Recent Trends in Civil Engineering and Built Environment Vol. 4 No. 1 (2023) 331-341

RTCEBE

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/rtcebe
e-ISSN :2773-5184

Stream Flow Forcasting on Pahang River by Time Series Models, ARMA, ARIMA and SARIMA

Yap Zi Ning[1], Sabariah Musa[2]*

[1]Faculty of Civil Engineering and Built Environment,
Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/rtcebe.2023.04.01.035
Received 06 January 2022; Accepted 15 January 2023; Available online 01 May 2023

Abstract: Stream flow forecasting is essential in resources planning and flood response. However, it is a challenge to generate an accurate forecast. In this study, time series models, ARIMA, ARMA and SARIMA are adopted to develop forecast series of two stations, Temerloh and Lubok Paku of Pahang River. SARIMA was chosen as the best forecasting model. SARIMA $(2,0,3)(0,1,0)^{12}$ for Temerloh generated forecast series with MAPE of 18.35 % which indicative of good accuracy modelling. Besides, SARIMA $(0,0,0)(1,1,2)^{12}$ for Lubok Paku demonstrated its MAPE values of 6.54 % which characterized the model as high accuracy. The ability of SARIMA deal with the seasonality of streamflow had increased the precision of forecasting. Besides, ARMA models revealed a lower MAPE values than ARIMA model due to the over-differencing of ARIMA model that lower its accuracy.

Keywords: Time Series Model, Streamflow Forecasting, MAPE, Minitab

1. Introduction

Streamflow forecasting is defined as the prediction of the water amount discharged on a certain waterway during specific period of time [1]. There is a need for hydrologists to achieve an accurate forecasting and modelling of river flow as a it is essential in effective allotment and management of water resources at hydraulic structure. Streamflow is heavily influenced by climate, geographical environment and anthropogenic factors as well. The aims of this study are twofold. Firstly, to demonstrate the role of time series model in develop short-term river flow forecast model that based on historical observed data of River Pahang. Secondly, to compare performances of SARIMA, ARIMA and ARMA model in order to determine the best model for short-term prediction. Monthly streamflow data of River Pahang from January 2010 to September 2019 are applied to develop a short-term period forecast by using time series models, whereas the three-monthly streamflow from October 2019 to December 2019 were also collected to compare with the forecast series. The time series models are parsimonious with one input, which is the streamflow data that obtained from Department of Irrigation and Drainage Malaysia.

In this study, 2 stream flow stations along Pahang River were observed namely Temerloh and Lubok Paku station. Time series model, SARIMA, ARIMA and ARMA model are employed by using

*Corresponding author: sabariah@uthm.edu.my

statistical software, Minitab. The ARIMA models are denoted as (p,d,q) and it will be reduced to the ARMA (p,q) models [2].. Seasonal ARIMA model was denoted with the notation SARIMA(p,d,q)(P,D,Q)S. where (p, d, q) refers to a non-seasonal part of the model whereas the latter (P, D, Q) refers to seasonal part of the model with a seasonality term S. S refers to period of the seasonality.

The nomenclature of parameters of time series model following:

Table 1: Nomenclature

Hyperparameters	Definition
p	number of autoregressive terms (AR)
d	Degree of differencing
q	number of lagged forecast errors in the prediction equation (MA)

The hyperparameters P, D and Q for SARIMA is analogous to seasonal moving average, autoregression and integration term respectively.

2. Methodology

The process for the time series modelling is described in a flowchart which shown in Figure 1. The fundamental four steps for time series modelling including model identification, parameter estimation, diagnostic checking and forecasting.

