International Journal of Innovative Research and Scientific Studies, 8(3) 2025, pages: 4176-4188

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

A comparison of Markov Chain, Sarima and multiple linear regression for forecasting discharge

Ira Mulyawati[1*], Indah Rachmatiah Siti Salami[2], Mariana Marselina[3], Arwin Sabar[4]

[1,2,3,4]*Environmental Engineering Program, Institut Teknologi Bandung, Indonesia.*

Corresponding author: Ira Mulyawati (*Email: iramulyawati@gmail.com*)

Abstract

Developing a hydrological forecasting model based on past records is crucial for effective hydropower reservoir management and scheduling. Numerous popular discharge forecasting models have been developed; however, real-time forecasts remain challenging. This study evaluates discharge forecasts using the Markov Chain model, Seasonal Autoregressive Integrated Moving Average (SARIMA), and Multiple Linear Regression (MLR) models for forecasting monthly discharge time series. This study compares the accuracy of the discharge forecast results produced by the Markov Chain, SARIMA, and Multiple Linear Regression using five statistical indicators. Based on the simulation results, the Markov Chain, SARIMA, and MLR have accuracy levels of probability in discharge of 63%, 66%, and 76%, respectively. In comparison to other models, the highest correlation (r) is found in the MLR model (0.76) with MAPE (0.19), followed by SARIMA and Markov Chain. Therefore, the most accurate, precise, and representative water source model alternative for forecasts is the MLR model. The Markov Chain model and the SARIMA model are time series generation models, while the MLR model is a statistical regression model. In addition, this model is to be selected as the basis for modeling in forecasting river flow or optimal management of a reservoir, as well as determining future discharge, especially in monsoon climate regions.

Keywords: Forecasting discharge, Markov chain, Multiple linear regression (MLR), SARIMA.

DOI: 10.53894/ijirss.v8i3.7465
Funding: This study received no specific financial support.
History: Received: 8 April 2025 / **Revised:** 13 May 2025 / **Accepted:** 15 May 2025 / **Published:** 30 May 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Text Body

The identification of suitable models for forecasting future monthly inflows to hydropower reservoirs is a significant precondition for effective reservoir management and scheduling. The results, especially in long-term predictions, are useful in many water resources applications such as environmental protection, drought management, operation of water supply utilities, and optimal reservoir operation involving multiple objectives of irrigation, hydropower generation, and sustainable development of water resources. As such, hydrologic time series forecasting has always been of particular interest in operational hydrology. It has received tremendous attention from researchers in the last few decades, and many models for hydrologic time series forecasting have been proposed to improve hydrology forecasting [1-3].

These models can be broadly categorized into three groups: regression-based methods, time series models and AI-based methods. Multiple linear regression (MLR) analysis is among the most widely utilized statistical techniques [3-5]. A regression model that involves more than one regressor variable is called a multiple linear regression model [5]. Multiple linear regression modelling has been widely utilized for modelling such as urban runoff pollutant load, wash load sediment concentrations, suspended-sediment discharge, and prediction of swell potential of clayey soils [6-8]. The is suitable for predicting discharge data, as well as for optimal reservoir management [3-5].

Markov chain models, often recognized as autoregressive stochastic models, are widely regarded as effective tools for discharge forecasting, particularly when the discharge at a given time depends on the flow from previous periods. Over time, these models have undergone significant improvements, making Markov chains more reliable for predicting both river and reservoir discharges [9, 10].

Markov Chain is used to evaluate discharge predictability by leveraging key hydrological characteristics. However, achieving accurate forecasts requires advanced data processing capabilities to refine the model's performance. Recent studies highlight the suitability of Markov chain models for estimating inflow discharge, tackling challenges like managing excess water discharge and supporting optimized reservoir operations. Built on established frameworks such as the French EDF model, these models prove to be essential for predicting discharge patterns and promoting more efficient, sustainable reservoir management [10-12].

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a powerful tool in hydrology, widely used to forecast time series data that show both trends and seasonal patterns — such as river inflows, rainfall, and reservoir levels. Its ability to account for recurring climatic cycles makes it especially effective for modeling hydrological processes influenced by seasonal changes. Research has shown SARIMA's effectiveness in predicting monthly river inflows, daily monsoon rainfall, and groundwater levels, providing reliable forecasts that are essential for managing water resources, controlling floods, and mitigating droughts. Additionally, SARIMA's flexibility allows it to be combined with advanced machine learning models, further improving its accuracy when dealing with complex hydrological systems [13-15].

