Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 18, 2023, page 1 – 10

Rainfall Modelling and Forecasting Using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing in Pahang and Terengganu, Malaysia

Nur Hafizzah Samsulaiman, Norazlina Ismail*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: i-norazlina@utm.my

Abstract
Forecasting rainfall is crucial when making important decisions and conducting strategic planning. The management of water-related difficulties, such as those caused by extreme rainfall conditions like floods and droughts, among other challenges, is influenced by the ability to statistically anticipate and forecast rainfall (Otieno et al., 2022). The purpose of this study is to identify the best performing forecasting model by predicting the amount of rainfall using the most appropriate forecasting model. Two stations from Terengganu and Pahang were chosen as the research area and the monthly rainfall data, with the range from January 1980 to December 2021, were obtained from Department of Irrigation and Drainage Malaysia (DID) and the Department of Meteorology Malaysia. Terengganu's rainfall pattern exhibits a trend and seasonality for both stations, but Pahang only has one station with trend. In this study, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Holt-Winters' Exponential Smoothing method were proposed to forecast the rainfall data. The performance of the models had been evaluated based on performance indicator which is Root Mean Square Error (RMSE). It has been proven that the Holt-Winters' Exponential Smoothing model gives a more accurate result and can be used to predict future rainfall.

Keywords: Rainfall; SARIMA; Holt-Winters; Time Series Forecasting; Exponential Smoothing

1. Introduction

Malaysia is known as a country having a hot climate all year round because of its proximity to the tropical. Peninsular Malaysia, which is in the north between Singapore and Thailand, and the two Borneo states of Sabah and Sarawak, which are in the south, are the two distinct regions of Malaysia. Malaysian Metrological Department, Ministry of Environment and Water reports that there are four seasons can be distinguished. For northeast monsoon, it usually will occur on early November until the end of March. For southwest monsoon and inter-monsoon, it usually established in the latter half of May or early June and ends in September and starting in late March to early May and October to mid-November respectively.

Monsoon, mud, and flash floods are three different types of flood disasters that can happen in Malaysia (Wong, 2020). A major part of Malaysia, including Johor, Melaka, Pahang, Kelantan, Terengganu, Sabah, and Sarawak, is typically affected by monsoon flooding. The first wave of flooding in Terengganu happened in November 2014 after a week of heavy rain in Kuala Terengganu and the surrounding area caused the soil's surface to get saturated and the groundwater level to rise. Because of this, all low-lying regions were flooded. The road acts as a temporary drainage system and causes flooding in the development area when the ground water level rises and there was no effective drainage system to direct water to the sea. At the end of 2014, there will then be a second wave of flooding. Kemaman, Dungun, Kuala Terengganu, Hulu Terengganu, Besut, and Marang are the Terengganu

districts that have been most severely impacted (Buslima et al., 2018).

Time series forecasting occurs when you make scientific predictions based on historical time stamped data, especially for financial trends or coming weather. Time series are important for predicting since they must be considered during the decision-making process. It applies in many areas, but it's particularly important in weather forecasting.

Thus, this study is conducted to (1) model and analyse rainfall data using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Holt-Winters' method, (2) to propose the best forecasting method between best fitting model using Root Mean Square Error (RMSE) and (3) to forecast the changes of monthly time series rainfall pattern by forecasting using the best model chosen.

2. Materials and methods

2.1 Data Collection

In this research, secondary data has been used for the data collection from Department of Irrigation and Drainage Malaysia (DID) and the Department of Meteorology Malaysia. The rainfall data used in this study consists of daily rainfall amounts which are then categorized as monthly rainfall data for 40 years from 1980 to 2021. For the evaluation process, four rainfall stations from the state of Pahang and Terengganu have been chosen.

2.2 Seasonal Autoregressive Integrated Moving Average (SARIMA)

A seasonal ARIMA model is written as SARIMA $(p,d,q)(P,D,Q)m$, where m is the number of periods in each season. The lowercase notation part stands for the non-seasonal part of the model while the uppercase letters P, D, and Q stand for the seasonal ARIMA model's respective moving average (MA), autoregressive (AR), and differencing (I) factors (Dimri et al., 2020).

