JOURNAL OF COMPUTATIONAL INNOVATION AND ANALYTICS
https://e-journal.uum.edu.my/index.php/jcia

How to cite this article:
Tee H.Y. & Mansor R. (2024). Forecasting rainfall volume in Selangor with combined ARIMA model. *Journal of Computational Innovation and Analytics*, *3*(1), 83-103. https://doi.org/10.32890/jcia2024.3.1.5

FORECASTING RAINFALL VOLUME IN SELANGOR WITH A COMBINED ARIMA MODEL

[1]Tee Huey Yin & [2]Rosnalini Mansor
School of Quantitative Sciences,
Universiti Utara Malaysia, Malaysia

[1]Corresponding author: tee_huey_yin@uumkl.uum.edu.my

Received: 10/4/2023 Revised: 5/10/2023 Accepted: 31/10/2023 Published: 31/1/2024

ABSTRACT

Flash flood is the most hazardous type of flooding, mainly caused by extensive rainfall. It also can cause significant harm to a community's economy, ecology, and society without warning at an irrational pace. Therefore, this study was conducted to detect the time series element within the rainfall data, select the optimal model, and make predictions about the volume of rainfall in Selangor. A variety of univariate time series models were utilized, including the naïve model, decomposition model, Autoregressive Integrated Moving Average (ARIMA) model, exponential models, and combined models. Historical monthly rainfall data collected from Petaling station and Subang station from 2018 to 2022 were used to estimate the parameters of the models, and the model was evaluated for the smallest error of measurements. Previous research mostly focused on complex methodologies for forecasting rainfall. However, this research aimed to identify a simple tool for fast prediction of rainfall. The results showed that the combination of the ARIMA (2,0,3) model from Petaling Station and the ARIMA (4,0,4) model from Subang station were able to capture the trends

and seasons in the time series with the lowest error of measurement on short-term predictions of rainfall volume. Furthermore, the study delves into the concept of combined time series models, which are blended using weighted performance measures to enhance prediction accuracy further. The research acknowledges certain limitations of univariate time series models, notably their inability to account for intricate interactions among environmental variables and potential long-term trends, such as those stemming from climate change. Overall, the study explores the potential of combining models to refine predictions for forecasting rainfall volume in Klang Valley.

Keywords: ARIMA, Combined Models, Forecasting, Rainfall Volume, Time Series Model.

INTRODUCTION

In recent decades, global climate change has significantly impacted the Earth's ecological system, causing long-term shifts in temperature and variable weather patterns. According to the World Meteorological Organization (WMO) in World Bank Group (2021), high-impact events such as flooding have been recorded across the world. Severe flooding occurred in Africa, Sudan, Kenya, India, and Southeast Asia due to heavy monsoon rainfall, resulting in considerable losses to these nations. The impact of climate change has been substantial, and it had devastating consequences on the affected countries' infrastructure, society, and economy. Moreover, rapid urbanization and land development have resulted in significant changes to the land structure, potentially leading to ecological and environmental problems (Reza, 2016). When the ecological system is disrupted, the outflow of water cannot be dispersed, causing more severe flooding in the country.

The effects of climate change and rapid land development had significant consequences on Malaysia, a developing nation located in a hot and humid tropical climate with heavy tropical rainfall. Additionally, the country is highly reliant on the monsoon seasons for its livelihood. Malaysia experiences two periods of monsoons, namely the Northeast Monsoon (NEM) season from early November to March and the Southwest Monsoon (SWM) season from late May to September (World Bank Group, 2021). During these monsoon seasons, Malaysia experiences long-haul rainfall of approximately 200 mm in June and July and 350 mm in November and December.

The precipitation in Malaysia has increased year by year during the monsoon season, from 3053.99 mm in 2020 to 3297.34 mm in 2021, resulting in frequent flooding and flash floods (World Bank Group, 2021). The Department of Statistics of Malaysia reported that the recent flood that occurred in December 2021 resulted in 50 fatalities, the evacuation of about 400,000 people, and an estimated financial loss of RM6.1 billion (Department of Statistics Malaysia, 2022).

Extreme rainfall leading to severe flood events can significantly impact a nation's society and economy. Vehicle owners may incur significant repair expenses, while infrastructure such as roads and buildings can suffer damages. Schools may also have to be utilized as evacuation centers, disturbing the usual schedules of students and teachers. The residents in flood-prone regions can experience ongoing anxiety and trauma, while businesses may face losses of products and disturbances to their services due to power and water supply outages. The cost of compensating for damages and repairing infrastructure is typically high, leading to losses for the country. Due to the vast increase in flood frequency, this study aims to forecast the rainfall volume in Selangor to mitigate the occurrence of flash floods. The paper proposes to identify the time series components of the rainfall volume, later to determine the most suitable time series forecasting model for the rainfall volume. Last but not least, the paper forecasts the rainfall volume in the coming year with the identified model.