2. Method

The study was conducted at the Ciberang watershed, the Ciujung sub-watershed, in Lebak Regency, Banten Province (Figure 1). The Ciberang watershed covers a catchment area of 458 ㎢. The Ciberang River, originating at the Karian dam site and extending 78.3 km from the mountain's summit, flows through the central Karian basin and converges with the Ciujung River at Rangkasbitung [16].

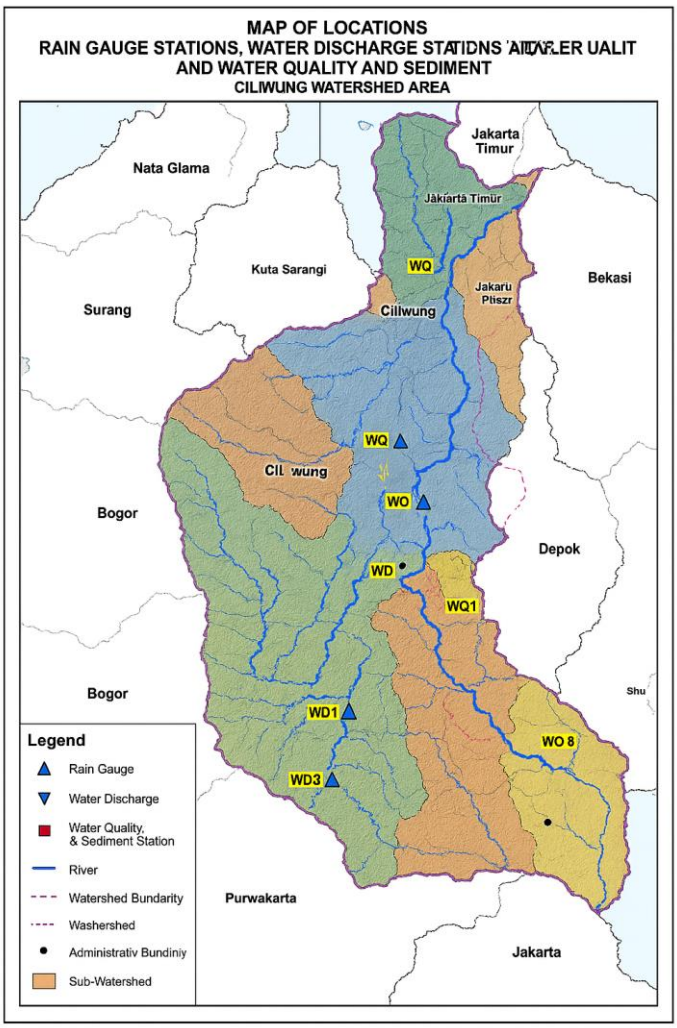

Figure 1.
Map Location of Ciberang-Karian watershed area in Indonesia.

This study utilized discharge (Q) and rainfall (P) data acquired directly from existing records maintained by pertinent agencies, including the office of the Cidanau-Ciujung-Cidurian River Area (BBWSC3) and the Korean International Cooperation Agency (KOICA) [16].

All the discharge data Figure 1 utilized data recorded from the discharge gauge station Q1 Bojongmanik (6º 34' 33.031" S- 106º 10' 9.354" E), Q2 Leuwidamar (6º 30' 49.982" S- 106º 11' 37.136" E), Q3 Jahe Cilaki (6º 25' 47.568" S -106º 14' 26.664" E), Q4 Karian (of 5°50′ 7°10′S-105°48′ 07°28′ E), Q5 Sabagi (6º 23' 50.91" S-106º 15' 14.58" E), Q6 Jembatan Keong (6º 21' 30.661" S -106º 14' 34.523" E), Q7 Jembatan Rangkas (6º 20' 55.223" S-106º 14' 49.578" E). In this study, Karian Q4 (1997-2023) input discharge data was used in Table 1 [16].

