APPLIED MATERIALS AND TECHNOLOGY

Jamt.ejournal.unri.ac.id

Research Article

A comparative study of data-driven models for discharge forecasting: a study case of Siak river, Pekanbaru water gauge station

Manyuk Fauzi*, Bambang Sujatmoko, Igeny Dwiana Darmawan, Siswanto, Ermiyati, Merley Misriani

ABSTRACT: The availability of long-term river discharge data covering at least 30 years is needed for proper hydrological studies, so the ability to predict river discharge is a matter of concern in the field of civil engineering. The Siak river in Pekanbaru c ity experiences overflowing water during the rainy season. One of the steps to prevent flooding on the Siak River is to utilize river discharge information, data-driven models utilize historical data to train or derive useful insights for predicting outputs, some data-driven models that are suitable for generating monthly historical data into new data include the Autoregressive Integrated Moving Average (ARIMA) method and the Thomas-Fiering method. The research begins by conducting the Rescaled Adjusted Partial Sums (RAPS) test to test the homogeneity of the data, then the prediction of discharge data with several schemes using the ARIMA and Thomas-Fiering methods, then the performance comparison between the two models is carried out using MAPE, RMSE, Nash-Sutcliffe, and correlation coefficient r. From the research results, it was found that the Thomas-Fiering method tends to be more accurate for predicting 1-year monthly discharge as well as long-term discharge, namely periods of 3, 5, and 7 years, with the best prediction being 1-year discharge prediction using 5 years of observed discharge with MAPE, RMSE, Nash-Sutcliffe, and correlation coefficient r values of 7.42%, 26.76 m^3s^{-1}, 0.92, and 0.96, respectively. This study could be a valuable reference for future studies in selection and further modification of data driven discharge simulation models.

Keywords: *Siak watershed, Stream flow prediction, Data driven model, ARIMA, Thomas-fiering*

OPEN ACCESS

Affiliation
Departmen of Civil Engineering, Riau University, Pekanbaru 28293, Indonesia.

***Correspondence**
Email: manyukfauzi@lecturer.unri.ac.id

ORCID
Manyuk Fauzi: 0000-0002-7798-1420

Received: December 11, 2024
Revised: April 22, 2025
Accepted: April 24, 2025

How to cite: Fauzi, M., Sujatmoko, B., Darmawan, I., D, Siswamto, Ermiyati, Misriani, M., (2025). comparative study of data-driven models for discharge forecasting: a study case of siak river, pekanbaru water gauge station. *Journal of Applied Materials and Technology*, 6(2), 47–57. https://doi.org/10.31258/Jamt.6.2.47-57.

1. INTRODUCTION

The availability of long-term river discharge data covering at least 30 years is required for proper hydrological studies [1]. Due to the importance of river discharge data, so the ability to predict river discharge data is also a matter of concern in the field of civil engineering, this has led to an increasing need for time series hydrological modeling that is able to imitate a series of historical discharge data of a river [2]. The evaluation and planning of water projects necessitate the forecasting of water resources using hydrological data spanning a sufficient time series. However, hydrology researchers encounter challenges due to the limited availability of monitored hydrological series over extended periods. Monitoring stations are sparse, and in many cases, data records are insufficient or absent [3].

The Siak River is a river that is administratively located entirely in Riau Province, passing through Rokan Hulu Regency, Kampar Regency, Bengkalis Regency, Pekanbaru City and Siak Regency [4]. According to the Decree of the Minister of Public Works Number 424/KPTS/M/2013 concerning the Water Resources Management Pattern of the Siak River Basin, the Pekanbaru City area is an area that is often affected by flooding from the Siak River and the flood-prone area in Pekanbaru City reaches 8,755 ha, but the water level data recorded at the pekanbaru water gauge station is only 23 years of data. One of the steps to prevent flooding is to utilize river discharge information, the generated synthetic streamflow discharge is as important as the historical river discharge to study several feasible alternatives for planning, design, and operation of water resources projects [5]. There are two main categories of prediction models: (i) knowledge-driven models and (ii) data-driven models. Knowledge-driven models are effectively applied to known physical characteristics of catchments such as area, shape, slope, stream length, altitude, etc. These models, exemplified by rainfall-runoff modeling and empirical relationships, rely on understanding the internal mechanisms of the system between input and output data. In contrast, data-driven or black box models do not explicitly account for these internal mechanisms. They perform well even when there is limited information available about the physiographic characteristics of the catchment [6], data-driven models utilize historical data to train or derive useful insights for predicting outputs. The accuracy of forecasting is primarily influenced by two key factors: the duration and quality of historical data available [7]. Examples of black-box or data-driven models include artificial intelligence techniques, regression models, and stochastic models. Some stochastic methods that are suitable for generating monthly historical data into new data include the Autoregressive Integrated Moving Average (ARIMA) method and the Thomas Fiering method [8]. An important reason for using statistical analysis in the field of hydrology is related to natural phenomena, which involve uncertainties in space and time. There is no single hydrological model that gives an absolute percentage of success to the natural event process. The use of statistical methods accommodates the probailistic nature of the natural event process and the use of statistical methods is often very relevant.

The ARIMA method is known to have the advantage of involving seasonal factors in the modeling [2], researchers prefer the ARIMA method because of its systematic procedures including identification, estimation, and diagnostic checking. In the process of modeling using the ARIMA method, software assistance is needed and the ARIMA method also requires stationary data, efficient execution and prediction with the ARIMA method requires a lot of research experience because it is a complex modeling technique, this is another weakness associated with this technique [6]. In contrast to this, the Thomas Fiering method is known to be simpler, where Thomas and Fiering developed a model for generating monthly river flows that implicitly allows for non-stationarity in monthly flow data [9]. The Thomas-Fiering method treats the flow in each period as a linear function of the flow in the previous period, is flexible and easy to use in Microsoft Excel [10]. However, compared to other methods such as ARIMA, the Thomas-Fiering method has a tendency to overestimate river flow especially in low flow months [11].

Previous research that has been done succeeded in generating 62 years synthetic stream flow data used Thomas-Fiering for Ofu River and demonstrated that the generated synthetic data is about 95.9% reliable implying that it could be used for hydrological studies and projects [11]. More so, Thomas-Fiering model generally underestimated streamflow in most simulation months. However, based on measures of goodness-of-fit, the model's performance was deemed adequate for generating synthetic monthly streamflow data for the Jakham river [10]. Furthermore, prediction of long-term (monthly) streamflow for an intermittent river using SARIMA, the Thomas-Fiering model, and ANN models and it was found that the Thomas-Fiering model, which is regression-based, was unable to perform adequately during periods of low flow. Additionally, the ARIMA model demonstrated strong performance in predicting streamflow for the intermittent river, particularly in capturing peak or high discharge events better than other models [6]. ARIMA and a modified Thomas-Fiering model have also been used to analyze the time series of monthly rainfall data from the Tallafar station, the findings indicated that the modified Thomas-Fiering model was the most appropriate for representing the characteristics of the station [12].