RELATED WORK

The study highlighted the importance of understanding the rainfall patterns in Malaysia, which are the trends of rainfall and seasonal monsoons during a year. The NEM and SWM are the two primary seasons for rainfall in Malaysia, which lasts from November to March and May to September, respectively. The rainfall volume typically ranges between 2000 mm to 3000 mm annually, with the northwestern areas having the highest mean rainfall during the SWM season. Recent studies suggest that Malaysia has observed a rise in both the frequency and severity of heavy rainfall occurrences, increasing the prevalence of flash floods in densely populated regions like Petaling Jaya and Subang (Diya et al., 2014; Suparta et al., 2015; Syafrina et al., 2015). The Titiwangsa Range is one potential factor that affects rainfall patterns in Malaysia by blocking northeasterly winds. However, recent rainfall trend analysis shows a rise in the number of days having rainfall and extremely heavy rainfall in Klang Valley, which requires attention as it is a heavily populated area.

In time series analysis and forecasting, several research were reviewed on predicting rainfall volumes to mitigate flood occurrence. In Adnan et al.'s (2012) research, this paper proposed an Artificial Neural Network (ANN) model to predict the flood water level. In Hong and Hong (2016) 's study, it has evaluated the use of Multi-Layer Perceptron (MLP) neural network models to forecast water levels of a gauging station located at the Kuala Lumpur city center in Malaysia using records of upstream multiple stations. Moreover, Mishra et al. (2018), the research had also proposed an Artificial Neural Network (ANN) technique to develop one to two- months ahead forecasting of rainfall in Northern India. Besides, in Mustapha and Ismail's (2021) research, they compared the use of two models of univariate time-series analysis, the Autoregressive Integrated Moving Average (ARIMA) and SARIMA models, which were applied to model and forecast the monthly time series rainfall in Kelantan, Malaysia. It was proven that the SARIMA model was a good model for forecasting monthly rainfall time series, resulting in a lower measurement error.

Understanding the complexity of multivariate analysis and selecting the most appropriate univariate time series model is crucial to ensure accurate predictions. However, with a small or moderate number of observations, models that are close to each other can be challenging to distinguish, and the model selection criterion values can be similar. Therefore, choosing the model with the lowest criterion value may not always be reliable, and small shifts in the data can influence the choice of a different model. This instability in model selection can lead to high variability in the forecast using the selected algorithm. To overcome this issue, this paper studied the use of univariate models and proposed alternative methods to combine forecasting models, including model averaging, ensemble methods, and combination of residuals.

Research has shown that combined models can improve the accuracy and stability of forecasts. For example, the ARIMA-AR model combines two ARIMA models with different orders to improve forecast accuracy, with the forecasts from the two models being combined using a weighted average determined by a genetic algorithm (Zhang et al., 2019). Similarly, the combined ARIMA time series model with the ARIMA-ARIMA model proposed by Tunc et al. (2016) combines two ARIMA models to capture both long-term trends and short-term fluctuations, with the forecasts from both models being combined using a weighted average approach. This paper considers the methodology of combining models by assigning

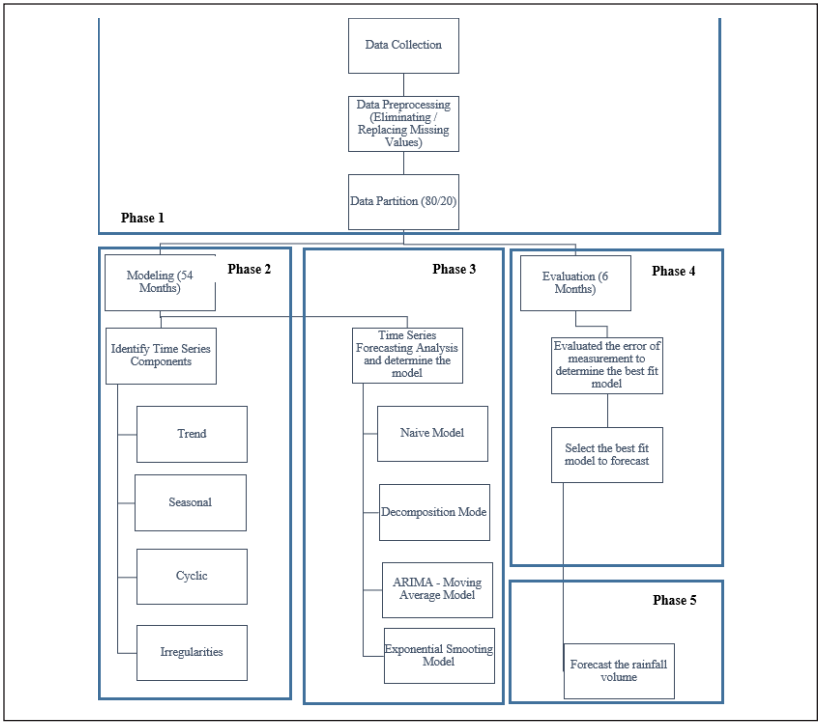
weights to the selected models, which can effectively overcome the instability of model choosing in time series analysis and forecasting.