The input discharge data in this study was built upon our prior study published in E3S Web of Conf. Volume 485, 2024. For the input discharge data from previous research, this will be referred to as historic discharge. Furthermore, this research aims to deepen the understanding of forecasting input discharge, while the previous study aimed to deepen the understanding of generating input discharge.

Table 1.
Water discharge data at Karian (1997–2023).

Year	January	February	March	April	May	June	July	August	September	October	November	December
1997	20	13	16	17	18	6	4	6	4	5	9	10
1998	12	20	21	20	18	10	11	4	10	16	18	16
1999	43	23	16	11	10	8	8	5	7	12	5	16
2000	19	35	15	9	10	9	9	5	8	8	18	9
2001	20	36	17	12	12	10	11	8	10	11	12	7
2002	12	23	13	18	10	8	10	6	10	5	9	13
2003	12	28	13	15	18	9	8	3	10	9	9	37
2004	10	19	15	15	10	4	9	4	8	8	21	12
2005	16	33	15	14	18	14	14	10	8	9	19	15
2006	19	34	17	14	13	9	7	7	7	14	14	13
2007	20	29	16	14	13	6	8	9	15	16	16	20
2008	18	26	17	17	13	8	9	8	7	20	14	16
2009	13	18	12	18	14	10	9	10	4	18	10	15
2010	13	21	11	9	8	9	13	16	11	23	9	14
2011	18	20	19	16	14	7	10	6	5	11	19	15
2012	26	19	14	15	10	8	7	4	6	10	13	17
2013	45	22	13	15	17	10	13	10	9	9	6	15
2014	25	25	13	14	14	8	11	11	8	13	21	13
2015	22	28	13	16	16	9	8	4	4	4	12	18
2016	20	30	15	14	10	12	13	9	4	7	5	22
2017	20	27	14	13	13	12	12	9	8	14	12	13
2018	13	23	11	11	10	7	7	6	7	5	13	10
2019	15	19	15	12	8	8	7	6	10	7	8	15
2020	8	22	11	10	5	8	8	6	6	10	6	34
2021	16	28	10	15	8	8	10	8	12	7	5	12
2022	8	20	16	8	11	13	10	10	7	9	6	23
2023	20	22	12	7	11	6	10	7	6	8	12	8

All the rainfall data Figure 1 utilized data recorded from the Rainfall gauge station of P1 Banjar Irrigation (6º 34' 9.12" S-106º 24' 39.96" E), P2 Ciminyak Cilaki (6º 32' 22.92" S-106º 18' 29.16" E), P3 Sampang Peundeuy (6º 30' 6.12" S-106º 11' 21.84" E), P4 Sajira (6º 29' 58.6" S- 106º 21' 57.38" E), P5 Cimarga (6º 25' 26.04" S- 106º 14' 7.08" E) and P6 Pasir Ona (6º 22' 9.84" S-106º 15' 56.88" E). The rainfall data used is in the range of 1997-2023 (Figure 1) [16].

2.1. Rainfall Analysis

For the analysis of the rainfall data used in this study, monthly rainfall data were recorded at nine selected rain gauge stations (Figure 1). The first step for processing rainfall data is filling in rainfall data and testing consistency for data adjustment [17, 18]. The second stage is to calculate the regional rainfall using the arithmetic method [19].

Water Source Model Analysis

2.2. Multiple Linear Regression (MLR)

The multiple linear regression model can be utilized to fill data gaps, generate, and forecast data. The MLR model examines the relationship between rainfall and discharge, employing multiple linear regression statistics to forecast input discharge [3-5].

In multiple linear regression analysis, variables are categorized as dependent and independent. These variables need to be assessed on an MLR scale (i.e., an interval or ratio variable) [3-5]. Therefore, in this research, the Karian input discharge generation model uses multiple linear regression analysis, referred to as the MLR model.

Multiple linear regression includes binary, ternary, and quaternary models. The most suitable model was selected from these various models based on the one with the highest determination value. Previous research indicates that the quaternary variation model has a higher determination value compared to the ternary and binary variations. This method is preferred due to its dynamic statistical nuances, which capture variations in hydrological components, such as rainfall and discharge [3-5, 20, 21].