The aim of this study is to predicting the discharge data of the Siak watershed, Pekanbaru water gauge station, with two data driven models that belong to stochastic models, namely the ARIMA method with the help of Minitab software and the Thomas Fiering method, which aims to see the accuracy of the two models in forecasting the discharge data of the Siak watershed (Pekanbaru Water Post). In general, application of Thomas-Fiering and ARIMA methods for generating synthetic streamflows is rare in Riau region, Indonesia. Therefore, this study was undertaken with the objective of developing and validating the monthly streamflow of Siak river, specifically Pekanbaru water gauge stasion using ARIMA dan Thomas-Fiering methods. Based on the previous studies that have been conducted, the ARIMA and Thomas-Fiering methods, which are data driven models, have different performance depending on the pattern, trend, and length of data at each research location. The advantages and disadvantages of the two methods have also been presented. This study will look at how the two models work to predict discharge at one of the Siak River water gauge stations by creating several discharge prediction schemes with different variations of predicted and observed discharge.

2. MATERIALS AND METHODS

2.1. Research location. Administratively, Pekanbaru water gauge station of Siak watershed located in Kampung Bandar Village, Senapelan District, Pekanbaru City, Riau Province, with a geographical location of 00° 32' 27.2“ S Latitude 101° 26' 14.5” E Longitude. Figure 1 is location map of the Pekanbaru water gauge station, where the water level data is measured.

2.2. Data availability. The data needed in this study was secondary data, specifically Automatic Water Level Recorder (AWLR) data from the Pekanbaru water gauge station, obtained from January 1, 2013, to December 31, 2022, through the Sumatra III River Basin Center. The water level data was subsequently converted into discharge data using the calibration curve equation as follows size:

$$Q = 6.029 \times (H + 2.481)^{2.738} \quad (1)$$

Where Q is the river flow discharge (m^3s^{-1}), and H is the water level (m).

2.3. Discharge data forecasting. Automatic Water Level Recorder (AWLR) data for 10 years (2013-2022) was converted into discharge data with a calibration equation, then consistency testing was carried out on the discharge data using Rescaled Adjusted Partial Sums (RAPS), after the data was known to be consistent, several discharge prediction schemes were made as shown in Table 1. The schemes were created with various lengths of observed discharge, ranging from a short period of 3 years to increasingly longer periods of 5 years, 7 years, and 9 years. The

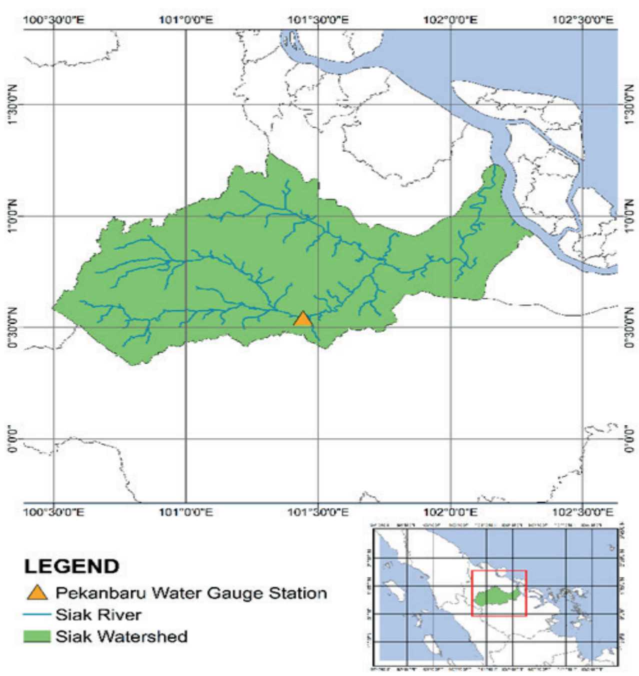

Figure 1. The Study Location.

observation discharge data were utilized to predict 1 year period discharge, aiming to assess the ability of observation discharge with different lengths in predicting 1-year period discharge. Furthermore, the observation discharge was also employed to predict long-term discharge, this was carried out to evaluate the ability of observation discharge with varying lengths to predict long-term discharge as well. The selection of the observation discharge length and the length of discharge to be predicted, as listed in Table 1, was based on the consideration that periods of 3, 5, 7, and 9 years were deemed capable of representing the prediction of discharge, starting from short discharge data to long discharge data. After that, discharge prediction was carried out using the ARIMA and Thomas-Fiering methods according to the scheme. Then, after obtaining the predicted discharge, the accuracy level was measured with several performance indicators such as MAPE, RMSE, Nash-Sutcliffe coefficient, and correlation coefficient r to assess the accuracy and capability of the method.

Table 1. Discharge prediction scheme.

Scheme	Period of observed discharge ussed	Period of discharge to be predicted
I	3 years (2013-2015)	1 year (2016)
II	5 years (2013-2017)	1 year (2018)
III	7 years (2013-2019)	1 year (2020)
IV	9 years (2013-2021)	1 year (2022)
V	3 years (2013-2015)	7 years (2016-2022)
VI	5 years (2013-2017)	5 years (2018-2022)
VII	7 years (2013-2017)	3 years (2020-2022)

2.4. Rescaled Adjusted Partial Sums (RAPS) consistency Test. To perform statistical analysis of hydrological data, it was assumed that the data was consistent. Therefore, before starting statistical analysis on hydrological data, it was important to identify and remove inconsistencies and inhomogeneities in the data. This action was necessary because these characteristics were undesirable to persist in the future [13]. The data consistency test used for this research was Rescaled Adjusted Partial Sums (RAPS), with the equation as follows:

$$RAPS = \frac{\sum_{i=1}^{k}(Qi-\bar{Q})}{\sqrt{\sum_{i=1}^{n}\frac{(Qi-\bar{Q})^2}{n}}} \tag{2}$$

Where Qi is i-th discharge data, $\bar{Q}$ is average discharge data, and n is total of the data.

2.5. Thomas-Fiering method. Basically, the Thomas Fiering method is a natural Markovian with periodic parameters, namely the mean, standard deviation and correlation between consecutive data. The model consists of 12 regression equations, one for each month[13]. The equation is generally written as follows [9]:

$$Q_{ij}=\bar{Q}_j+b_j\left(Q_{ij-1}-\bar{Q}_{j-1}\right)+t_{ij}s_j\sqrt{1-r_j^2} \tag{3}$$

Where Q_{ij} is forecasted discharge (in the i-th year and j-th month), $\bar{Q}_j$ is the j-th average discharge, Q_{ij-1} is discharge in the i-th year (in the previous month), $\bar{Q}_{j-1}$ is average discharge in the previous month, b_j is regression coefficient, r_j is correlation coefficient, s_j is standard deviation, and t_{ij} is normally distributed random number (in the i-th year and j-th month).

All parameters used in the Equation 2 such as regression coefficient, correlation coefficient, standard deviation were calculated using the equations below as follows:

$$S_j=\sqrt{\frac{\sum_{i=1}^{n}(Q_{ij}-\bar{Q}_j)^2}{n-1}} \tag{4}$$

$$r_j=\frac{\sum(Q_{ij}-\bar{Q}_j)\times(Q_{ij-1}-\bar{Q}_{ij-1})}{\sqrt{\sum_{i=1}^{n}(Q_{ij}-\bar{Q}_j)^2\times\sum_{i=1}^{n}(Q_{ij-1}-\bar{Q}_{ij-1})^2}} \tag{5}$$

$$b_j=\frac{r_j\times S_j}{S_{j-1}} \tag{6}$$

A random number generator is an algorithm used to create a series of random numbers, either through manual counting or electronic (computer) computation. Random numbers are generated by obeying probabilities in the range of 0 to 1 and have a uniform distribution. The requirements for the generation of random numbers include being random, non-repeating (Degenerative) and the return period usually appears very long [14]. The uniform random component provided by the computer could be converted into a normal distribution with the Box-Muller equation as follows:

$$t_1=(-2 \ln u_1)^{\frac{1}{2}} \times \cos(2\pi \times u_2) \quad (7)$$

$$t_2=(-2 \ln u_1)^{\frac{1}{2}} \times \sin(2\pi \times u_2) \quad (8)$$

Where t_1 and t_2 are normally distributed random number and u_1 and u_2 are uniformly distributed random numbers.