METHODOLOGY

Two rainfall stations, Subang station (ID 48647) and Petaling station (ID 48648), located in the Klang Valley of Peninsular Malaysia, were chosen as the research area based on the completeness of the data and the length of records. The study consulted the Department of Irrigation and Drainage for secondary data and collected monthly rainfall data for five years from 2018 to 2022. They are partitioned for the modeling and evaluation process using the 80:20 rule. Therefore, the first 54 months (Jan 2018 – Jun 2022) were used for modeling, and 6 months (Jul 2022 – Dec 2022) were used for testing the error of measurements.

Figure 1

Research Framework of this Study on Modeling, Evaluating, and Forecasting Rainfall Data



This study continued a time series analysis to comprehend how rainfall patterns evolve over time and evaluated various time series statistical-based modeling theories, such as the naïve model, decomposition model, moving average, and exponential smoothing. It followed five phases of the study: preprocessing and partitioning the data, identifying the components of the data, listing down potential time-series techniques, modeling and evaluating the data, and forecasting using the best possible model. The framework of this study and the summary of the sequence are shown in Figure 1.

Phase 1: Partitioning the Data

In this phase, the data is partitioned into two sets, with one set used for modeling and the other for evaluating the forecasting models. The data is also preprocessed to remove any outliers and ensure data continuity. One of the types of data with missing values, Missing Completely at Random (MCAR), is normally applied in realistic situations (Little and Rubin, 2002). Since the data in a specific area have no bearing on the occurrence of missing rainfall datasets in an area, hydrological data, particularly in the case of missing rainfall datasets, is classified as MCAR (Shaharudin et al., 2020). When dealing with a continuous value under MCAR conditions, the Last Observation Carried Forward (LOCF) method is frequently used (Houck et al., 2004). After preprocessing the data, data partitioning follows the 80/20 rule, where 80% of the data is used for modeling and 20% for evaluation.

Phase 2: Identifying the Components of Data

This phase involves identifying the four components of time series data, namely the seasonal component (S), trend component (T), cyclical component (C), and irregular component (I). The seasonal component is affected by the monsoon season, while the cyclical component is influenced by fluctuations without a fixed period. The trend component is captured over long periods, and the irregular component is a variable that cannot be explained seasonally or cyclically. The components are identified using different techniques, such as Sequence Chart, Autocorrelation Function, and Kruskal Wallis test for seasonality.

Phase 3: Listing Potential Time-series Techniques

Several time-series techniques can be used based on the components identified in Phase 2. The Naïve model, Decomposition model,

ARIMA model, and Exponential Smoothing model are the techniques that will be modeled in this research.

Naïve Model

This method represents a simple forecasting model that requires the least work effort and data manipulation. The value of the naive forecast is set based on the value of the last observation. However, in this project study, we were forecasting using a naïve model but using the strategy of adding a variance value depending on the components of the data, as shown in Equation (1).

$$\hat{Y}_{t+1} = Y_{t-11} \frac{Y_t - Y_{t-12}}{12}. \quad (1)$$

Decomposition Model

This method compromises a sophisticated forecasting technique that integrates the historical data into different components to perform forecast value. All-time series components of trend, seasonal, cycle, and irregular were considered in the model. First, the trend component and seasonal indices must be calculated before irregular and cyclic components. Both later components have to be isolated to better forecast the value.

Autoregressive Model and Moving Average Model

This model connects the time series' present value to random errors in previous periods. It consisted of a combination of the Autoregressive (AR) and Moving Average (MA) models. AR uses data from previous time steps as input to a regression equation to forecast the value at the next stage. The AR model takes in the number of previous time steps by looking at the Partial Autocorrelation Function (PACF) as *p*. Meanwhile, the MA model is expressed as a function of the error term. The Autocorrelation Function (ACF) plot of the time series is used to estimate *q*. By abstraction with integrated (I), which expressed the differencing raw data of the time series from nonstationary to stationary, combining both AR and MA, it designed the autoregressive integrated moving average ARIMA (*p, d, q*) model. If the series is nonstationary, the process of identifying differencing degree, *d*, is repeated until the data is stationary (Osarumwense, 2014; Singh et al., 2019). Therefore, the ARIMA model's General Equation is presented in Equation (2).

$$X_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} - \sum_{i=1}^{p} \theta_i \in_{t-i,} \quad (2)$$

where,
ϕ_i is the AR parameter, θ_i is the MA parameter, and $\in_t$ is the series of random unknown errors.

Exponential Smoothing Method

This model is explicitly a forecasting model with an exponentially weighted average of prior observations. In this method, the weights were inversely related to the data collection time. The range for the weight of alpha (α) is between 0 and 1. Each weight corresponding to an observation exhibits a declining trend. More weightage will be given to past observations if the value of α is close to zero. Vice versa, more weightage will be given to immediate observations if the value of α is close to 1 (Jain & Mallick, 2017). It also can be computed using double parameters with α and β smoothing factors, which control the rate of influences on the past observations.
The General Equation is given as Equation (3).