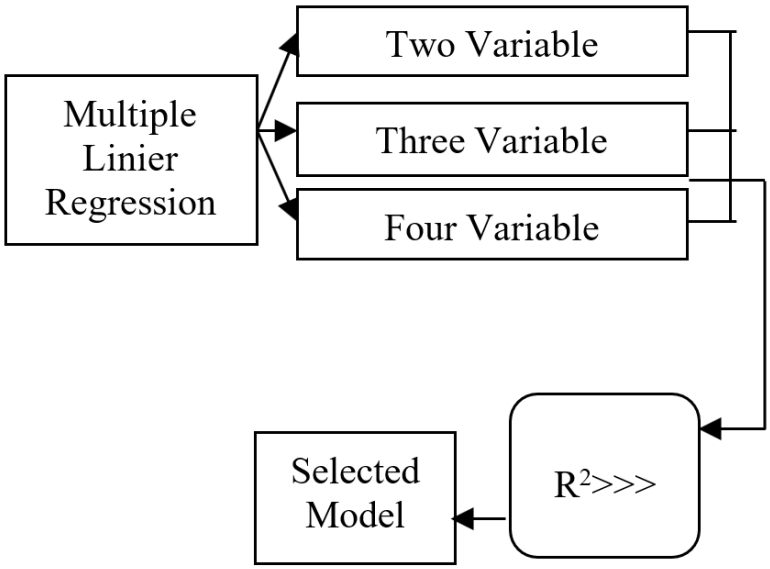

Figure 2.
Multiple Linear regression of rainfall models.
Source: Sabar [22]**,** Marselina et al. [23] **and** Dar [24]

2.3. Formula and Calculation of Multiple Linear Regression Rainfall-Discharge Dar [24]**;** Patel, et al. [25] and Gupta and Kumar [2]*:*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \cdots + \epsilon \quad (5)$$

Where, for *i*=*n* observations:

yi = dependent variable (monthly average discharge at Reservoir)

xi = explanatory variables (monthly average discharge at neighboring stations)

zi = explanatory variables (monthly average at rainfall neighboring stations)

*β*0 = y-intercept (constant term)

β1, β2 =slope coefficients for each explanatory

variable

ϵ = the model's error term (also known as the residuals)

2.3.1. Seasonal Autoregressive Integrated Moving Average (SARIMA)

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an advanced version of the ARIMA model, designed to handle time series data that show both regular (non-seasonal) trends and recurring seasonal patterns. It's represented as ARIMA(p,d,q)(P,D,Q)[s], where Eshragh et al. [26], Perez-Guerra et al. [27] and Lee and Kim [28]:

- p: Non-seasonal autoregressive order (how past values influence the present)
- d: Non-seasonal differencing (to make the data stationary)
- q: Non-seasonal moving average order (how past errors influence the present)
- P: Seasonal autoregressive order (captures seasonal effects from past values)
- D: Seasonal differencing (removes seasonal trends)
- Q: Seasonal moving average order (captures seasonal error patterns)
- s: The length of the seasonal cycle (e.g., 12 for monthly data with yearly seasonality)

This setup allows SARIMA to model time series data that displays both short-term fluctuations and longer-term seasonal behaviors. The model's multiplicative structure comes from combining non-seasonal and seasonal components. Mathematically, it is expressed [29, 30]:

$$\Phi_p(\mathrm{B})\Phi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \Theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Where :

$\Phi_{\mathrm{p}}(\mathrm{B})$ and $\Theta_{\mathrm{q}}(\mathrm{B})$ represent the non-seasonal autoregressive and moving average terms.

$\Phi_{\mathrm{p}}(\mathrm{B}^{\mathrm{s}})$and $\Theta_{\mathrm{Q}}(\mathrm{B}^{\mathrm{s}})$ capture the seasonal autoregressive and moving average parts.

$(1-\mathrm{B})^{\mathrm{d}}$ handles non-seasonal differencing, and $(1-B^s)^D$ manages seasonal differencing.

Here are the steps for modeling Seasonal Autoregressive Integrated Moving Average (SARIMA) [29, 30]:

1. Identify the Model: Start by exploring the time series data to spot trends, patterns, and seasonal cycles. If the data in level origin is stationary, then use the SARMA model, and if the data needs differencing, then use the SARIMA model.

Next, use visual tools like the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to help determine the right non-seasonal (p, d, q) and seasonal (P, D, Q) parameters.