The general form of the model ARIMA (p, d, q) is given by

$$\left(1-\phi_1 B-\cdots-\phi_p B^p\right)(1-B)^d y_t=\delta+(1-\theta_1 B-\cdots-\theta_q B^q)e_t \tag{1}$$

Equation (1) can be written as

$$\phi_p(B)(1-B)^d y_t=\delta+\theta_q(\mathrm{B})e_t \tag{2}$$

The ARIMA model can be extended to handle the seasonal components of data series. The seasonal ARIMA model, SARIMA $(p,d,q)\ (P,D,Q)_s$ can be defined as

$$\left(1-\phi_1 B-\phi_1 B^2-\cdots-\phi_p B^p\right)(1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps})(1-B^s)^D(1-B)^d x_t=\delta+(1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)(1-\Theta_1 B^2-\Theta_2 B^{2s}-\cdots-\Theta_Q B^{Qs})\varepsilon_t \tag{3}$$

Equation (3) can be written as

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t=\delta+\theta_q(\mathrm{B})\Theta_Q(B^s)\varepsilon_t \tag{4}$$

where

$\phi(B)$	:	Autoregressive component of order p, $AR\ (p)$
$\theta(B)$	:	Moving average component of order q, $MA\ (q)$
$\Phi_P(B^s)$	:	Seasonal autoregressive component of order $P, SAR\ (P)$
$\Theta_Q(B^s)$	:	Seasonal moving average component of order $Q, SMA\ (Q)$

$(1-B)^d$: Difference component of order $d, I(d)$
$(1-B^s)^D$: Seasonal difference component of order $D, I(D)$
S : Seasonal period

2.3 Additive Holt-Winters' Method

Exponential smoothing is a method which plays an important role in forecasting by accounting for fluctuations in the data. It is simple to understand and use to make decisions depending on the user's past assumptions, such as seasonality. Holt-Winters' models can be divided into two categories based on the seasonality. In this research, the additive Holt-Winters' was used.

The Additive Holt-Winters approach takes seasonal fluctuations into account regardless of the level of the series; there is no pattern or hint that the seasonal pattern is affected by the amount of data. These are the equations applied in the additive model:

Level : $L_t = \alpha(y_t - S_{t-S}) + (1-\alpha)(L_{t-1} + b_{t-1})$ (5)

Trend : $b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$ (6)

Seasonal : $S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-S}$ (7)

Forecast : $F_{t+m} = L_t + b_t m + S_{t+m-s}$ (8)

where,

- α, β and γ are smoothing constants where α is for level, β is for trend and γ is for seasonal between 0 and 1
- s indicated the number of seasons for example $s = 12$ for monthly data
- L_t is level series, b_t is trend estimate and S_t is seasonality factor

3. Error of Measurement

3.1 Root Mean Square Error (RMSE)

RMSE represents the model's overall fit to the data, or how well the observed data points match the values predicted by the model. According to the following formula, a good performance model should include the lowest possible value where F_t is forecast value, y_t is actual value in time t and n is size of sample. The error measurement is calculated using the formula as shown below:

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(y_t - F_t)^2}{n}} \quad (9)$$

4. Results and discussion

In this section, there are two methods will be used to analyse the data set given. Firstly, the Box-Jenkins Algorithm is used to generate SARIMA models. Model identification, parameter estimate, diagnostic evaluation, and forecasting are the four processes involved. Then, the Holt-Winters' method will be carried out to be the comparison with SARIMA to find the best model using error measurements in time series analysis.

4.1 Time Series Plot

The time series plot for four stations are presented in Figure 1, Figure 2, Figure 3 and Figure 4 as shown below. The stations in Terengganu state which are Figure 1 and Figure 2, clearly visualize the trend existence in this dataset. Meanwhile, the data obtained in Sg. Lembing PCCL Mill station of Pahang

state, Figure 3 also shows trend, yet it is hardly shown the seasonal factor. Therefore, we assume the data contain trend and seasonality which leads the data to be non-stationary. Nevertheless, Kg. Sg. Yap in Figure 4, the data indicate that no trend presence.