$$y_{t+h|t} = \sum_{j=1}^{t} \alpha^{t-j}(1-\alpha)^t l_0. \quad (3)$$

Combined Models

A model that was close to one another was typically difficult to distinguish with a small or moderate number of observations, as anticipated, and the model selection criterion values are typically quite similar. In such a situation, selecting the model with the lowest criterion value is unreliable. A small shift could influence the choice of a different model in the data. As a result, the forecast made using the chosen algorithm may have a high level of variability. In common research practice, (1) model averaging, (2) Ensemble methods, and (3) Combination of residuals were the several methods of combining the forecasting models. Model averaging involves taking a weighted average of the forecasts generated by the two ARIMA models. The weights can be based on the performance of the models on past data or can be assigned based on expert judgment. In ensemble methods, this involves combining the forecasts generated by the two ARIMA Models using an ensemble method, such as bagging, boosting, or

stacking. This involves more complex machine learning methods that can help to improve the accuracy and stability of the forecasts by leveraging the strengths of multiple models and reducing the impact of any weaknesses in a single model. Last but not least, combining the residuals method involves combining the residuals (the difference between the observed and predicted values) generated by the two ARIMA Models and using them to fit a third ARIMA model (combined model). The third model can help capture any patterns or relationships not captured by the original models and improve the accuracy of the forecasts. Therefore, the general combined model in this study is represented by Equation (4).

$$\hat{Y}^* = \sum_{i=1}^{M}(1 - \%performance\ weightage_{Model\ i}) * Forecast\ value_{Model\ i}, \quad (4)$$

where

$$\%performance\ weightage_{Model\ i} = 1 - \left(\frac{Measurement\ error_{Model\ i}}{\sum_{i=1}^{M} Measurement\ error_{Model\ i}}\right)$$

M = number of models to combine

$\hat{Y}^*$ = Combined model forecast value

Phase 4: Modeling and Evaluation of the Data

The chosen time-series technique is modeled in this phase, and the error measurements are calculated. The error measurements are used to compare the performance of the different techniques. The technique with the least error measurement is chosen as the best method for predicting rainfall trends. The techniques are evaluated using four error measurements, which are Mean Absolute Deviation (MAD), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

MSE is the average of the squared differences between the observed and predicted values. It penalizes larger errors more heavily than smaller errors, making it a more sensitive measure of prediction accuracy. On the other hand, MAD is a statistical measure of the average distance between each data point and the mean of the dataset. It is a measure of the variability or spread of the data. MAD is useful in evaluating the accuracy of a forecasting model, such as an ARIMA model, by measuring the deviation between the actual and forecasted

values. The smaller the MAD, the better the accuracy of the model. One limitation of MAD is that it gives equal weight to each data point, regardless of its magnitude or importance. In cases where some data points are more important than others, weighted MAD can be used, where each data point is assigned a weight based on its importance. Other measures of error include MAPE, which measures the percentage difference between the observed and predicted values. Meanwhile, MPE is the mean percentage error (or deviation), which essentially scales mean error to be in percentage units instead of the variable's units. Table 1 presents the formula of MSE, MAD, MAPE, and MPE.

Moreover, advanced time series analysis on ARIMA models was also tested and evaluated using Ljung-Box Q-test, Bayesian Information Criterion (BIC), and R-squared. The Ljung-Box Q-test was a diagnostic tool used to assess the presence of autocorrelation in the residuals of an ARIMA model. A significant Q-statistic suggests that there are remaining patterns in the residuals that the model has not accounted for, indicating the need for further model refinement. Consequently, BIC is a statistical criterion used for model selection, particularly when comparing different ARIMA models with varying orders of autoregressive and moving average components. A lower BIC value indicates a better-fitting model, making it a valuable tool for choosing the most appropriate ARIMA configuration. R-squared, also known as the coefficient of determination, measures the proportion of variance in the dependent variable (i.e., the time series data) explained by the ARIMA model. A higher R-squared value indicates a better fit, suggesting that the ARIMA model effectively explains the observed variations in the time series.

These diagnostic tests collectively assist in evaluating the goodness-of-fit, adequacy, and predictive power of ARIMA models, aiding researchers in making informed decisions regarding the suitability and performance of their time series forecasting models.

Table 1

Equation of Measurements Error

Error of Measurement	*MSE*	*MAD*	*MAPE*	*MPE*
Equation	$\frac{\sum e^2}{n}$	$\frac{\sum \lvert e \rvert}{n}$	$\frac{\sum \frac{\lvert e \rvert}{y^t} \times 100}{n}$	$\frac{\sum \frac{e}{y^t}}{n}$

Phase 5: Forecast the Rainfall Volume

In the last phase of the study, when the best-fit model is chosen based on the smallest of measurements, alternate methods of combining models are considered to result in a lower measurement of errors for predicting the values. Once the methods of combining models are chosen, the future rainfall data is forecasted in a simple way using the most minimum error measurements methodology for the year 2023.

ANALYSIS AND RESULTS

Understanding Rainfall Patterns

The study analyzed the time series component of rainfall volume data in Petaling and Subang stations in Malaysia. Figures 2 and 3 illustrate the data, which was found to have a slight upward trend but was assumed to be stationary. Based on Figures 4 and 5, autocorrelations and partial autocorrelation analyses were performed to identify the optimal parameters for ARIMA models used for forecasting. The results showed that both data had seasonal peaks across the months, and the mean distribution of the rainfall volume was not the same, indicating the presence of seasonality. A non-parametric test of Kruskal Wallis was used to validate the seasonality. The p-values were discovered at 0.019 and 0.014, respectively, to be less than 0.05 significance level, which led to the rejection of the null hypothesis and concluded that all the mean distribution of the rainfall volume were not the same and contained seasonality as shown in Figure 6.