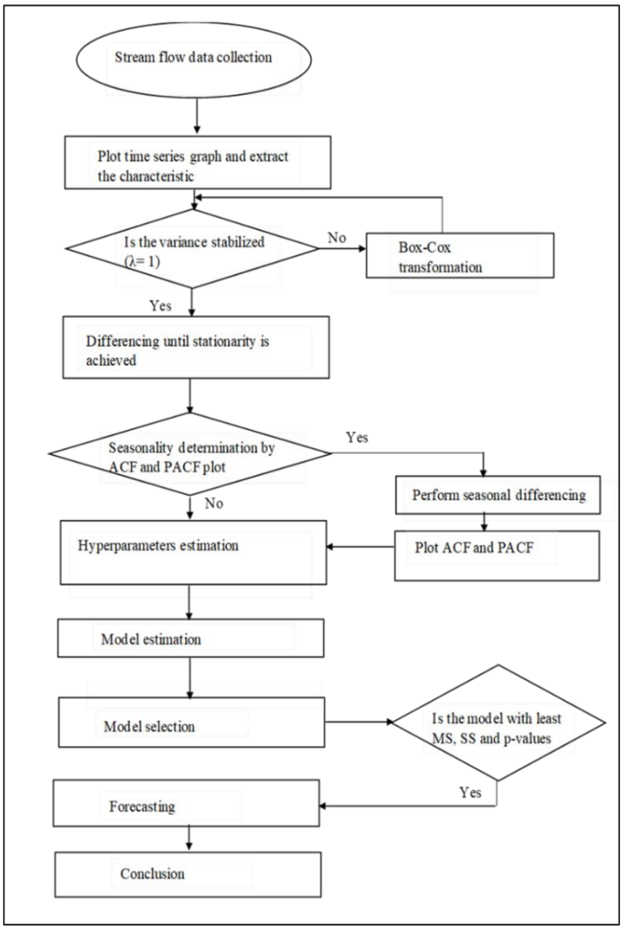

Figure 1: Flow chart of time series modelling

The discharge data from Department of Irrigation and Drainage Malaysia is plotted into graph for identification of data pattern. Next, Box-Cox transformation is required to be performed until the value of λ (power transformation parameter) is equal to 1 to achieve stationary in variance. A time series may be unfixed in the mean, variance or both. Hence, differencing was used to remove linear trend. Next, the ACF and PACF graph should be observed for seasonality identification and hyperparameters identification.In order to determine the most suitable SARIMA, ARIMA and ARMA model, the comparison is made based on the value of p-value, mean squares (MS) and sum of square (SS). The three different time series model with the least MS, SS and p-value will be chosen as the best fitting model. A statistical performance evaluation measure, mean absolute percentage error (MAPE) is adopted. The forecasted results that obtained from modelling are compared with the observed value to compute the value of MAPE. The accuracy of MAPE was defined in terms of percentage error. Hence, the model with lowest MAPE value will be chosen as the best time series model with better performance. The formulas of MAPE will be shown as follow:

$$\text{MAPE (\%)} = \frac{1}{N}\left[\frac{[z_t' - x_t]}{x_t}\right] \text{x } 100\ \%$$

3. Results and Discussion

3.1 Data visualization

According to Figure 2 and Figure 3, both graphs recorded the highest streamflow in January 2015. However, Lubok Paku station had experienced higher stream flow when compared to Temerloh station in January 2015. The flood events on 2014/ 2015 was one of the extreme flood incidents occurred [3]. Based on published report in 2016, the Municipal of Temerloh described this flood event as the longest period for flood inundation which lasted for 23 days (from 28 December 2014 to 15 January 2015) [4]. This disaster affected 7,052 households and 29,204 victims were required to be relocated.

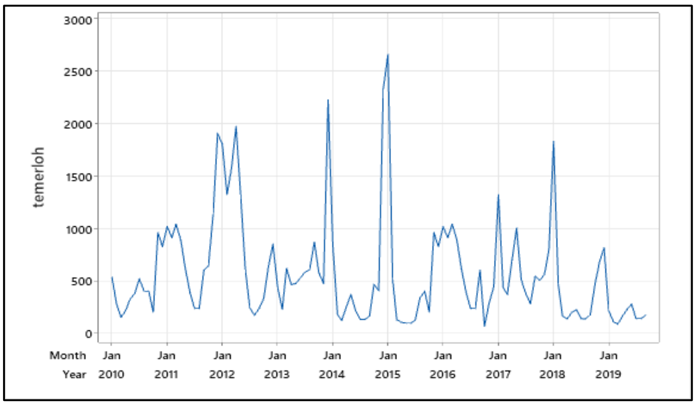

Figure 2: Time series plot of stream flow in Temerloh from January 2010 to September 2019

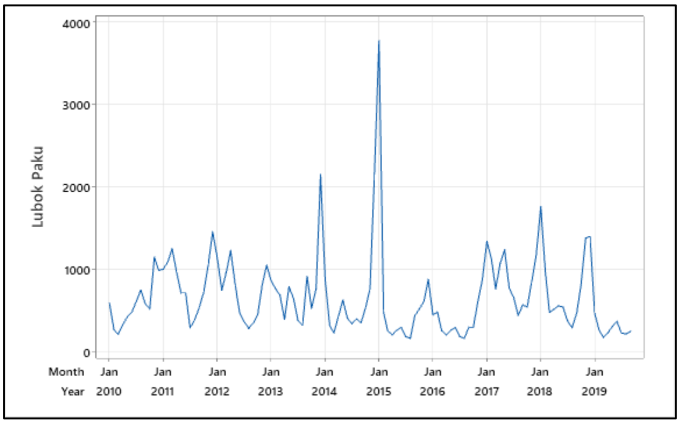

Figure 3: Time series plot of stream flow in Lubok Paku from January 2010 to September 2019

Figure 4 and Figure 5 demonstrated the trend analysis plot for both stations. The best-fit linear trend is plotted to detect significant changes in stream discharge and to approximate the presence of trend. Visually assessment of Figure 4 and Figure 5 suggests that the plots exhibit non-stationarity behavior as the overall streamflow of Temerloh had dipped while Figure 5 presented the upward trend for Lubok Paku . A series with trend is non-stationary series as its mean and variance are vary over time span.