2. Estimate the Parameters**:** Once you've identified the model structure, estimate the model's parameters using statistical techniques. This step involves fitting the SARIMA model to the data and finding the best coefficients that describe the observed patterns.
3. Check the Model's Fit: After fitting the model, evaluate how well it performs. Analyze the residuals (the differences between the model's predictions and actual data); they should look like random "white noise" with no remaining patterns. If the residuals show structure or correlation, it's a sign the model may need adjusting.
4. Make Forecasts: Once the model checks out, use it to make predictions. Compare the forecasts to actual data to see how accurate they are. If the performance isn't satisfactory, revisit the previous steps and tweak the model to improve its accuracy.

Following these steps provides you with a structured, reliable way to build SARIMA models, making it easier to analyze time series data and capture those tricky seasonal patterns for better forecasting.

2.3.2. Markov Chain Models

The Markov Chain model is one of the stochastic models that utilizes a time series of discrete variables [9-12]. It is essential to conduct a reliable discharge analysis of the discharge's reliability before implementing the Markov Chain generation model using the Weibull method. The Weibull probability formula calculates the probability (after data has been sorted from largest to smallest) of the event being greater than the discharge value [10, 31, 32]. And is given below.

$$P\ (X \geq x) = \frac{m_x}{n+1} \quad (2)$$

Figure 1 Ciberang Watershed research location (Office of Cidanau-Ciujung-Cidurian River Area BBWS-C3)

P ($X \geq x$) is the value of the probability occurrences of all events 'X' greater than or equal to data 'x', m_x represents the ranking of data x after sorting from largest to smallest, and n denotes the total amount of data.

The probability of data reliability from several data points is the probability value of an event where the value that occurs is equal to or above the data value. Furthermore, the calculated reliability is categorized as dry, normal, or wet discharge. After calculating the discharge reliability using the Weibull method, the discharge generation analysis is carried out using the Markov Chain method. In the Markov Chain process, the probability at a certain time is determined only from the events of the previous time [10, 11]. The formula of the Markov Chain model is given below [10, 11, 32, 33]:

$$qi = di + ei \quad (3)$$

d_i represents the deterministic component and e_i indicates the random component.

The simplification of water discharge in Markov Chain models was carried out by categorizing it into three classes [17]. In this model (three-class Markov), a stochastic matrix that classifies the historical data into three classes: Dry discharge (represented by 0), Normal discharge (represented by 1), and Wet discharge (represented by 2), can be created monthly. The class intervals for each class division were obtained by dividing the probability curve of the distribution of the selected population into three equal parts, namely 0.333, 0.667, and 1. The guidelines' trajectory was determined with the concept of planned discharge by examining the behavior of historical water discharge and subsequently determining the threshold for the magnitude of future water flow events [10, 11].

2.4. Statistical Analysis

To assess the simulation discharge generation models, a comparison was conducted utilizing the value of correlation, MAPE [3-5, 27]. The correlation coefficient indicates the extent of association in regression analysis with a cause-and-effect relationship, ranging from 1 to -1, where 1 signifies a strong positive correlation, 0 denotes no correlation, and -1 indicates a strong negative correlation [3-5]. Meanwhile, MAPE, or Mean Absolute Percentage Error, is a way to measure how well a model's predictions match reality, expressed as an easy-to-understand percentage. In hydrological modeling, it plays an important role in evaluating how accurately the model simulates phenomena such as river flow, rainfall, or water levels. Simply put, the lower the MAPE, the better the model's performance. A low MAPE means the model's predictions are close to the actual data, making it a reliable tool for identifying models that are underperforming or overfitted. In this study, each model's accuracy level was determined using the formula below [3-5, 34, 35]:

$$\boldsymbol{r} = \frac{(\boldsymbol{n})(\sum \boldsymbol{xy}) - (\sum \boldsymbol{x})(\sum \boldsymbol{y})}{\sqrt{(\boldsymbol{n}.\sum \boldsymbol{x^2} - \sum(\boldsymbol{x})^2.(\boldsymbol{n}.\sum \boldsymbol{y^2} - \sum(\boldsymbol{y})^2))}} \quad (6)$$

$$MAPE = \frac{\mathbf{1}}{\boldsymbol{n}} \mathrm{x} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left| \frac{\boldsymbol{Q_{obs} - Q_{mod}}}{\boldsymbol{Q_{obs}}} \right| \quad (7)$$