Figure 2

Sequence Chart for Petaling Station

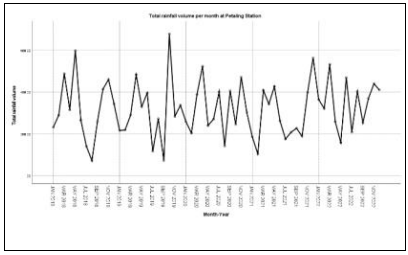


Figure 3

Sequence Chart for Subang Station

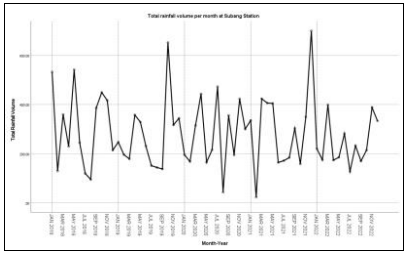


Figure 4

ACF & PACF Graphs for Petaling Station

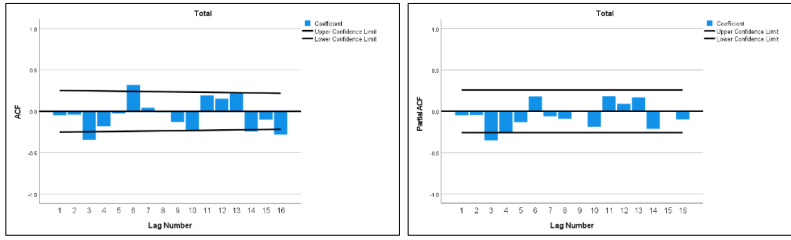


Figure 5

ACF & PACF Graphs for Subang Station

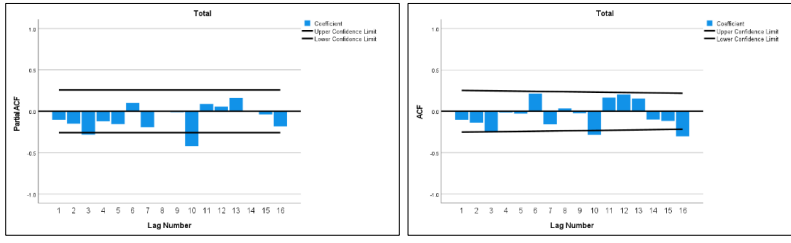
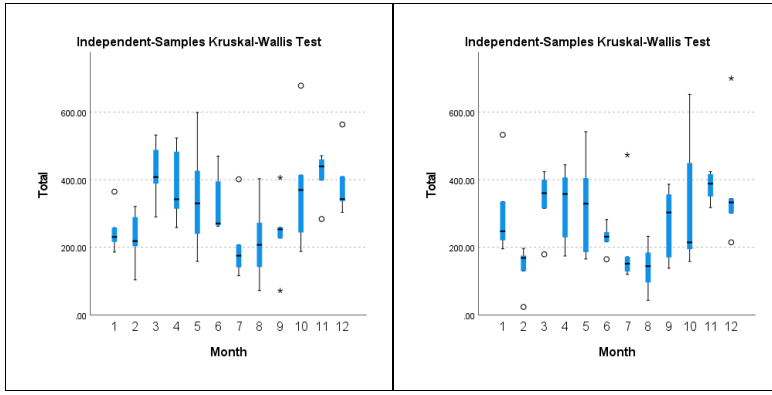


Figure 6

Kruskal-Wallis Test for Petaling (left) and Subang (right) Station



Performance Comparison Between Listed Models

This data was tabulated according to the listed models, and all of the listed models had been evaluated on the last 6 months of the data. A

summary of the performance evaluation on measurement of errors (MAD, MSE, MAPE & MPE) was presented in Tables 2 and 3 of both stations, respectively. The data showed huge mean square errors for all the models. The MAPE value also takes an extreme value if this value is exceedingly tiny or huge. It works better with data free of zeros and extreme values because of the in-denominator. Note that MSE is the average of the squared differences between the observed and predicted values in this scenario. It penalizes larger errors more heavily than smaller errors, making it a more sensitive measure of prediction accuracy.