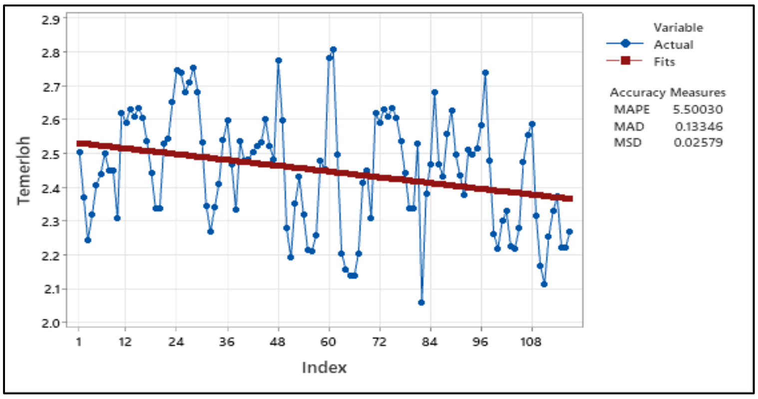

Figure 4: Trend analysis plot for Temerloh station

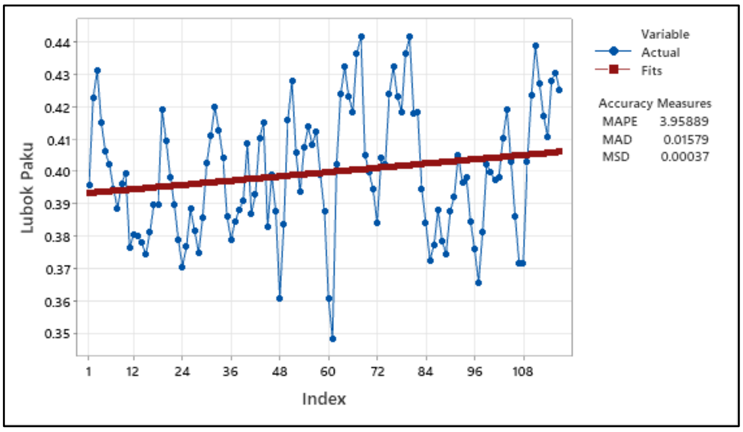

Figure 5: Trend analysis plot for Lubok Paku station

3.2 Box-Cox transformation

Based on the results of Box-Cox testing after transformation which shown in Figure 6 and Figure 7, the plots are stationary to the variance since the rounded λ values were equal to 1 after the logarithmic transformation was performed. Thus, there was no any further transformation required for both stations.

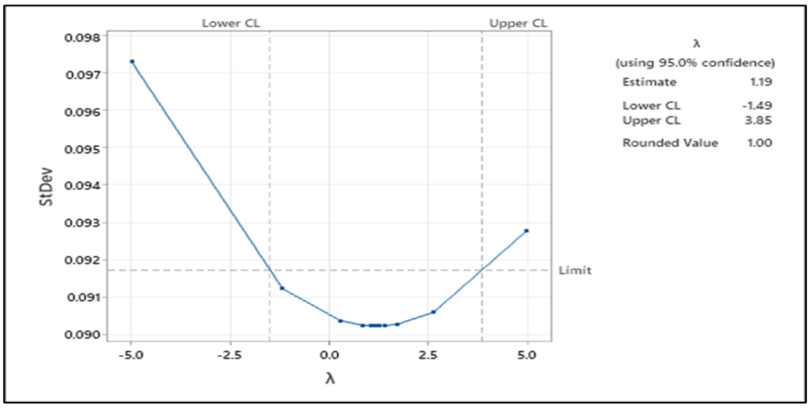

Figure 6: Box- Cox plot for Temerloh station (after stabilization of variance)

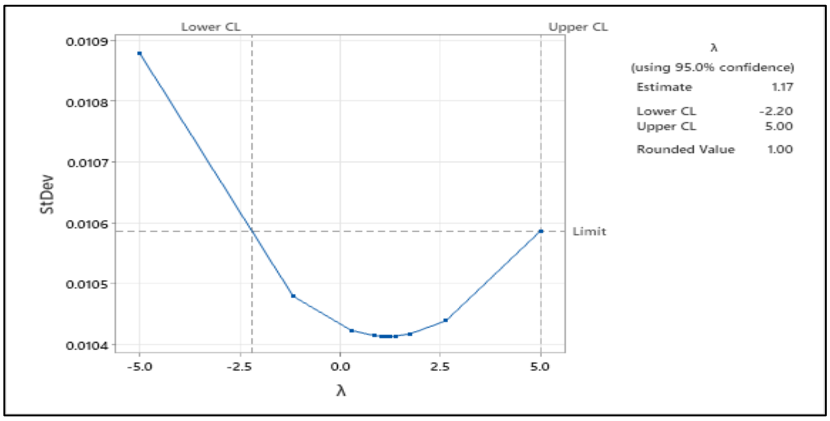

Figure 7: Box- Cox plot for Lubok Paku station (after stabilization of variance)