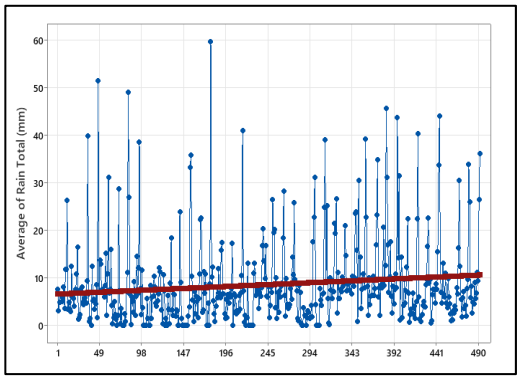

Figure 1: Kg. Dura

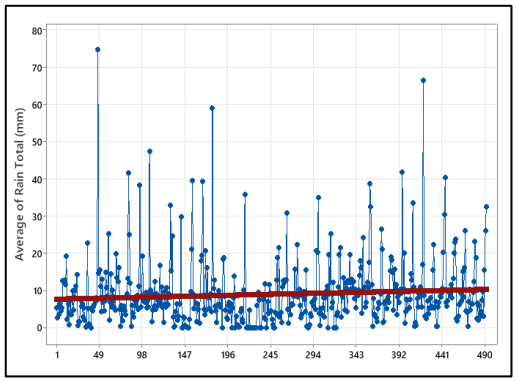

Figure 2: Al-Muktafi Billah Shah

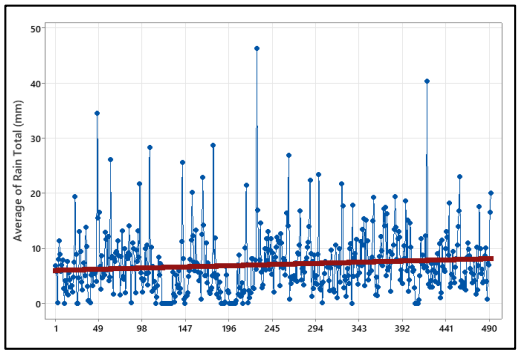

Figure 3: Sg. Lembing PCCL Mill

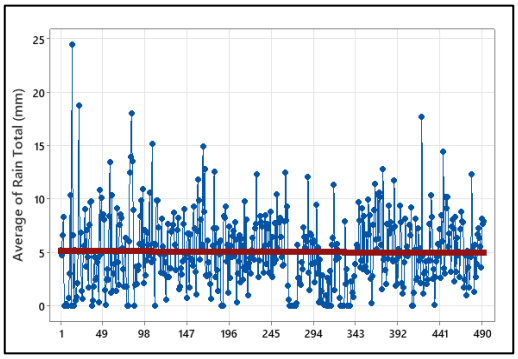

Figure 4: Kg. Sg. Yap

4.2 SARIMA Model

A non-stationary monthly rainfall time series is converted to a stationary time series as the initial step in the development of the SARIMA model. The correlogram plot which is ACF and PACF for the data set on rainfall is shown below. It should be noted that even if the best SARIMA model is found, alternative SARIMA models with values for the parameters AR and MA that are less than those of the evaluated SARIMA models may still be taken into consideration (Ali, 2013). After estimating the model's parameters, AIC and BIC were chosen to test the optimal model.

Figure 5 and Figure 6 show finalized ACF and PACF plot of two stations at Terengganu which are Kg. Dura and Al-Muktafi Billah Shah stations. The raw dataset of ACF revealed that the data was non-stationary, so the differencing is needed to find the suitable models. The data in the area have trend and seasonality existence so we need to difference twice. These figures are the outcomes after differencing for trend, d=1 and seasonality, D=1.