Table 2

Measurement Errors of Petaling Stations from the Evaluation Part of the Data

Petaling Station	*MSE*	*MAD*	*MAPE*	*MPE*
Naïve	23,326.36	130.09	46.4427%	-17.4518%
Addictive	7,497.28	82.20	26.8388%	-3.2488%
Multiplicative	7,634.29	83.91	26.7658%	-1.0533%
ARIMA (1,0,1)	9,196.99	86.73	23.6507%	12.7799%
ARIMA (2,0,2)	9,115.73	86.15	23.4329%	12.8535%
ARIMA (2,0,3)	**6,581.61**	**67.22**	**16.8901%**	**16.3738%**
ARIMA (3,0,2)	7,821.09	74.82	19.4625%	16.5957%
ARIMA (3,0,3)	8,850.30	83.18	22.1177%	14.2369%
ARIMA (4,0,4)	9,041.48	83.97	22.2436%	14.4869%
Holt	7,666.67	83.62	26.9972%	-2.2020%
Brown	7,607.58	83.16	26.9417%	-2.5284%

Table 3

Measurement of Errors of Subang Station from Evaluation Part of Data

Subang Station	*MSE*	*MAD*	*MAPE*	*MPE*
Naïve	11,070.17	89.39	43.3463%	-8.1774%
Addictive	12,855.42	100.12	55.7371%	-47.6658%
Multiplicative	8,968.49	87.30	44.8023%	-29.5605%
ARIMA (1,0,1)	9,386.79	89.41	46.4870%	-32.2003%

(continued)

Subang Station	*MSE*	*MAD*	*MAPE*	*MPE*
ARIMA (2,0,2)	9,545.57	89.90	46.9172%	-32.6616%
ARIMA (2,0,3)	10,273.31	87.27	47.1982%	-33.0993%
ARIMA (3,0,2)	9,473.68	89.42	46.6840%	-32.4282%
ARIMA (3,0,3)	9,465.32	89.44	46.6695%	-32.4121%
ARIMA (4,0,4)	**8,481.22**	**81.34**	**44.1147%**	**-33.3135%**
Holt	9,492.53	89.93	46.6984%	-32.1500%
Brown	9,001.72	87.46	44.7058%	-29.0403%

The study compared the error of measurements of different models for forecasting rainfall from both stations. Naïve and decomposition models were not suitable for modeling continuous trend and seasonal data, and ARIMA models had relatively smaller errors. The naïve model is not able to consider the time series data with more complex patterns or dynamics. Meanwhile, the decomposition model assumes that the trend and seasonality components are constant over time, which may not be true for all-time series data (Chatfield & Xing, 2019; Hyndman & Athanasopoulos, 2018). The best ARIMA models were ARIMA (2,0,3) for Petaling Station and ARIMA (4,0,4) for Subang station, which passed the L-jung Box-Q test and had moderate fitness of data shown in Tables 4 and 5. Exponential smoothing models did not perform well in capturing trend and seasonality patterns in the data. This is because it assumed that the time series data have no trend or seasonality components and rely solely on past observations to forecast future values. If the time series data exhibits a clear trend or seasonality pattern, exponential smoothing models may not be able to capture these patterns effectively and may produce inaccurate forecasts (Hyndman & Athanasopoulos, 2018).

Table 4

ARIMA Model Test of Petaling Stations

ARIMA	*L-jung Box-Q*	*BIC*	*R-squared*
(1,0,1)	0.0010	9.9240	0.1590
(2,0,2)	0.0000	10.0900	0.1770
(2,0,3)	**0.0650**	**10.1040**	**0.2400**
(3,0,2)	0.0520	10.1020	0.2420
(3,0,3)	0.1040	10.0700	0.3320
(4,0,4)	0.0280	10.2470	0.3420

Table 5

ARIMA Model Test of Subang Station

ARIMA	*L-jung Box-Q*	*BIC*	*R-squared*
(1,0,1)	0.5400	9.9490	0.228
(2,0,2)	0.0230	10.135	0.229
(2,0,3)	0.2070	10.118	0.310
(3,0,2)	0.1560	10.151	0.288
(3,0,3)	0.1080	10.245	0.288
(4,0,4)	**0.0980**	**10.436**	**0.295**

Combining ARIMA Models

After evaluating the error of measurements for both stations, it was found that the errors were still relatively high for the models to be used for prediction. Therefore, combining both stations' selected models based on performance was considered to generate a single algorithm for forecasting. The aim was to reduce the variability of the combined forecast through a suitable weighting scheme to increase accuracy (Zou & Yang, 2004). The weightage assigned to each ARIMA model depended on its performance and contribution to the accuracy of the combined forecast. Different methods were available for determining the weightage, and the choice depended on the specific situation and objectives of the analysis. The paper considered performance weightage based on three error measurements and the effectiveness of the model in reducing the ratio of errors between the two models. The weightage tabulation was provided as Equation (5).

$$\%performance\ weightage = 1 - \left(\frac{Error\ of\ measurement\ of\ the\ station}{Total\ error\ of\ measurement\ of\ both\ stations}\right). \quad (5)$$

Table 6

Weightage of the Error of Measurement on the Selected Models

	MSE	*MAD*	*MAPE*	*MPE*
Petaling (2, 0, 3)	43.69%	45.25%	27.69%	-96.66%
Subang (4, 0, 4)	56.31%	54.75%	72.31%	196.66%

Table 7

Evaluation of the Last 6 Months on Combined ARIMA Models

Weightage of performance	*MSE*	*MAD*	*MAPE*	*MPE*
Weightage % MSE Performance	**3916.03**	55.70	**19.6265%**	-1199.7369%
Weightage % MAD Performance	3921.15	**55.40**	19.7950%	-1228.3665%
Weightage % MAPE Performance	4262.73	58.76	18.0125%	**-1081.9640%**

Table 6 presents the model performance ratio based on the generated errors, which revealed notable differences in MAPE and MPE readings. The evaluation of error measurements on the weightage applied to the combined models is provided in Table 7. The calculations showed that the weightage using the ratio of MSEs resulted in the lowest MSE for the combined model, which reduced the errors compared to ARIMA (2,0,3) and (4,0,4) models. Consequently, the combined models were selected with a weightage of 56.31% for the ARIMA (2,0,3) model at Petaling Station and 43.69% for the ARIMA (4,0,4) model using Equation (5).