3.3 Differencing

3.3.1 Temerloh station

According to Figure 4 and Figure 5, the analysis exhibited trend in the plot which was indicative of non-stationarity. Hence, they are required to be corrected through first non-seasonal differencing to remove trend. After the first differencing process, trend analysis plots of first differenced data for Temerloh and Lubok Paku stations were shown as Figure 8 and Figure 9. There were no signs of elevation or decreasing for trend and also appeared horizontal along the x-axis, hence no any further differencing was required as stationarity in mean was ascertained.

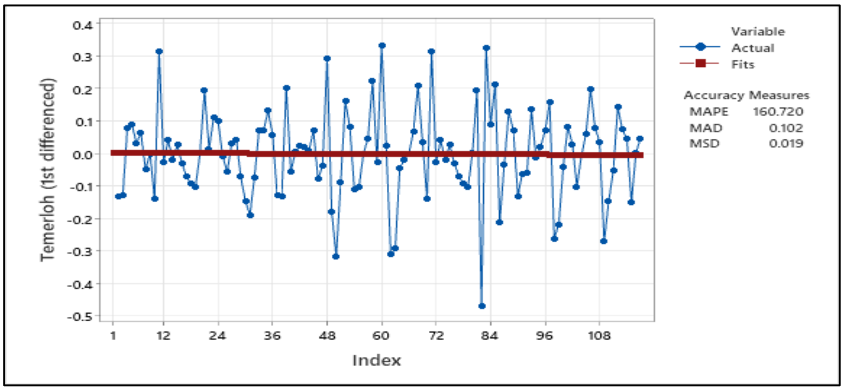

Figure 8: Trend analysis of first difference data for Temerloh station

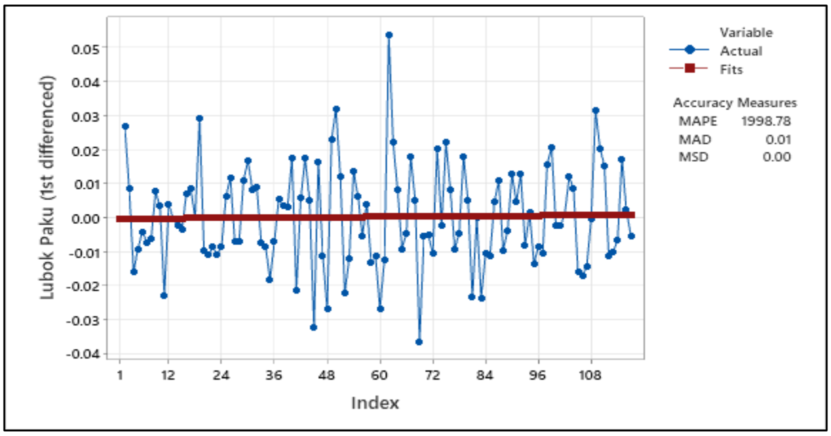

Figure 9: Trend analysis of first difference data for Lubok Paku station

3.4 ACF and PACF

3.4.1 Temerloh station

3.4.1.1 Non-seasonal models

The ACF and PACF of the series with first non-seasonal differenced data, d = 1 were plotted as Figure 10 and Figure 11. A rapid decay was observed for the transformed data as shown, which support the evident of reduction non-stationarity. Both ACF and PACF plots were damping out with presence of significant spike implied that the model can be constructed as a combination of both AR and MA processes. The PACF plot cutoff at the order of 3. Hence, the possible order of autoregressive is equal to 3 (p =3). Similar to PACF plot, the ACF plot also cut off at the order of 3, it can be concluded that the possible order of moving average is 3 (q= 3) as well.