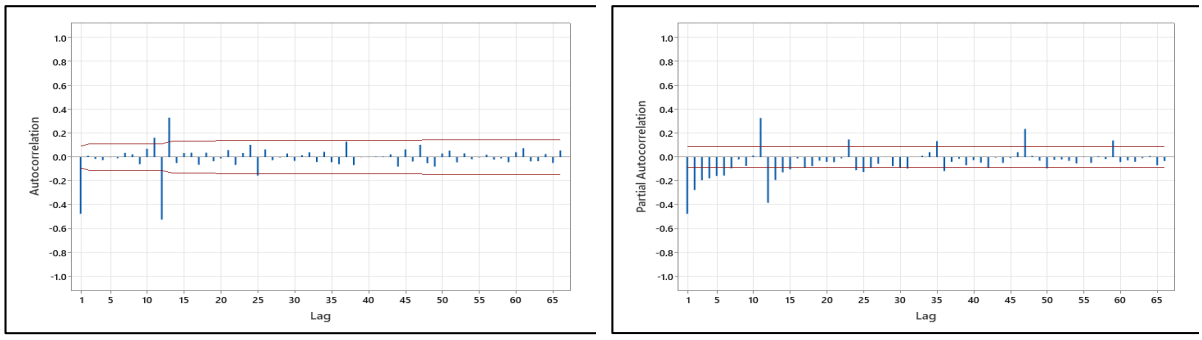

Figure 5: ACF and PACF Plot of differenced for Kg. Dura Station

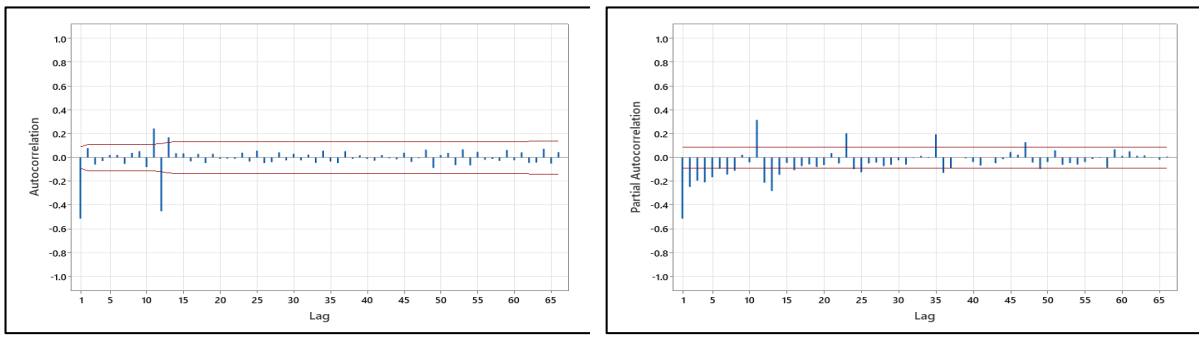

Figure 6: ACF and PACF Plot of differenced for Al-Muktafi Billah Shah Station

The finished ACF and PACF plot of the Sg. Lembing PCCL Mill and Kg. Sg. Yap stations in Pahang are shown in Figures 7 and Figure 8. Since the ACF raw dataset showed that the data was non-stationary, differencing is required to identify the most appropriate models. The data in the area only have trend existence but no seasonality shown so we need to difference once. These results were modified for trend, d=1 to produce these figures.