MAPE < 10% → Excellent forecasting accuracy
10% ≤ MAPE < 20% → Good forecasting accuracy 20% ≤ MAPE < 50% → Reasonable or Moderate forecasting accuracy
MAPE ≥ 50% → Poor forecasting accuracy

x represents the mean value of x_i, y indicates the mean value of y_i, MAPE stands for the average error rate, n denotes the Number of data, $_i$ symbolizes the month-I, Q_{obs} represents the historical discharge, Q_{mod} symbolizes the discharge models.

3. Result and Discussion

3.1. Rainfall Analysis

The rainfall data used in this study is the monthly average rainfall data from 1997-2023 of six stations in the Ciujung watershed, Karian. Figure 2 shows the trend of average rainfall from six rain gauge stations at the Ciujung-Karian watershed. Based on an assessment of the average rainfall as shown in Figure 4, it can be proven that this research area is a Monsoon zone type, with rainfall characteristics that have two peaks of rain concentrated in the rainy season (October to May), while rainfall occurs in the dry season (June to September). The recapitulation of the average rainfall from 1997 to 2023 can be seen in Table 2.

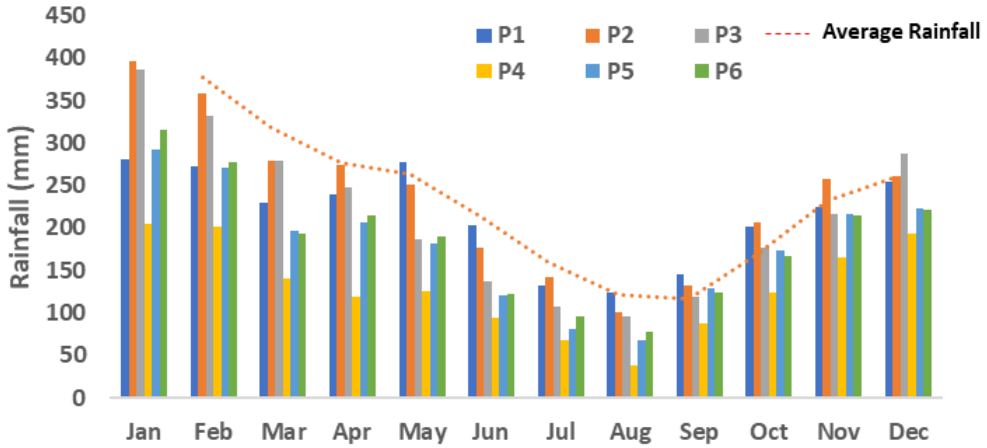

Figure 3.
The average monthly rainfall and regional rainfall at the observation locations.

Table 2.
Recapitulation of average rainfall (mm).

Month	P1	P2	P3	P4	P5	P6
Jan	281	396	386	204	292	315
Feb	272	358	332	201	271	277
Mar	230	280	279	140	196	194
Apr	239	274	247	118	207	214
May	278	251	187	125	181	190
Jun	204	176	137	94	121	123
Jul	133	141	108	68	81	96
Aug	124	101	96	38	68	77
Sep	145	132	119	88	129	124
Oct	201	206	177	123	174	166
Nov	224	257	217	166	217	214
Dec	255	262	287	193	223	221
Average	216	236	214	130	180	184

3.2. Water Sources Model

Figure 4, Figure 5, Figure 6, depicts the simulation results of historical data discharge comparisons with Markov Chain, Autoregressive Moving-Average models SARIMA and Multiple Linear Regression (MLR) models' predictions. Each water source model has advantages; for example, the Markov Chain, Seasonal Autoregressive Integrated Moving-Average models SARIMA and Multiple Linear Regression (MLR) models can predict discharges in both river catchments and dams. The Markov Chain, SARIMA, and MLR models are also appropriate for generating discharge data and predicting future flow rates [3-5, 11, 26, 27].