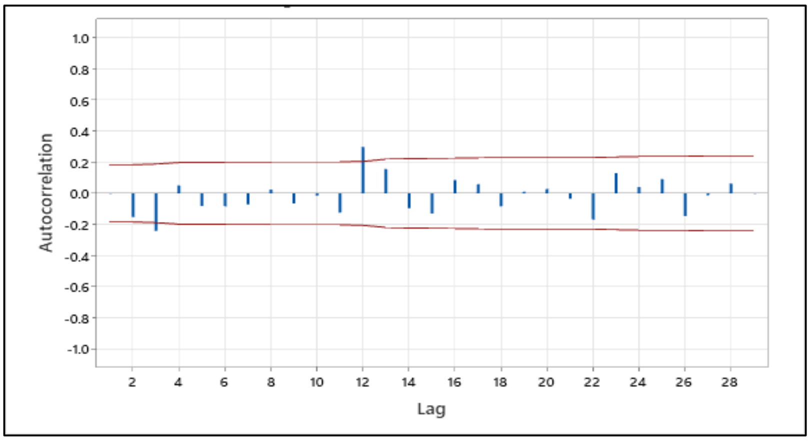

Figure 10: Autocorrelation function of differenced stream discharge data in Temerloh station

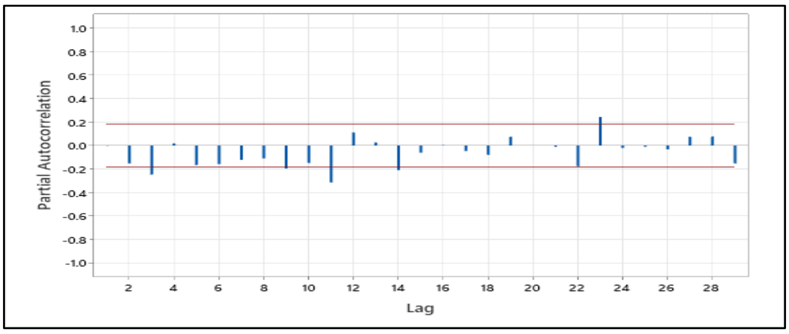

Figure 11: Partial autocorrelation function of differenced stream discharge data in Temerloh station

3.4.1.2 Seasonal model

Figure 12 and Figure 13 indicated the plots of the ACF and PACF for the first seasonal differenced stream flow. ACF plot decayed fast with a sine wave (alternating curve with negative and positive sign) that converged to 0, so q = 0 whereas PACF plot dies down after lag 1 and hence suggesting that p = 1. For the seasonal parameter, spikes can be observed in the PACF plot at lags 12, whereas in the ACF plot, there were no significant spike at lag 12. Therefore, both of the PACF and ACF plots suggest the seasonal AR (1) and MA (0) term (P=1, Q=0).

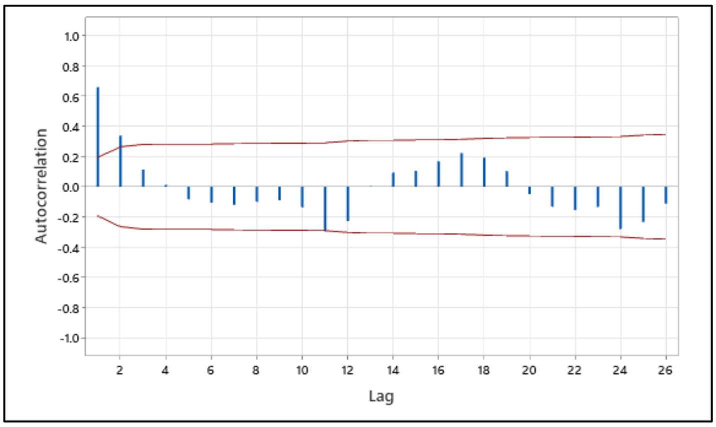

Figure 12.: ACF plot for seasonal differencing of the first differenced data (Temerloh)

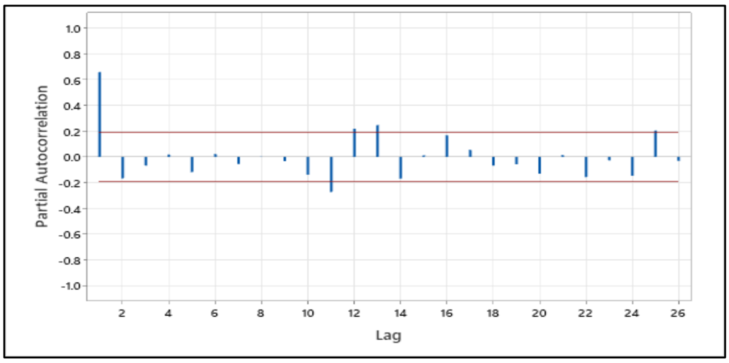

Figure 13: PACF plot for seasonal differencing of the first differenced data (Temerloh)

3.4.2 Lubok Paku station

3.4.2.1 Non-seasonal models

Figure 14 and Figure 15 had demonstrated ACF and PACF for the first differenced stream flow of Lubok Paku station. Both ACF plot and PACF plot dies down after lag 2 and hence suggesting a combination of ARMA model which indicated that q= 2 and p= 2 respectively. Both plots decay rapidly with a sine wave. The rough model was identified as MA (2) and AR (2). For ARIMA and SARIMA models, the d term is equal to one as first differencing was performed.