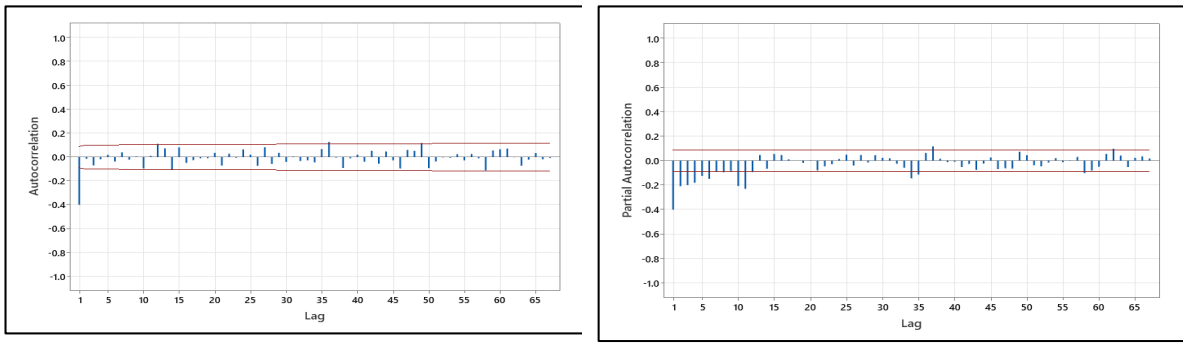

Figure 7: ACF and PACF Plot of differenced for Sg. Lembing PCCL Mill Station

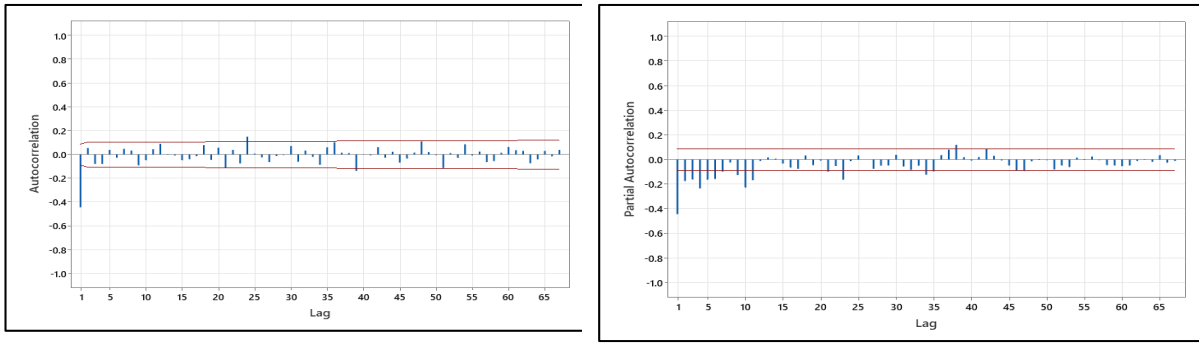

Figure 8: ACF and PACF Plot of differenced for Kg. Sg. Yap Station

Several models were suggested based on the outcomes of the model identification charts, and the values for AIC and BIC were determined to penalize the models and prevent over-fitting.

Table 1: AIC and BIC Values

Station	SARIMA Model	AIC	BIC	Best Model
Kg. Dura	$(1,1,1)(1,1,1)^{12}$	3.8972	3.9320	$(0,1,1)(0,1,1)^{12}$
	$(0,1,1)(0,1,1)^{12}$	3.8932	3.9106	
Al-Muktafi Billah Shah	$(0,1,1)(0,1,1)^{12}$	3.9987	4.0161	$(0,1,1)(0,1,1)^{12}$
	$(1,1,1)(0,1,1)^{12}$	4.0023	4.0284	
Sg. Lembing PCCL Mill	$(1,1,1)(1,0,1)^{12}$	3.2917	3.3266	$(0,1,1)(1,0,1)^{12}$
	$(0,1,1)(1,0,1)^{12}$	3.2941	3.3202	
Kg. Sg. Yap	$(0,1,1)(0,0,1)^{12}$	2.4911	2.5085	$(0,1,1)(0,0,1)^{12}$
	$(1,1,0)(0,0,2)^{12}$	2.6508	2.6769	

For Kg. Dura station, we found that the best model was SARIMA $(0,1,1)(0,1,1)^{12}$ while the best SARIMA model for Al-Muktafi Billah Shah station was SARIMA $(0,1,1)(0,1,1)^{12}$. Other than that, the model SARIMA $(0,1,1)(1,0,1)^{12}$ and SARIMA $(0,1,1)(0,01)^{12}$ is the most significant for Sg. Lembing PCCL Mill and Kg. Sg. Yap station respectively.

4.3 Additive Holt-Winters' Method

In the time series plot, there are occasionally many fluctuations, although they appeared to have a constant amplitude throughout time. Thus, the Additive Holt-Winters approach is chosen for use. Using the solver tool in Microsoft Excel, the optimal set of parameters was determined. Both stations in Terengganu which in Figure 9 exhibit the same value for $\alpha = 0.9938, \beta = 0.0126$ and $\gamma = 0.1870$.