2.6. ARIMA method. Autoregressive Integrated Moving Average (ARIMA) is a time series model introduced by Box and Jenkins in 1976. The model in forecasting required data that was stationary in both mean and variance. Data that had not been stationary in variance needed to be transformed, and the transformation used was the Box-Cox transformation. Data that was not yet stationary in the mean needed to be differentiated. The differentiation process was a process of looking for differences between data from one period to another in sequence, and the level of differentiation process done was represented by d in ARIMA (p,d,q). In general, the form of ARIMA (p,d,q) as follows [15]:

$$\phi(B)\,(1\text{-}B)^d\hat{Z}_t = \theta_0+\theta(B)a_t \quad (9)$$

$$\phi(B)=1\text{-}\phi_1 B\text{-}\dots\text{-}\phi_p B^p \quad (10)$$

$$\theta(B)=1\text{-}\theta_1 B\text{-}\dots\text{-}\theta_q B^q \quad (11)$$

Where ϕ (B) is AR operator, θ (B) is MA operator, p is AR order, q is MA order, and d is the level of differencing process.

If the discharge data contained a seasonal element, which could be defined as a pattern that repeated regularly within a consistent time interval. It included the tendency to repeat patterns of behavior in seasonal periods, generally a year for monthly data. Seasonal ARIMA models could solve time series consisting of non-seasonal and seasonal parts [16]. The general form of the seasonal ARIMA model or ARIMA (p,d,q) (P,D,Q)S with the equation as follows [17]:

$$\phi_p(B)\Phi_P(B^S)(1\text{-}B)^d(1\text{-}B^S)^D Z_t=\theta_q(B)\Theta_Q(B^S)a_t \quad (12)$$

Where p, d, and q is non-seasonal order, P, D,and Q is seasonal order, S is number of seasonal periods, and a_t is residual at time t.

To obtain the value of AR and MA coefficients used in ARIMA, it is calculated using the non-linear least square method developed by Marquardt with a long iteration process [18]. In this study, the coefficient estimation process was carried out using Minitab software. The ARIMA modeling process consisted of several steps as follows:

2.6.1. Model identification. Model identification was done to see whether the time series data was stationary or not, checking data stationarity could be done by looking at the visual form of the time series plot or by examining the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot. To identify stationarity in variance, the Box-Cox plot could also be used, data that was stationary in variance was characterized by a value of $\lambda = 1$, if the data was found to be non-stationary in variance, it could be made stationary through Box-Cox transformation, if it was found that the data was non-stationary in mean, it could be made stationary through the differencing process. Data that was not yet stationary in mean was indicated by the presence of a pattern that died down very slowly on the ACF and PACF plots [13]. Additionally, it could also be characterized by the presence of lags that still exhibited patterns and contained seasonality [19]. Through the ACF and PACF plots, it could also be determined whether the data had a seasonal pattern, the seasonal pattern would be evident in the ACF and PACF lags that were significant in multiples of the season.

2.6.2. Temporary model estimation. Temporary Model Estimation was done by analyzing the Autocorrelation Function (ACF) plot and Partial Autocorrelation Function (PACF) plot, the determination of p (AR order) was based on the partial autocorrelation function (PACF), and the determination of q (MA order) was based on the autocorrelation function (ACF). Autocorrelation Function is the term autocorrelation is used to explain the association or mutual dependence between the values of the same periodic series in different periods [18] and Partial Autocorrelation Function is the partial autocorrelation measure is used to show the magnitude of the relationship between the current value of a variable and the previous values of the same variable (values for various time lags) assuming the influence of all other time lags is [18]. The way to determine the ARIMA Model based on the ACF and PACF plots was done according to the Table 2 [20].

Table 2. Determination of ARIMA model based on ACF and PACF plots.

ACF	PACF	Model
Cut off after lag 1 or lag 2, seasonal lag is not significant	Dying down	Non seasonal MA (q = 1 or q = 2)
Cut off after lag L, non seasonal lag insignificant	Dying down	Seasonal MA (Q = 1)
Cut off after lag L seasonal, non-seasonal lag cut off after lag 1 or 2	Dying down	Non seasonal-seasonal MA (q = 1 or q = 2; Q = 1)
Dying down	Cut off after lag 1 or lag 2, seasonal lag is not significant	Non seasonal AR (p = 1 or p = 2)
Dying down	Cut off after lag L, non seasonal lag insignificant	Seasonal AR (P = 1)
Dying down	Cut off after lag L seasonal, non-seasonal lag cut off after lag 1 or 2	Non seasonal-seasonal AR (p = 1 or p = 2; P = 1)
Dying down	Dying down	ARMA

2.6.3. Potential model estimation. Model parameter estimation aimed to determine the estimated values of the ARIMA model parameters. The estimated parameters then had to be tested to determine their significance in the model with hypothesis testing to test the significance of the parameters. At this estimation stage, mathematical calculation techniques were relatively complex, so the help of software was used [21], in this research Minitab 20 was used. The Hypothesis: (a) $H_0 = \theta_1 = 0$ (Parameter is not significant), (b) $H1 = \theta_1 \neq 0$ (Parameter is significant), with level of significance or α= 5%, H_0 is rejected if p-value< α.

2.6.4. Diagnostic check. Diagnostic checks were done to determine the accuracy of the model after significant parameters were obtained so that a valid model could be obtained [22]. The diagnostic check done was the white noise residual assumption test with the Ljung Box test, which was carried out with the help of Minitab 20 software. The basic assumption that residuals were white noise meant that there was no correlation between residuals with a mean equal to zero and constant variance. The residual independence test (white noise) could be done using the Ljung-Box test statistic [23]. Hypothesis: (a) $H_0 = \rho_1 = \rho 2 = \ldots = \rho k$ (residual white noise), (b) H1 = there is at least one value of $\rho k \neq 0$; k= 1, 2, …, k (residuals are not white noise), with significance level $\alpha = 5$, H_0 is rejected if the p-value< α.

2.6.5. Performance indicator. 2.6.5.1. MAPE. Mean Absolute Percent Error (MAPE) is a measure of average error in percentage terms. MAPE is the average absolute difference between the predicted value and the actual value, expressed as a percentage of the actual value [24]. MAPE accuracy criteria can be seen in Table 3 [25] . The MAPE value was calculated with the following eqution:

$$\text{MAPE}=\frac{100}{\text{n}}\sum\frac{\left|Q_{\text{observation}} - Q_{\text{prediction}}\right|}{Q_{\text{observation}}} \tag{13}$$

Table 3. Mape accurary criteria.