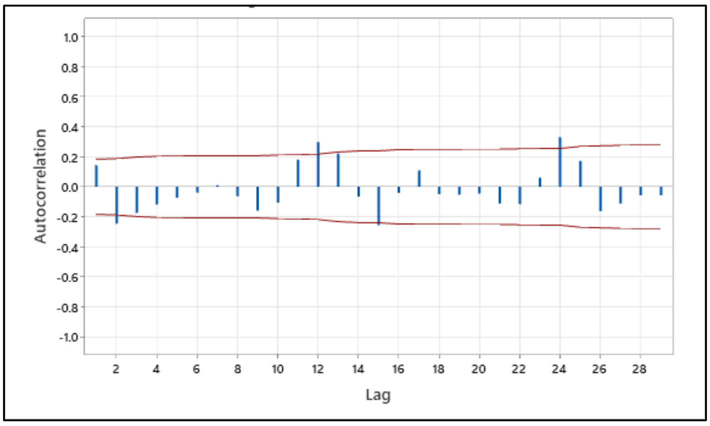

Figure 14: Autocorrelation function of differenced stream discharge data in Lubok Paku station

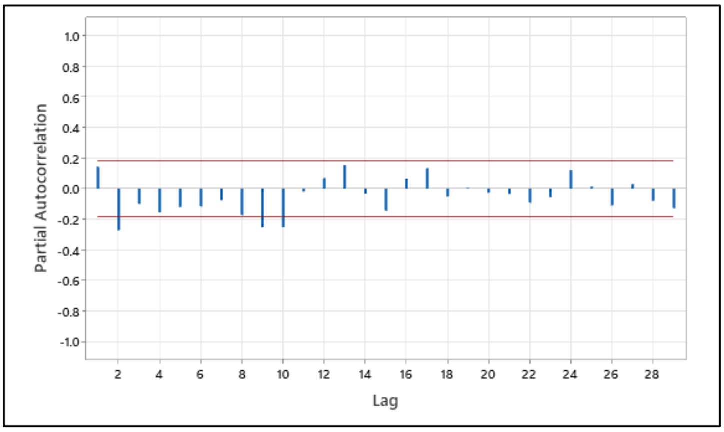

Figure 15: Partial Autocorrelation function of differenced stream discharge datain Lubok Paku station

4.4.2.2 Seasonal model

Figure 16 and Figure 17 indicated the plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) respectively for the first seasonal differenced in stream flow. The ACF plot while damped out with a sine wave while the PACF plot dies down after lag 2 and hence suggesting that Q= 0 and P= 1 respectively.

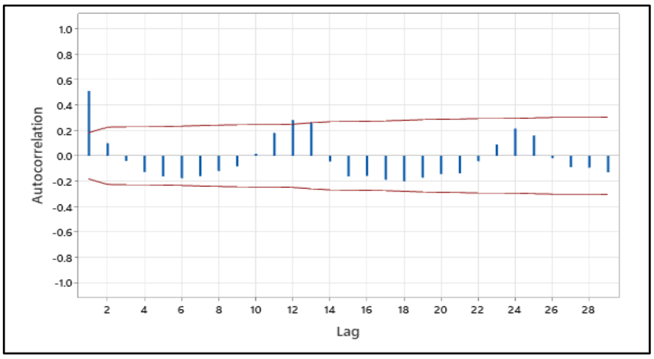

Figure 16: ACF plot for seasonal differencing of the first differenced data (Lubok Paku)

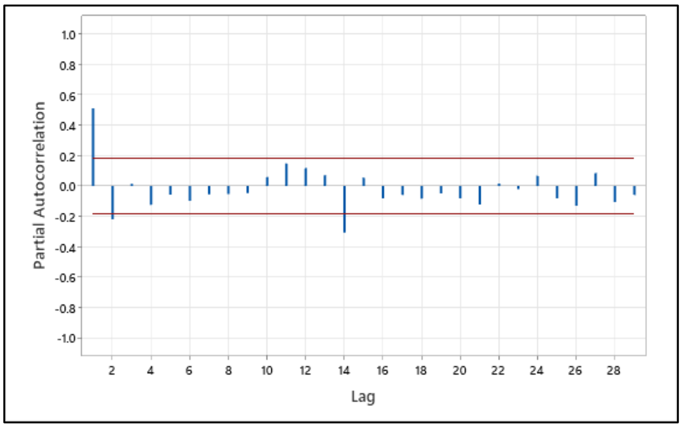

Figure 17: PACF plot for seasonal differencing of the first differenced data (Lubok Paku)

3.5 ARIMA, ARMA and SARIMA model

3.5.1 Temerloh station

SARIMA $(2,0,3)(0,1,0)^{12}$ revealed the lowest p-value which is equal to 0.001 as well as the lowest value for sum of square (SS) and mean square (MS) when compared to other models. The SS and MS of this model were 16108397 and 153317 respectively. Besides that, among the ARMA model candidates, ARMA (3,1) model demonstrated the lowest p-value which is equal to 0.003 with a lower value for sum of square (SS) and mean square (MS) when compared to other models. The SS and MS of this model were 19363798 and 172891 respectively. ARIMA (3,1,1) model outperform its candidate models with least p-value, 0.002 and lower value for SS and MS, 18423378 and 165976.