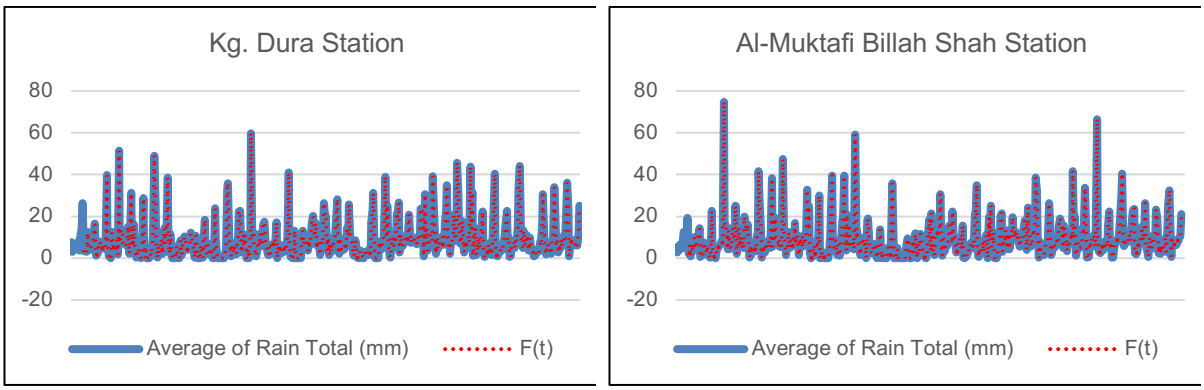

Figure 9: Plots of actual value and forecast value for Terengganu's station

Based on Figure 10 below, the parameters value for Sg. Lembing PCCL Mill station using Excel's Solver is $\alpha = 0.9999, \beta = 0.0001$ and $\gamma = 0.1090$ meanwhile for Kg. Sg. Yap station is $\alpha = 0.9938, \beta = 0.0126$ and $\gamma = 0.1870$.

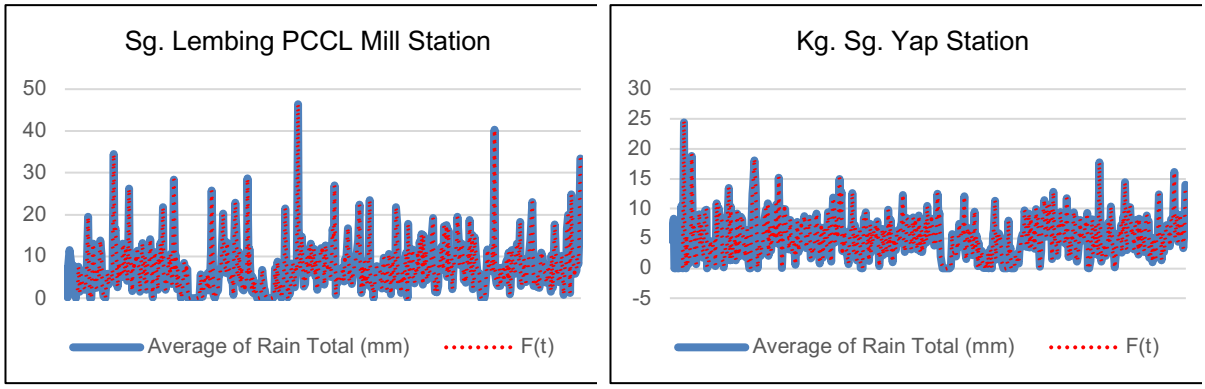

Figure 10: Plots of actual value and forecast value for Pahang's station

As can be seen in Figure 9 and Figure 10, it was noticed that the original data (blue colour in line) fits into and follows the trend of the forecast data (red dot colour in line). In light of the results, it may be concluded that the Holt-Winters model is a suitable one.