MAPE value	Criteria
<10%	Very good
10% - 20%	Good
20% - 50%	Average
>50%	Bad

2.6.5.2. RMSE. Mean Absolute Percent Error (MAPE) is a measure of The smaller the RMSE value (close to 0), indicating that the prediction results are more accurate [26]. RMSE value was calculated with the following equation:

$$\text{RMSE}=\sqrt{\frac{\sum (Q_{\text{observation}}-Q_{prediction})^2}{\text{n}}} \tag{14}$$

2.6.5.3. Nash-Sutcliffe. The Nash-Sutcliffe efficiency test shows the accuracy of the correlation relationship between measured and calculated data [27]. Used to assess the validity of the model by comparing the model simulation results with observational data [28]. The Nash-Sutcliffe coefficient value was calculated with the following equation equation:

$$\text{NSE}=1-\frac{\sum_{i=1}^{n} (Q_{\text{observation}}-Q_{\text{prediction}})^2}{\sum_{i=1}^{n} (\bar{Q}_{\text{observation}}-Q_{\text{prediction}})^2} \tag{15}$$

The Nash-Sutcliffe accuracy criteria can be seen in Table 4. below [29]:

Table 4. Nash-sutcliffe accuracy critea.

NSE value	Criteria
0.75 < NSE < 1.00	Very Good
0.65 < NSE < 0.75	Good
0.50 < NSE < 0.65	Average
NSE ≤ 0.50	Bad

2.6.5.4. Correlation coefficient r. The correlation coefficient illustrates the closeness of the relationship between these variables. It should be noted that a high or low correlation coefficient value does not describe the cause-and-effect relationship between two or more variables, but only describes the linear relationship between these variables [30]. Correlation coefficient r value was calculated with the following equation:

$$r=\frac{\sum (Q_{\text{hit}}-\bar{Q}_{\text{hit}})\times(Q_{\text{obs}}-\bar{Q}_{\text{obs}})}{\sqrt{\sum(Q_{\text{hit}}-\bar{Q}_{\text{hit}})^2\times\sum(Q_{\text{obs}}-\bar{Q}_{\text{obs}})^2}} \tag{16}$$

The grading criteria can be seen in Table 5 [31].

Table 5. Correlation coefficient accuracy criteria.

r coefficient value	Correlation
0.90-1.00	Very good
0.70-0.90	Good
0.50-0.70	Average
0.30-0.50	Bad
0.00-0.30	No correlation

3. RESULT AND DISCUSSION

3.1. Rescaled adjusted partial sums test. Statistical analysis of hydrological data requires the assumption that the data is consistent. Therefore, before performing statistical analysis of hydrological data, it was necessary to identify and eliminate data inconsistencies and inhomogeneities, as these characteristics were undesirable in future projections. Table 6 showed the results of data consistency test using RAPS, where the value of $Q\sqrt{n}$ was smaller than $Q\sqrt{n}$ table or the value of $R\sqrt{n}$ was smaller than $R\sqrt{n}$ table at the 99% trust level. Therefore, it can be concluded that the monthly discharge data from 2013-2022 was consistent and could be used for further analysis.

Table 6. RAPS data consistency test result at $\alpha = 0.01$.

Month	$Q\sqrt{n}$	$Q\sqrt{n}_{table}$	$R\sqrt{n}$	$R\sqrt{n}_{table}$	Conclusion
Jan	1.15	1.29	1.15	1.38	Consistent
Feb	0.96	1.29	1.02	1.38	Consistent
Mar	0.92	1.29	0.92	1.38	Consistent
Apr	1.32	1.29	1.32	1.38	Consistent
May	0.92	1.29	0.92	1.38	Consistent
Jun	1.13	1.29	1.13	1.38	Consistent
Jul	1.16	1.29	1.16	1.38	Consistent
Aug	1.21	1.29	1.21	1.38	Consistent
Sep	0.83	1.29	0.83	1.38	Consistent
Oct	0.78	1.29	0.78	1.38	Consistent
Nov	0.55	1.29	1.01	1.38	Consistent
Dec	0.97	1.29	0.97	1.38	Consistent

3.2. Thomas-Fiering method discharge prediction. To predict discharge using the Thomas-Fiering method, several parameters were first calculated such as average discharge, standard deviation, correlation coefficient, and regression coefficient. Here is one of the calculations of thomas fiering discharge prediction for scheme prediction IV, in Table 7 listed the parameters of the discharge data for a period of 9 years (2013-2021). tj or normally distributed random numbers were obtained from Microsoft Excel in the form of uniformly distributed random numbers, which were then converted to normal distribution using the Box-Muller equation, the automatic random number generation process was repeated until the best predicted discharge was obtained. After obtaining all the parameters needed, the calculation of discharge prediction was carried out with the Thomas-Fiering equation. The comparison graph between predicted discharge and observed discharge can be seen in Figure 8 through Figure 14.

Table 7. Parameters required for Thomas-Fiering method-discharge prediction.

Month	$\overline{Q}_j$ m^3s^{-1}	S_j m^3s^{-1}	r_j	b_j	t_j
Jan	386.58	75.59	0.44	0.26	0.29
Feb	377.85	143.15	0.69	1.32	-1.10
Mar	329.71	175.47	0.85	1.04	-0.45
Apr	304.00	130.60	0.67	0.50	0.35
May	322.67	107.86	0.65	0.54	1.02
Jun	237.09	48.17	0.62	0.28	1.06
Jul	233.84	68.01	0.33	0.47	0.59
Aug	254.98	62.43	0.91	0.84	0.15
Sep	244.89	52.27	0.72	0.60	1.65
Oct	294.41	71.61	0.97	1.33	0.19
Nov	497.28	148.33	0.52	1.07	1.15
Dec	566.35	128.39	0.31	0.27	-0.12

3.3. ARIMA method discharge prediction. The Box Cox plot performed on the discharge data for a period of 9 years (2013-2021) for prediction schemes IV, showed the value of $\alpha = 0.00$ as shown in Figure 2, indicating that the data had not been stationary in variance, therefore a Box-Cox transformation was done to stationarize the data, the transformation results showed that the data had become stationary in variance, indicated with $\alpha = 1$ as shown in Figure 3.

Furthermore, after the data had become stationary in variance, it was checked whether the data had also become stationary in mean by looking at the Autocorrelation Function (ACF) plot and Partial Autocorrelation Function (PACF) plot. The ACF and PACF plots could be seen in Figure 4 and Figure 5.

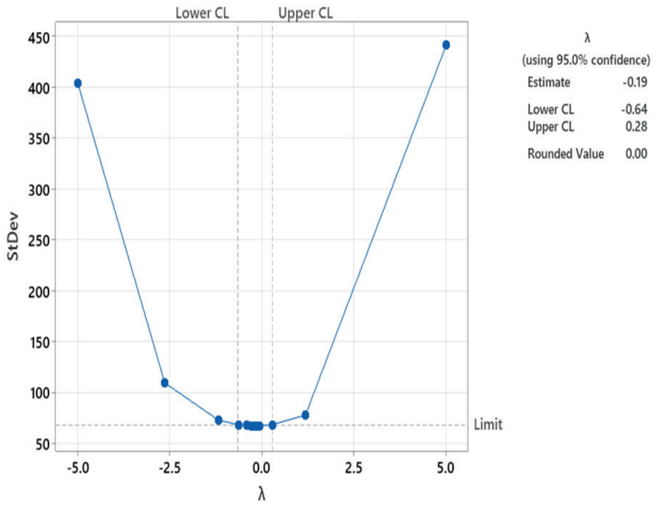

Figure 2. Box Cox plots of 9 year average monthly discharge data (2013-2021).