3.5.2 Lubok Paku station

SARIMA$(0,0,0)(1,1,0)^{12}$ model revealed the lowest p-value which is equal to zero as well as the lower value for sum of square (SS) and mean square (MS). The SS and MS of this model were 16186455 and 153655 respectively. SARIMA $(0,0,0)(1,1,0)^{12}$ model was consisted of non-seasonal ARIMA model (0,0,0). It was considered as a white noise model which is uncorrelated and the random variables are identically distributed with mean zero and constant variance [5], which satisfied all the criterion required for stationary condition. The ARMA model that superior to its model candidates was ARMA(1,2) model with p-value which equal to 0.017 as well as values of SS and MS of 19211144 and 170010. The ARIMA (0,1,1) model recorded the highest p-value of 0.025 when compared to ARMA and SARIMA model with values of SS and MS of 16472427 and 172214. The mentioned model would be used to fit the time series regression of stream flow discharge of Lubok Paku station from January 2010 to September 2019.

3.6 Stream flow forecasting

3.6.1 Temerloh station

Based on Table 2, SARIMA model $(2,0,3)(0,1,0)^{12}$ depicted low value of error as its overall MAPE value is insignificant (less than 20 % deviation from the actual streamflow data) and also in good fit with the observed discharge data on Temerloh station. According to Table 3, a value of MAPE which less than 20 % as a good forecasting. This implies that models with a MAPE values within 20 % of the best model enter the model portfolio. For ARIMA (3,1,1) and ARMA (3,1) models, they both demonstrated the MAPE values that more than 50 %, which classified them as in accurate forecasting as shown as Table 3. Both ARIMA and ARMA models do not consist of seasonal part and hence fail to achieve stationary for seasonality of streamflow while stationarity is the main criterion of time series model which required to be adhered. Additionally, a visual plot of the comparative forecast performance for the best-fitted model SARIMA $(2,0,3)(0,1,0)^{12}$ as shown as Figure 18. It showcased that the SARIMA generalised forecast series with slightly deviation from historical stream flow and are approximately identical as the forecast streamflow are tracking relatively closely to observed data.

Table 2: Example of presenting data using a table

Models	MAPE (%)
SARIMA $(2,0,3)(0,1,0)^{12}$	18.35
ARIMA (3,1,1)	60.16
ARMA (3,1)	50.88

Table 3: Example of presenting data using a table [6]

MAPE	Evaluation
MAPE ≤ 10%	High accuracy forecasting
10% < MAPE ≤ 20%	Good forecasting
20% < MAPE ≤ 50%	Reasonable forecasting
MAPE > 50%	Inaccurate forecasting

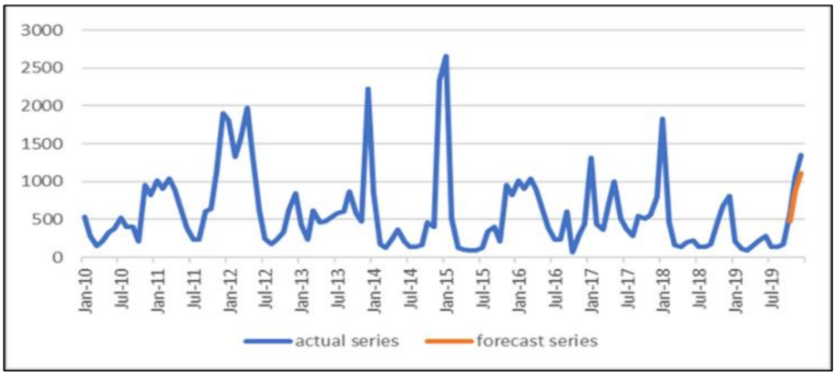

Figure 18: Representative plots of observed and predicted stream flow (Temerloh station)

3.6.2 Lubok Paku station

Based on Table 4, SARIMA model $(0,0,0)(1,1,0)^{12}$ exhibited lowest value of error as its MAPE value less than 10 % which indicated the forecast series maintained a good concurrence with observed data on Lubok Paku. The MAPE that less than 10% is acts as an indicative of very precise prediction. ARIMA (0,1,1) model demonstrated the highest MAPE value of 46.26 % whereas the MAPE of ARMA (1,2) model is 22.49 %. Similarity, this might due to the over-differencing of model and hence lower the precision of forecast model as Temerloh station. This can be supported remarkably by the best fitted model SARIMA model with MAPE value of $(0,0,0)(1,1,0)^{12}$ which did not exhibit differencing term in its non-seasonal part as well. Both ARIMA and ARMA models exhibited MAPE values within the range from 20 % to 50% and hence they were categorised as reasonable forecasting. A visual plot of the comparative forecast performance for the best-fitted model SARIMA model $(0,0,0)(1,1,0)^{12}$ was presented in Figure 19. It had showcased that the SARIMA and its deviation from historical actual stream flow are identical as the forecast streamflow are tracking relatively closely to observed data.