4.4 Model Evaluation

Table 2: Performance comparison based on RMSE value

Station	RMSE	
	SARIMA	**Holt-Winters'**
Kg. Dura	6.9755	0.0751
Al-Muktafi Billah Shah	7.3536	0.0859
Sg. Lembing PCCL Mill	5.0957	0.0077
Kg. Sg. Yap	3.4178	0.0936

The prediction model's accuracy increases as RMSE levels decrease. Due to its lower metric readings than the SARIMA model, the Holt-Winters model performed better in terms of performance.

Conclusion

In this research, Holt-Winters' and SARIMA models were proposed and used to forecast the rainfall data. Based on the Root Mean Square Error (RMSE), the effectiveness of each model has been assessed. The data have been evaluated and the results show that the Holt-Winters' model has smaller values of RMSE than the SARIMA model, indicating a higher level of accuracy. As a result, forecasting by additive Holt-Winters' method has been carried out for the year 2022. For both stations Kg. Dura and Al-Muktafi Billah Shah, April is the driest month while December is the wettest. Besides, December is the wettest month and July is the driest month in Sg. Lembing PCCL Mill station. The Kg. Sg. Yap station experiences the least amount of rainfall in December meanwhile February is the driest month. For future research, another forecasting method is needed such as Artificial Neural Network (ANN) because it was a popular choice among researchers and also able to get more accurate forecasting.

Acknowledgement

The researcher wishes to thank all those who have supported the research and the Department of Irrigation and Drainage Malaysia (DID) and the Department of Meteorology Malaysia for providing the rainfall data. Special thanks are also dedicated to Universiti Teknologi Malaysia for having this research.

References

[1] Afrifa-Yamoah, E., Krzyszczak, J., Saeed, B. I. I., & Karim, A. (2016). *Sarima Modelling and*

Forecasting of Monthly Rainfall in the Brong Ahafo Region of Ghana. https://doi.org/10.5923/j.env.20160601.01

[2] Ali, S. M. (2013). Time Series Analysis of Baghdad Rainfall Using ARIMA Method. In *Ali Iraqi Journal of Science* (Vol. 54).

[3] Boshnakov, G. N. (2016). Introduction to Time Series Analysis and Forecasting, 2nd Edition, Wiley Series in Probability and Statistics, by Douglas C.Montgomery, Cheryl L.Jennings and MuratKulahci (eds). Published by John Wiley and Sons, Hoboken, NJ, USA, 2015. Total number of pag. *Journal of Time Series Analysis*, *37*(6), 864. https://doi.org/10.1111/jtsa.12203

[4] Buslima, F. S., Omar, R. C., Jamaluddin, T. A., & Taha, H. (2018). Flood and flash flood geo-hazards in Malaysia. *International Journal of Engineering and Technology(UAE)*, *7*(4), 760–764. https://doi.org/10.14419/ijet.v7i4.35.23103

[5] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[6] Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, *129*(1). https://doi.org/10.1007/s12040-020-01408-x

[7] Harfizi, N. A., Elianie, N. I., Farhanah, N., Bakar, A., & Sahat, S. (2022). Trend Analysis of Rainfall Characteristics in Kota Tinggi. *Multidisciplinary Applied Research and Innovation*, *3*(1), 282–288. https://doi.org/10.30880/mari.2022.03.01.033

[8] Joe, T. T., Arina, N., Kamisan, B., & Binti Kamisan, B. (2022). *Rainfall Forecasting with Time Series Model in Senai, Johor* (Vol. 11).

[9] Kokilavani, S., Pangayarselvi, R., Ramanathan, S. P., Dheebakaran, Ga., Sathyamoorthy, N. K., Maragatham, N., & Gowtham, R. (2020). SARIMA Modelling and Forecasting of Monthly Rainfall Patterns for Coimbatore, Tamil Nadu, India. *Current Journal of Applied Science and Technology*, 69–76. https://doi.org/10.9734/cjast/2020/v39i830594

[10] Lakshminarayana, S. V. (2020). Rainfall Forecasting using Artificial Neural Networks (ANNs): A Comprehensive Literature Review. *Indian Journal of Pure & Applied Biosciences*, *8*(4), 589–599. https://doi.org/10.18782/2582-2845.8250