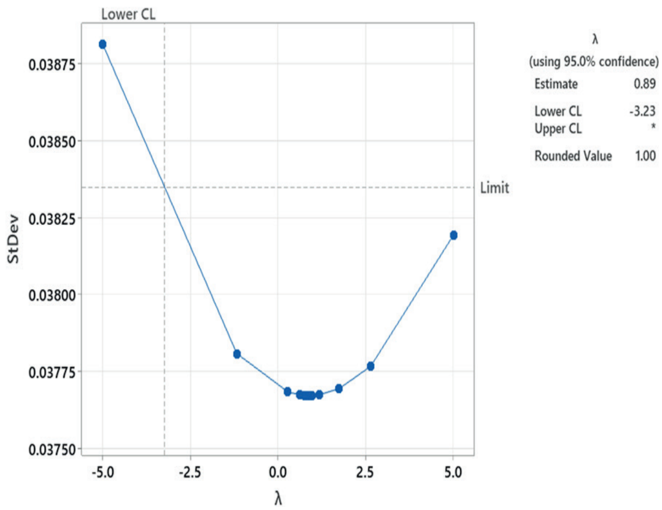

Figure 3. Box Cox plots discharge data after Box Cox transformation.

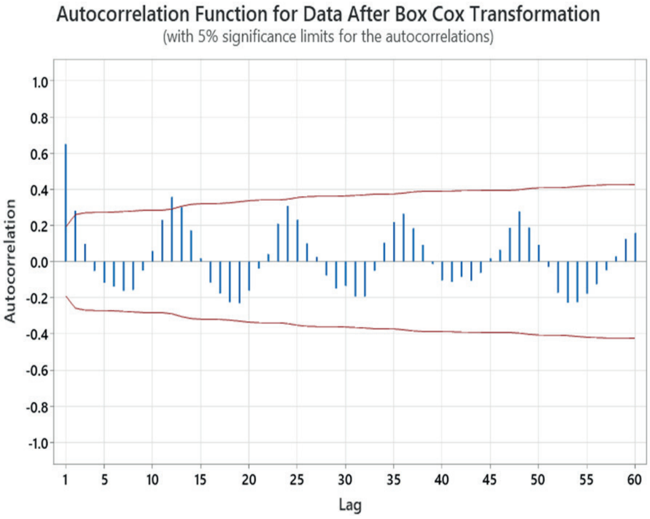

Figure 4. ACF Plot of Data after Box Cox transformation.

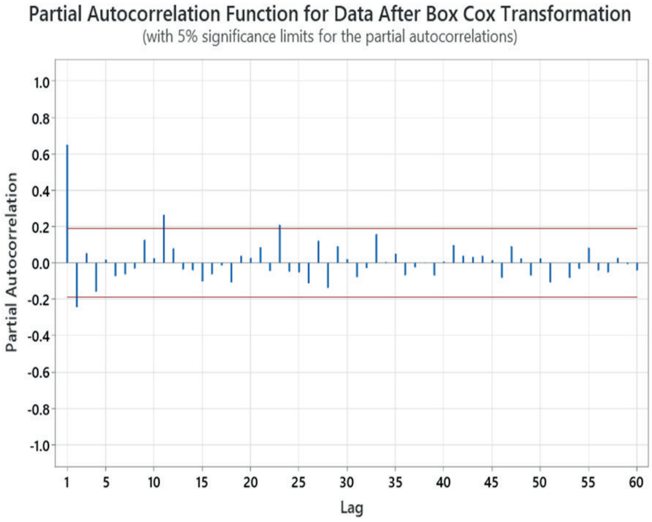

Figure 5. PACF Plot of Data after Box Cox transformation.

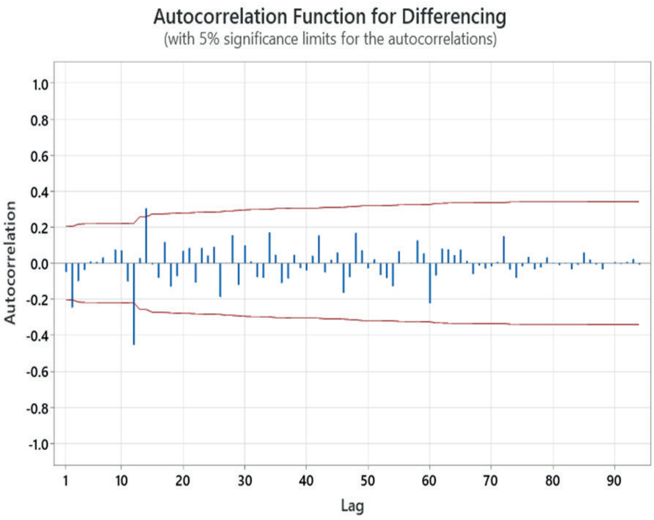

Figure 6. CF Plot of Data after differencing d = 1 and D =1.

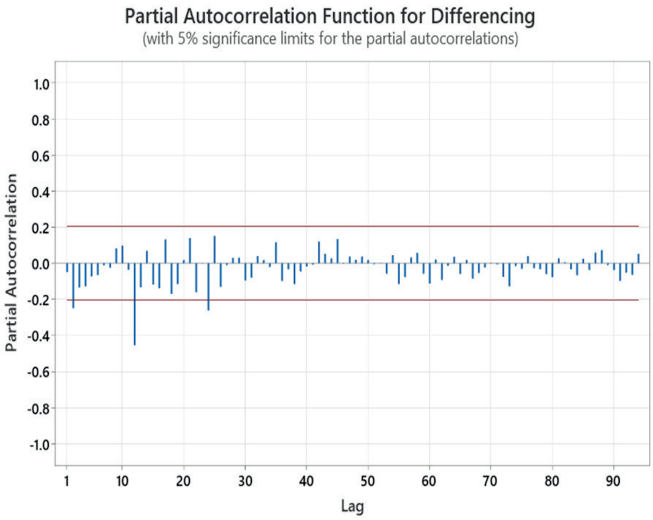

Figure 7. PACF Plot of Data after differencing d = 1 and D = 1.

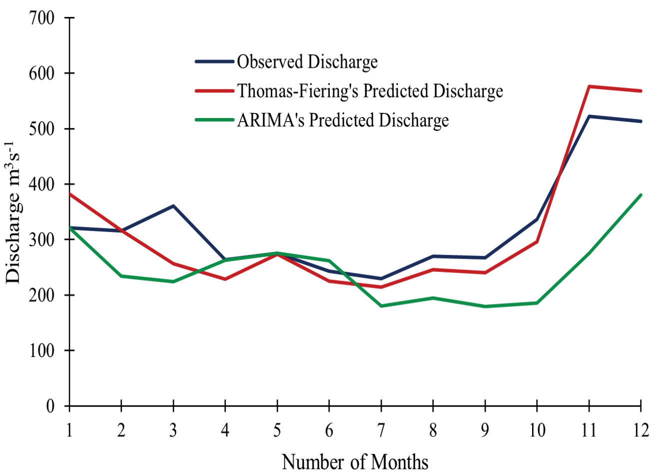

Figure 8. Comparison graph of observed and predicted discharge of scheme I.