Table 4: Example of presenting data using a table

Models	MAPE (%)
SARIMA $(0,0,0)(1,1,0)^{12}$	6.54
ARIMA (0,1,1)	46.26
ARMA (1,2)	22.49

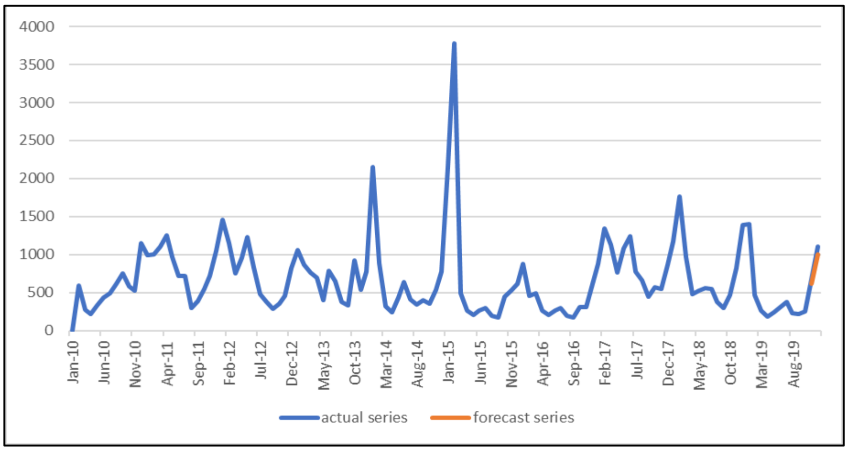

Figure 19: Representative plots of observed and predicted stream flow (Lubok Paku station)

4. Conclusion

In conclusion, SARIMA models outperformed the remaining models which indicated the least MAPE values of 18.35 % and 6.54 % for Temerloh station and Lubok Paku station respectively. Both SARIMA models generate good forecasting with MAPE lower than 20 %. However, the ARMA models obtain a lower MAPE values than ARIMA models for both stations. This might imply the over-differencing of models. The application of time series model should be verified by performing the modelling at varying time span (daily or annually). Besides, multivariate time series modelling which input with exogenous data (temperature and climate variables) also suggested as an attempt to improve accuracy of forecast. Besides, a hybrid forecasting that combines data-driven and process driven models is proposed to enhance the accuracy of SARIMA models too. For this study, there are some recommendations to improve the results of forecasted value of stream flow for River Pahang. The application of time series model should be verified by performing the modelling at varying time span such as daily or annually. Besides, multivariate time series modelling which input with various exogenous data including temperature and climate variables also proposed to be performed as an attempt to enhance the precision of forecast values. Notwithstanding, the results demonstrated that the accuracy of obtained forecast series is acceptable as the MAPE values of SARIMA model are less than 20%, a hybrid forecasting that comprised of data-driven and process driven parameters to strengthen the efficiency of SARIMA models.

Acknowledgement

The authors would also like to thank the Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia for its support.

References

[1] Frausto-Solis, Juan & Pita, Esmeralda & Lagunas, Javier. (2008). Short-term streamflow forecasting: ARIMA vs neural networks, 402.

[2] Nochai, Rangsan & Nochai, Titida. (2006). ARIMA Model Forecasting Oil Palm Price. Ong, Chorng-Shyong & Huang, Jih-Jeng & Tzeng, Gwo-Hshiung. (2005). Model identification of ARIMA family using genetic algorithm. Applied Mathematics and Computation, 164, 885-912.

[3] Alias, Nurul & Chamhuri, Siwar & Ismail, Mohd Khairi & Mohd Idris, Nor. (2018). Flood Disaster Management in Sungai Pahang Basin: Case of Temerloh: Case Studies in Asia. Community, Environment and Disaster Risk Management] Improving Flood Management, Prediction and Monitoring Volume 20, 91–102.

[4] Temerloh Municipal Council. (2016). Slide presentation: Briefing on preparation for floods disaster Temerloh district 2015/2016

[5] Moffat, I. and Akpan, E. (2019) White Noise Analysis: A Measure of Time Series Model Adequacy. Applied Mathematics, 10, 989-1003.

[6] Lewis, C.D. (1982). Industrial and business forecasting methods. London: Butterworths. p.40.