[11] Mohamad Fudzi, F., Md Yusof, Z., & Misiran, M. (2021). Rainfall Forecasting With Time Series Model In Alor Setar, Kedah. *Universiti Malaysia Terengganu Journal of Undergraduate Research*, *3*(1), 37–44. https://doi.org/10.46754/umtjur.2021.01.005

[12] Mohd Firdaus Azis, T., Wai Sum, L., Aizzat Adnan, A., Sapiri, H., & Misiran, M. (2019). The Use of Correspondence Analysis on the Visualization of Locality and Seasonal Behaviour on the Flood Pattern in Malaysia. *Journal of Advanced Research in Applied Sciences and Engineering Technology Journal Homepage*, *15*, 1–7. www.akademiabaru.com/araset.html

[13] Otieno, V. A., & Wanyonyi, S. W. (2022). Modeling and Forecasting of Rainfall Trends based on Historical Data in Bungoma County, Western Kenya using Holt Winters Method. *Asian Journal of Probability and Statistics*, 38–44. https://doi.org/10.9734/ajpas/2022/v17i430431

[14] Pahang Media. (2022). Gelinciran Tanah Dan Aliran Puing Punca Utama Bencana Banjir 2021. *Pahang Media*. https://pahangmedia.my/gelinciran-tanah-dan-aliran-puing-punca-utama-bencana-banjir-2021/

[15] Panda, A., & Sahu, N. (2019). Trend analysis of seasonal rainfall and temperature pattern in Kalahandi, Bolangir and Koraput districts of Odisha, India. *Atmospheric Science Letters*, *20*(10). https://doi.org/10.1002/asl.932

[16] Papalaskaris, T., Panagiotidis, T., & Pantrakis, A. (2016). Stochastic Monthly Rainfall Time Series Analysis, Modeling and Forecasting in Kavala City, Greece, North-Eastern Mediterranean Basin. *Procedia Engineering*, *162*, 254–263. https://doi.org/10.1016/j.proeng.2016.11.054

[17] Pertiwi, D. D. (2020). Applied Exponential Smoothing Holt-Winter Method for Predict Rainfall in Mataram City. *Journal of Intelligent Computing and Health Informatics*, *1*(2), 45. https://doi.org/10.26714/jichi.v1i2.6330

[18] Pongdatu, G. A. N., & Putra, Y. H. (2018). Seasonal Time Series Forecasting using SARIMA and Holt Winter's Exponential Smoothing. *IOP Conference Series: Materials Science and*

Engineering, *407*(1). https://doi.org/10.1088/1757-899X/407/1/012153

[19] Puah, Y. J., Huang, Y. F., Chua, K. C., & Lee, T. S. (2016). *River catchment rainfall series analysis using additive Holt-Winters method*.

[20] Sinay, L. J., & Kembauw, E. (2021). Monthly rainfall components in ambon city: Evidence from the serious time analysis. *IOP Conference Series: Earth and Environmental Science*, *755*(1). https://doi.org/10.1088/1755-1315/755/1/012079

[21] Swain, S., Nandi, S., & Patel, P. (2018). Development of an ARIMA model for monthly rainfall forecasting over Khordha District, Odisha, India. *Advances in Intelligent Systems and Computing*, *708*, 325–331. https://doi.org/10.1007/978-981-10-8636-6_34

[22] Tan, K. C. (2018). Trends of rainfall regime in Peninsular Malaysia during northeast and southwest monsoons. *Journal of Physics: Conference Series*, *995*(1). https://doi.org/10.1088/1742-6596/995/1/012122

[23] Widodo*, A., Handoyo, S., Ariyanto, R., & Marji. (2020). The Data-Driven Fuzzy System with Fuzzy Subtractive Clustering for Time Series Modeling. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 3357–3362. https://doi.org/10.35940/ijitee.C9039.019320

[24] Wong, W. M. (2020). Flood Prediction using ARIMA Model in Sungai Melaka, Malaysia. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5287–5295. https://doi.org/10.30534/ijatcse/2020/160942020