From Figure 4 and Figure 5, it was known that the data was not stationary in mean. This could be seen from the pattern of slow decline on the ACF plot and the presence of a patterned lag that still contained seasonality. For this reason, regular differencing (d = 1) and seasonal differencing (D = 1, S = 12) were applied. Then, a temporary model estimation could be carried out by examining the ACF and PACF plots of the data that had become stationary, as shown in Figure 6 and Figure 7.

It could be seen in the PACF plot Figure 7 that it appeared to be dying down, indicating an MA process. The value of order q (MA) observed in the ACF plot (Fig. 7 was significant at lag 2, so it was assumed that q was between 1 and 2. Then, the ACF plot was cut off at seasonal lag 12, so Q = 1 or 2. Several potential ARIMA models could have been formed as ARIMA (0,1,[1,2]) and ARIMA (0,1,[1,2]),(0,1,[1,2])12.

Model parameter estimation was carried out using Minitab 20 software assistance on previously created ARIMA models. After the model parameters were estimated, they were then tested for

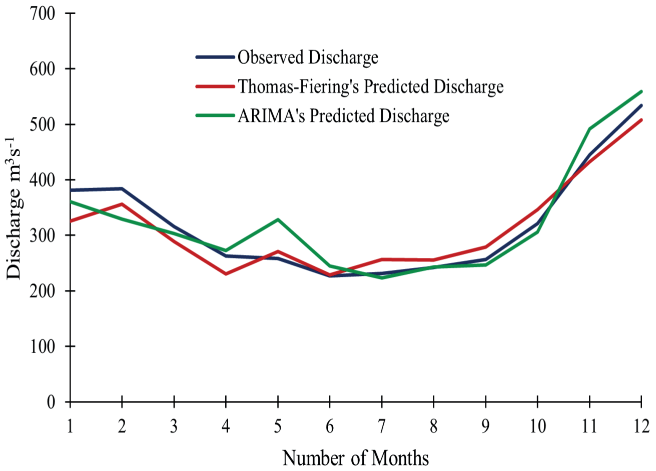

Figure 9. Comparison graph of observed and predicted discharge of scheme II.

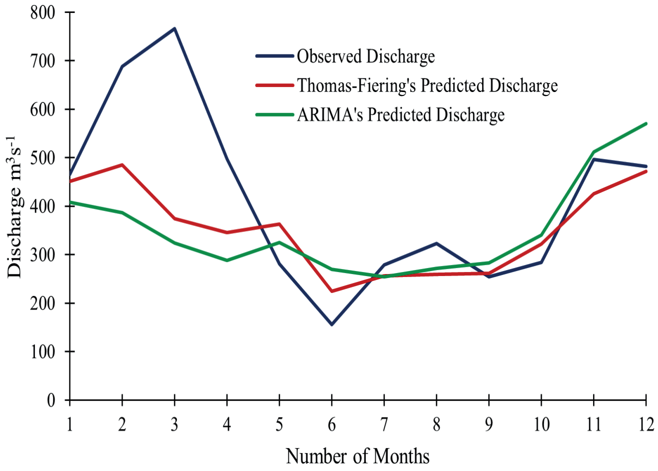

Figure 10. Comparison graph of observed and predicted discharge of scheme III.

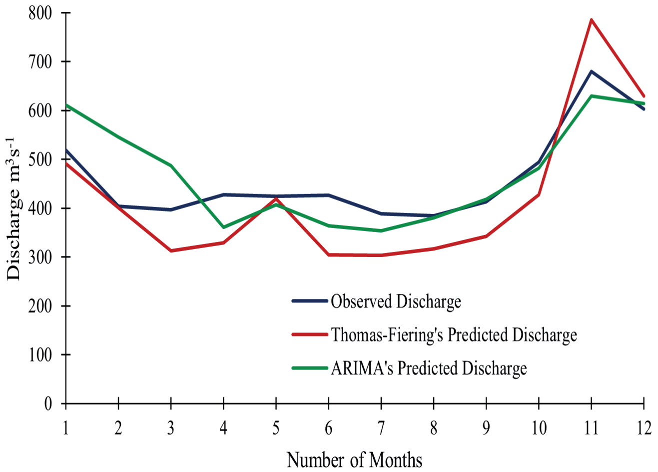

Figure 11. Comparison graph of observed and predicted discharge of scheme IV.

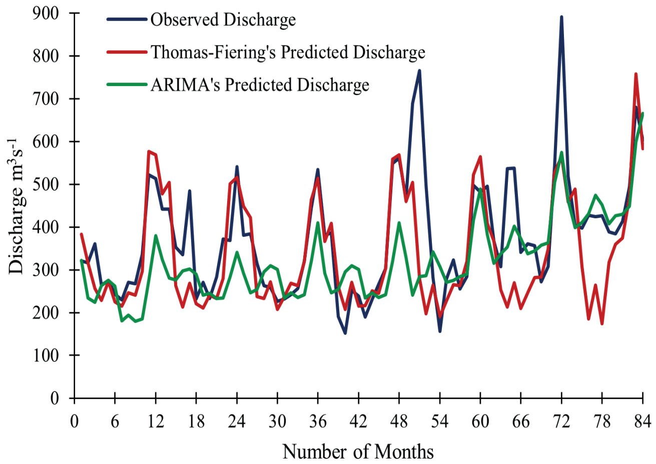

Figure 12. Comparison graph of observed and predicted discharge of scheme V.

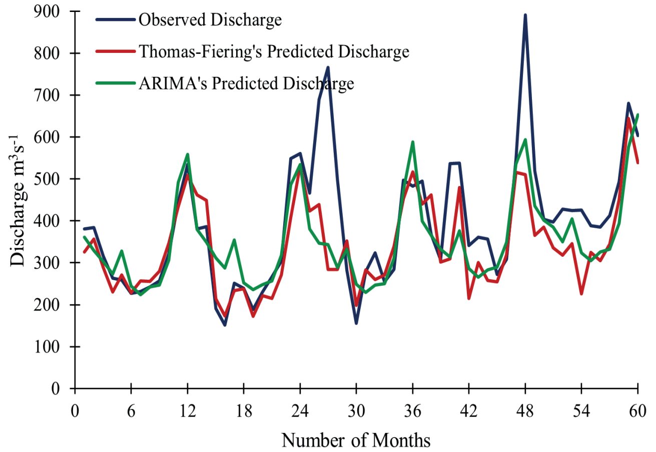

Figure 13. Comparison graph of observed and predicted discharge of scheme VI.

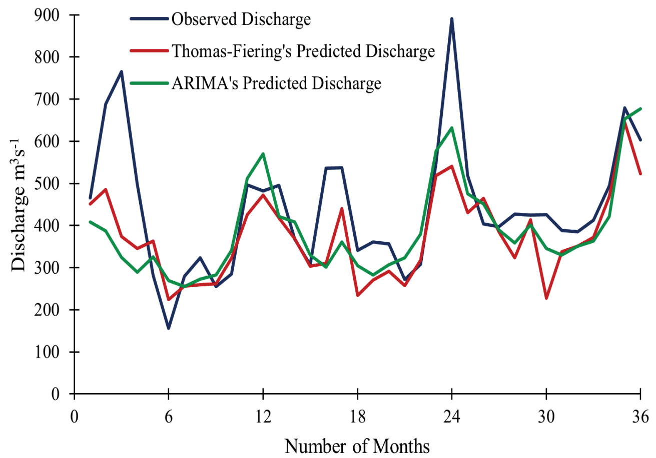

Figure 14. Comparison graph of observed and predicted discharge of scheme VII.