Forecast Rainfall Volume

During the final stage of the study, Phase 5, the forecast of rainfall volume was conducted for the next 12 months of 2023 based on the weightage assigned to the combined model as identified. The first step was to generate separate forecast values for the year 2023 using ARIMA (2,0,3) for Petaling Station and ARIMA (4,0,4) for Subang Station. The results were then re-tabulated with the assigned weightage using the provided equations. Finally, the forecasted values for the rainfall of 2023 were obtained by applying the weightage to both stations, as shown in Table 8.

Table 8

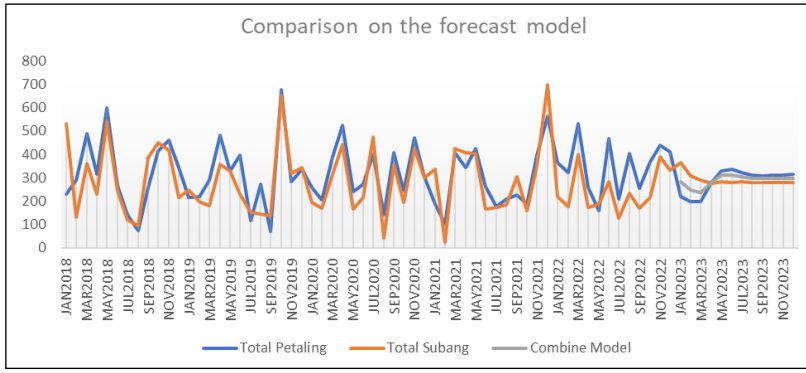
Forecast Values of ARIMA (2,0,3), (4,0,4), and Combined Models from Jan – Dec 2023

Model	*Jan-23*	*Feb-23*	*Mar-23*	*Apr-23*	*May-23*	*Jun-23*	*Jul-23*	*Aug-23*	*Sep-23*	*Oct-23*	*Nov-23*	*Dec-23*
ARIMA (2,0,3)	218.49	196.68	198.59	278.99	330.74	336.14	321.44	309.94	307.63	310.17	312.65	313.36
ARIMA (4,0,4)	364.42	309.26	289.83	276.16	283.4	280.44	283.4	277.78	280.05	277.71	280.67	278.71
Combined models	**282.25**	**245.86**	**238.45**	**277.75**	**310.05**	**311.81**	**304.82**	**295.89**	**295.58**	**295.99**	**298.68**	**298.22**

Journal of Computational Innovation and Analytics, Vol. 3, Number 1 (January) 2024, pp. 83-103

Figure 7

Forecasting Value of the ARIMA Models and Combined Models



The graphical representation in Figure 7 depicts the combined model's forecast value, which was well-fitted between both models at both stations. This approach was aimed to minimize the errors of the measurements, making them more appropriate and nearer to the actual rainfall readings between the stations. The upward trend of rainfall volume commences in March 2023, while a stable rainfall volume is evident from May 2023 to December 2023.

CONCLUSION

In conclusion, the study explored the use of univariate time series models to forecast rainfall volume in Selangor, using data from two stations. The study discovered that ARIMA (2,0,3) was the best model for the Petaling station, and ARIMA (4,0,4) was the best for the Subang station. To improve the accuracy of the forecast, the study combined the two models using a weighted approach and obtained a smaller error in measurements. The forecasted rainfall volume indicated a peak in May, June, and July 2023 and consistent rainfall until the end of 2023. Univariate time series models can be useful tools for understanding rainfall patterns in Selangor and supporting decision-making in agriculture, water resource management, and flood prevention. This study could enhance disaster preparedness, infrastructure planning, economic resilience, agricultural management, urban planning, environmental conservation, community safety, and scientific understanding. Accurate predictions empower authorities,

businesses, and communities to make informed decisions, mitigate the impact of flash floods, optimize resource allocation, ensure sustainable development, and safeguard lives and property.

However, the paper acknowledges its limitations, including the limited and incomplete data on rainfall volume in Klang region, Selangor, the need to reorganize the area of the collection to achieve data completeness, and the failure to capture the complex and dynamic interactions between different environmental factors that influence rainfall patterns in Selangor, such as deforestation. Furthermore, the study may not account for the impacts of climate change or other long-term trends that may affect rainfall volume over time, affecting the reliability and generalizability of research findings.

