural network and empirical mode decomposition based imbalance fault diagnosis of wind turbine using TurbSim, FAST and Simulink, *IET Renewable Power Generation*, Vol. 11, No. 6, pp. 889-902

[22] Isham MF, Leong MS, Hee LM & Ahmad ZAB (2017), Empirical mode decomposition: A review on mode selection method for rotating machinery diagnosis, *Internation Journal of Mechanical Engineering and Technnology (IJMET)*, Vol. 8, No. 6, pp. 16-26.

[23] Madhiarasan M & Deepa SN (2015), A novel criterion to select hidden neuron numbers in improved back propagation networks for wind speed forecasting, *Applied Intelligence*, Vol. 44, No. 4, pp. 878-893

[24] Mostafa F, Dillon TS & Chang E (2018), *Computational Intelligence Applications to Option Pricing, Volatility Forecasting and Value at Risk.* Springer.

[25] Lippmann RP (1987), An introduction to computing with neural network, *IEEE Assp Magazine*, Vol. 4, No. 4, pp. 4-22.

[26] Kim S & Kim H (2016), A new metric of absolute percentage error for intermittent demand forecasts, *International Journal of Forecasting*, Vol. 32, pp. 669-679.