Table 8. Accuracy assessment by performance indicator of Thomas-Fiering method.

Scheme	MAPE (%)	Description	Nash-Sutcliffe	Description	Correlation Coefficient (r)	Description	RMSE (m^3s^{-1})
I	10.68	Good	0.76	Very Good	0.95	Very Good	45.86
II	7.42	Very Good	0.92	Very Good	0.96	Very Good	26.76
III	20.60	Average	0.36	Bad	0.73	Good	141.55
IV	14.18	Good	0.32	Bad	0.97	Very Good	73.90
V	16.89	Good	0.34	Bad	0.70	Good	111.72
VI	15.83	Good	0.44	Bad	0.74	Good	112.13
VII	17.13	Good	0.35	Bad	0.74	Good	121.44

Table 9. Accuracy assessment by performance indicator of ARIMA method.

Scheme	MAPE (%)	Description	Nash-Sutcliffe	Description	Correlation Coefficient (r)	Description	RMSE (m^3s^{-1})
I	22.71	Average	-0.37	Bad	0.58	Average	109.17
II	7.48	Very Good	0.88	Very Good	0.95	Very Good	31.66
III	26.69	Average	0.03	Bad	0.41	Bad	173.78
IV	11.07	Good	0.50	Bad	0.78	Good	64.17
V	21.81	Average	0.18	Bad	0.61	Average	122.10
VI	17.73	Good	0.48	Bad	0.74	Good	98.40
VII	18.97	Good	0.29	Bad	0.64	Average	134.13

significance by examining whether the p-value was <5% or not. The model parameters were considered significant if the p-value was <5%. Following that, the significant ARIMA model was tested with the Ljung-Box test to assess whether the residuals were white noise. If the p-value was >5%, then the model was considered qualified to be used for discharge prediction.

The results of the two tests found that the ARIMA (0,1,2)(0,1,1)12 and ARIMA (0,1,2)(0,1,2)12 passed both tests, then the prediction of discharge was carried out with the two models, with the assessment of MAPE, RMSE, Nash-Sutcliffe, and the correlation coefficient r, it was found that the ARIMA (0,1,2)(0,1,2)12 was more accurate for predicting 1 year discharge (2022) using 9 years of observed discharge (2013-2021). The comparison graph between predicted discharge and observed discharge can be seen in Figure 8 through Figure 14.

3.4. Model Performance Comparison. In Table 8 and 9, it was evident that the ARIMA and Thomas-Fiering methods both had the best predicted discharge results for scheme II, specifically predicting 1 year of discharge using 5 years of observed discharge. To compare the ARIMA and Thomas-Fiering methods, the first compared the results of the assessment per parameter in all schemes, yielding the following results:

1. In the MAPE value of the entire scheme, the Thomas Fiering method led with 1 very good, 5 good, and 1 average, while the ARIMA method achieved 1 very good, 3 good, and 3 average.
2. In the Nash-Sutcliffe value of all schemes, the Thomas Fiering method led with 2 very good and 5 bad, while the ARIMA method achieved 1 very good and 6 bad.
3. In the correlation coefficient (r) of all schemes, the Thomas Fiering method led with ratings of 3 very good and 4 good correlations, while the ARIMA method achieved 1 very good, 2 good, 3 average, and 1 bad correlation.
4. In the RMSE of all schemes, the Thomas Fiering method also performed better, with 5 out of 7 RMSE values across all schemes closer to 0 compared to the ARIMA method, indicating better accuracy.

Then the next comparison was to compare the results of the accuracy assessment on the schemes with the best accuracy assessment results on each method, where the Thomas-Fiering method got the best results on schemes I and II, while the ARIMA method got the best results on schemes II and VI. Based on Table 8 and 9, the Thomas-Fiering method excelled in each parameter, with the acquisition of MAPE ratings of 1 very good and 2 good, Nash-Sutcliffe 2 very good and 1 bad, correlation coefficient 3 very good, and RMSE with 2 of the 3 RMSE values closer to 0 than the ARIMA method.

Some other things that could be seen from the results of the discharge prediction and performance comparison of the two methods that were carried out were as follows:

1. From the prediction of discharge that was done, the Thomas-Fiering method tended to be more accurate for predicting 1-year period discharge using 5-year period observation discharge or shorter.
2. The Thomas-Fiering method was more accurate in long-term prediction than the ARIMA method based on accuracy as-

sessment in scheme V, scheme VI, and scheme VII. This was thought to be because the Thomas-Fiering method was better at predicting peak discharge that occurred outside the seasonal period of 12, while the ARIMA method only focused on the 12 seasonal period. Although the seasonal period was indeed 12, there were times when the data trend changed to form other peaks outside the usual pattern. This could be seen from the comparison graph of observation discharge and predicted discharge in scheme VI and VII, in Figure 13 and Figure 14.

4. CONCLUSION

The Based on the prediction of monthly average discharge that has been carried out on the Siak River (Pekanbaru water gauge station) using two data-driven models which are stochastic models, which are the ARIMA and Thomas-Fiering methods, it is found that both models work best for predicting 1-year discharge using 5 years of observed discharge. Both methods did not perform better the longer the observation data was used, due to the increasingly random pattern and trend of the data as the length of the calibrated observation data period increased. The Thomas-Fiering method is more accurate in predicting discharge above 1 year than the ARIMA method based on accuracy assessment in schemes VI, and VII, as well as comparison graphs, this is because the Thomas-Fiering method is better at predicting peak discharge that occurs outside the seasonal period, which is 12. Based on these points, it is obtained that the Thomas-Fiering method tends to be more accurate for the prediction of monthly average discharge on the Siak River (Pos Duga Air Pekanbaru), which is expected to be the basis for further research in the selection and further modification of data driven discharge simulation models. The limitation of this research is that it is applied to monthly average discharge that is not too fluctuative, the suggestion for further research is to apply data driven models for 15 daily discharge predictions, and can also make performance comparisons between data-driven models and conceptual models.

ACKNOWLEDGEMENTS

The authors sincerely appreciate the invaluable support provided by the Laboratorium Hidro Teknik and the Civil Engineering Department of Riau University, whose facilities and technical assistance were essential to the successful completion of this research. Additionally, the authors extend their gratitude to the peer reviewers for their insightful evaluations, which significantly enhanced the quality of the manuscript.

CREDIT AUTHOR STATEMENT

Manyuk Fauzi: Supervision, Conceptualization, Methodology, Writing - Original draft preparation, Writing-Reviewing and Editing. **Bambang Sujatmoko**: Investigation, Data curation, Formal analysis, Writing-Original draft preparation. **Igeny Dwiana Darmawan**: Visualization, Data curation, Formal analysis. **Siswanto**: Data curation, Writing-Reviewing and Editing. **Ermiyati**: Data curation, Formal analysis. **Merley Misriani**: Formal analysis, Data curation.

DECLARATIONS

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] K. C. Patra, Hydrology and water resources engineering. 2008.

[2] A. Isnandi, M. Fauzi, and I. Suprayogi, Prediksi Debit Aliran Sungai Menggunakan Metode ARIMA (Auto Regresive Integrated Moving Average) Studi Kasus Sungai Tapung Kiri, SAINSTEK, vol. 11, no. 2, pp. 94–101, 2023.