In future studies, the researcher suggests more precise data collection for longer periods of time to improve rainfall volume forecasting. The model should also be re-evaluated periodically, such as every five years, to account for climate change and other long-term trends. Additionally, the researcher suggests considering more complex and flexible models incorporating external variables contributing to heavy rainfall, such as multiple regression integration. To confirm the accuracy and effectiveness of the proposed model, it is also recommended that its predictions be validated using additional datasets from other stations. This cross-validation process will help assess the model's performance in different scenarios and environments, ultimately contributing to its reliability and potential adoption for practical applications.

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The author gratefully acknowledges the support from myMET data through Jabatan Meteorologi Malaysia, providing secondary data for this study.

REFERENCES

Adnan, R., Ruslan, F. A., Samad, A. M., & Md Zain, Z. (2012). Flood water level modelling and prediction using artificial neural network: Case study of Sungai Batu Pahat in Johor. *2012 IEEE Control and System Graduate Research Colloquium*, 22–25. https://doi.org/10.1109/ICSGRC.2012.6287127

Chatfield, C., & Xing, H. (2019). *The analysis of time series: An introduction with R*. CRC press. https://shorturl.at/opQU7

Department of Statistics Malaysia. (2022). *Special report on impact of floods in Malaysia 2022*. Department of Statistics Malaysia. https://shorturl.at/gnNT4

Diya, S. G., BarzaniGasim, M., EkhwanToriman, M., & Abdullahi, M. G. (2014). Floods in Malaysia: Historical reviews, causes, effects and mitigations approach. *International Journal of Interdiciplinary Research and Innovations*, *2*(4), 59–65. https://shorturl.at/bkM19

Hong, J. L., & Hong, K. (2016). Flood forecasting for Klang River at Kuala Lumpur using artificial neural networks. *International Journal of Hybrid Information Technology*, *9*(3), 39–60. https://doi.org/10.14257/ijhit.2016.9.3.05

Houck, P. R., Mazumdar, S., Koru-Sengul, T., Tang, G., Mulsant, B. H., Pollock, B. G., & Reynolds, C. F. (2004). Estimating treatment effects from longitudinal clinical trial data with missing values: Comparative analyses using different methods. *Psychiatry Research*, *129*(2), 209–215. https://doi.org/10.1016/j.psychres.2004.08.001

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts. https://shorturl.at/aFP38

Jain, G., & Mallick, B. (2017). A Study of Time Series Models ARIMA and ETS. *SSRN*. https://ssrn.com/abstract=2898968 or http://dx.doi.org/10.2139/ssrn.2898968

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Wiley.

Mishra, N., Soni, H. K., Sharma, S., & Upadhyay, A. K. (2018). Development and analysis of artificial neural network models for rainfall prediction by using time-series data. *International Journal of Intelligent Systems and Applications*, *12*(1), 16–23. https://doi.org/10.5815/ijisa.2018.01.03

Mustapha, H., & Ismail, N. (2021). Time series modeling and forecasting of monthly rainfall using autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA). *Proceedings of Science and Mathematics*, 79–88.

Osarumwense, O.-I. (2014). Time series forecasting models: A comparative study of some models with application to inflation data. *Open Science Journal of Statistics and Application*, *2*(2), 24–29. http://www.openscienceonline.com/journal/osjsa

Reza, M. I. H. (2016). Southeast Asian landscapes are facing rapid transition: A study in the state of Selangor, Peninsular Malaysia. *Bulletin of Science, Technology and Society*, *36*(2), 118–127. https://doi.org/10.1177/0270467616668075

Shaharudin, S. M., Andayani, S., Kismiantini, Binatari, N., Kurniawan, A., Basri, M. A. A., & Zainuddin, N. H. (2020). Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1.4), 646–651. https://doi.org/10.30534/ijatcse/2020/9091.42020

Singh, N., Sharma, N., Sharma, A. K., & Juneja, A. (2019). Sentiment score analysis and topic modelling for GST implementation in India. *Advances in Intelligent Systems and Computing*, *817*, 243–254. https://doi.org/10.1007/978-981-13-1595-4_19

Suparta, W., Rahman, R., Singh, M. S. J., & Latif, M. T. (2015). Investigation of flash flood over the west Peninsular Malaysia by global positing system network. *Advanced Science Letters*, *21*(2), 153–157. https://doi.org/10.1166/asl.2015.5845

Syafrina, A. H., Zalina, M. D., & Juneng, L. (2015). Historical trend of hourly extreme rainfall in Peninsular Malaysia. *Theoretical and Applied Climatology*, *120*(1–2), 259–285. https://doi.org/10.1007/s00704-014-1145-8

Tunc, M., Bayraktar, D., & Gunalay, Y. (2016). A new approach for combining ARIMA and seasonal ARIMA models to improve forecasting accuracy. *Expert Systems with Applications*, *57*, 309–316.

World Bank Group. (2021). *Climate risk country profile: Malaysia*. Asian Development Bank. www.worldbank.org

Zhang, W., He, J., & Wang, J. (2019). A hybrid ARIMA-AR model for improved time series forecasting. *IEEE Access*, *7*, 140460–140472.

Zou, H., & Yang, Y. (2004). Combining time series models for forecasting. *International Journal of Forecasting*, *20*(1), 69–84. https://doi.org/10.1016/S0169-2070(03)00004-9