[3] F. K. Aswad, A. A. Yousif, and S. A. Ibrahim, Evaluation the Best Random Component in Modified Thomas-Fiering Model in Generating Rainfall Data for Akre station, Polytech. J., vol. 9, no. 2, pp. 186–192, 2019, doi: 10.25156/ptj.v9n2y2019.pp186-192.

[4] H. Ilva, I. Suprayogi, and M. Fauzi, Analisis Kondisi Hidrologi DAS Siak Bagian Hulu Berdasarkan Peta Tata Guna Lahan Tahun 2014 Menggunakan Model Flow Persistance, J. Tek., vol. 14, no. 1, pp. 22–26, 2020, doi: 10.31849/teknik.v14i1.3465.

[5] L. H. Wijayaratne and P. C. Chan, Synthetic Flow Generation with Stochastic Models, in Flood Hydrology: Proceeding of the International Symposium on Flood Frequency and Risk Analyses, 14–17 May 1986, Louisiana State University, Baton Rouge, USA, Springer, 1987, pp. 175–185.

[6] I. Dashora, S. K. Singal, and D. K. Srivastav, "Software Application for Data Driven Prediction Models for Intermittent Streamflow for Narmada River Basin," Int. J. Comput. Appl., vol. 113, no. 10, pp. 9–17, 2015, doi: 10.5120/19860-1817

[7] P. Gupta and R. Singh, PV Power Forecasting Based on Data-Driven Models: A Review, Int. J. Sustain. Eng., vol. 14, no. 6, pp. 1733–1755, 2021, doi: 10.1080/19397038.2021.1986590.

[8] S. Gunawan, S. E. Wahyuni, and Suharyanto, Kajian Panjang Data Historis yang Representatif pada Model Stokastik, Media Komun. Tek. Sipil, vol. 14, no. 2, pp. 129–141, 2006.

[9] R. Clarke, Mathematical Models in Hydrology. Rome: Food and Agriculture Organization of the United Nation, 1973.

[10] P. Sharma, S. R. Bhakar, Shakir Ali, H. K. Jain, P. K. Singh, and Mahesh Kothari, Generation of Synthetic Streamflow of Jakham River, Rajasthan Using Thomas-Fiering Model, J. Agric. Eng., vol. 55, no. 4, pp. 47–56, 2018, doi: 10.52151/jae2018554.1668.

[11] M. I. Alfa, M. A. Ajibike, and D. B. Adie, Reliability Assessment of Thomas Fiering's Method of Stream Flow Prediction, Niger. J. Technol., vol. 37, no. 3, p. 818, 2018, doi: 10.4314/njt.v37i3.35.

[12] A. A. Yousif, F. K. Aswad, and S. A. Ibrahim, "Performance of ARIMA Model and Modified Thomas- Fiering Model for Predicting the Monthly Rainfall Data for Tallafar Station," vol. 6, no. 1, pp. 293306, 2016.

[13] J. Suryanto, Perbandingan Kinerja Model ARIMA dan

Thomas-Fiering dalam Memprediksi Debit Sungai Loning, Magelang, Agrifor J. Ilmu Pertan. dan Kehutan., vol. 15, no. 1, pp. 65–74, 2016.
[14] P. Suryati and N. Henry, "Random Number Generator Dengan Metode Linear Congruent," Fahma J. Teknol. Inf. Dan Komput., vol. 13, no. 1, pp. 50–60, 2015.
[15] W. W. . Wei, Time Series Analysis (Univariate and Multivariate Methods) Second Edition. Pearson Addison Wesley, 2006.
[16] Salamah, Mutiah, Suhartono, Wulandari, and S. Pingit, Analisis Time Series. Surabaya: FMIPA-ITS, 2003.
[17] Aswi and Sukarna, 2006. Analisis Deret Waktu. Makassar: Penerbit Andira, 2006.
[18] S. Makridakis, S. C. Wheelwright, and V. E. McGee, Metode dan Aplikasi Peramalan. Jakarta: Bina Rupa Aksara, 1999.
[19] W. S. Rahayu, P. T. Juwono, and W. Soetopo, Analisis Prediksi Debit Sungai Amprong dengan Model ARIMA (Autoregressive Integrated Moving Average) sebagai Dasar Penyusunan Pola Tata Tanam, J. Tek. Pengair. J. Water Resour. Eng., vol. 10, no. 2, pp. 110–119, 2019.
[20] P. Gaynor and R. Kirkpatrick, Introduction to Time Series Modelling and Forcasting in Business and Economics. Singapore: Mc Grow Hill, 1994.
[21] R. Yuliyanti and E. Arliani, Peramalan Jumlah Penduduk Menggunakan Model ARIMA, Kaji. dan Terap. Mat., vol. 8, no. 2, pp. 114–128, 2022.
[22] M. Efendi, W. Soetopo, and P. T. Juwono, Pengaruh Panjang dan Lebar Data Debit Historis pada Kinerja Model Pembangkitan Data Debit Sungai Brantas dengan Metode ARIMA, J. Tek. Pengair., vol. 7, no. 1, pp. 37–46, 2016.
[23] H. Panjaitan, A. Prahutama, and S. Sudarno, Peramalan Jumlah Penumpang Kereta Api Menggunakan Metode Arima, Intervensi dan Arfima (Studi Kasus: Penumpang Kereta Api Kelas Lokal Ekonomidaop IV Semarang), J. Gaussian, vol. 7, no. 1, pp. 96–109, 2018.
[24] B. Putro, M. Tanzil Furqon, and S. H. Wijoyo, Prediksi Jumlah Kebutuhan Pemakaian Air Menggunakan Metode Exponential Smoothing (Studi Kasus: PDAM Kota Malang), J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 11, pp. 4679–4686, 2018.
[25] P.-C. Chang, Y.-W. Wang, and C.-H. Liu, The Development of a Weighted Evolving Fuzzy Neural Network for PCB Sales Forecasting, Elsevier, 32 (Expert Syst. with Appl., pp. 86–96, 2007.
[26] M. Mahyudin, I. Suprayogi, and T. Trimaijon, Model Prediksi Liku Kalibrasi Menggunakan Pendekatan Jaringan Saraf Tiruan (ZST)(Studi Kasus: Sub DAS Siak Hulu). Riau University, 2014.
[27] Indarto, Hidrologi: Dasar Teori dan Contoh Aplikasi Model Hidrologi. Jakarta: Bumi Aksara, 2010.
[28] L. Alby and E. Suhartanto, Perbandingan Metode Alih Ragam Hujan Menjadi Debit dengan FJ. MOCK dan NRECA di DAS Kemuning Kabupaten Sampang, J. Tek. Pengair., vol. 2, no. 1, pp. 1–8, 2018.
[29] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, Trans. ASABE, vol. 50, no. 3, pp. 885–900, 2007.
[30] R. A. Wibowo and A. A. Kurniawan, Analisis Korelasi dalam Penentuan Arah antar Faktor pada Pelayanan Angkutan Umum di Kota Magelang, Theta Omega J. Electr. Eng. Comput. Inf. Technol., vol. 1, no. 2, pp. 45–50, 2020.
[31] K. H. Hinkle et al., Phoenix spectrograph at Gemini South, in Discoveries and Research Prospects from 6-to 10-Meter-Class Telescopes II, SPIE, 2003, pp. 353–363.