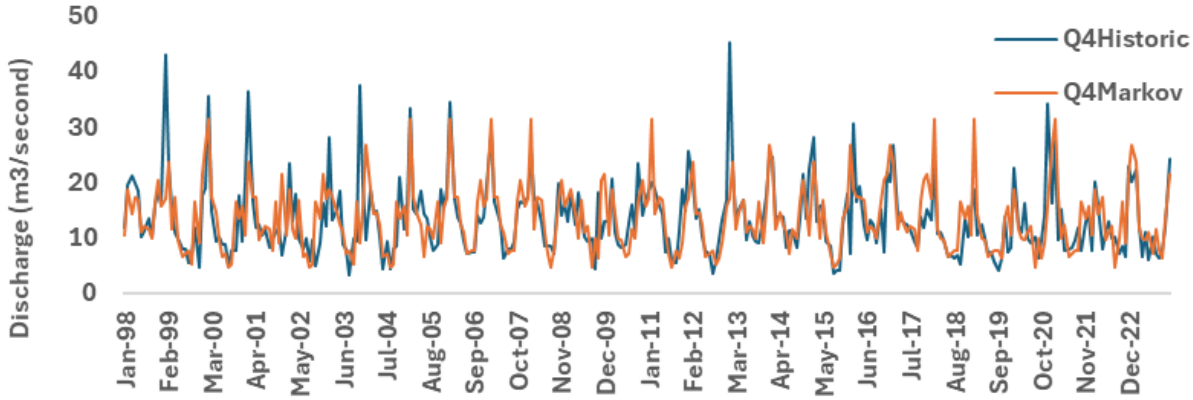

Figure 4.
The simulation results of historic discharge compared to the Markov Chain models.

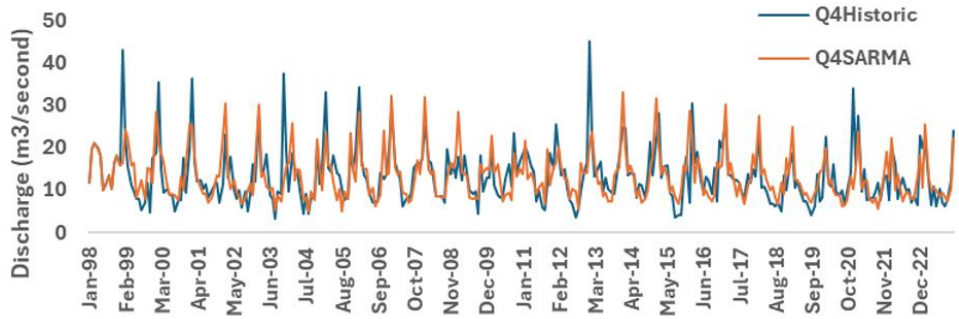

Figure 5.
The simulation results of historic discharge compared to the SARIMA.

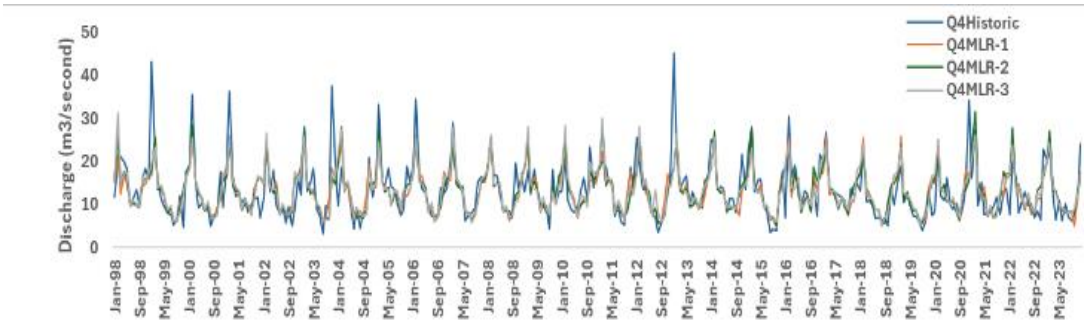

Figure 6.
The simulation results of historic discharge compared to the MLR-1, MLR-2, MLR-3.

3.2.1. Markov Chain Models

In the Markov process, the probability at a certain time is determined only from previous events. The Markov Chain process is as follows [36]:

1. Data is sorted from smallest to largest. Calculate the probability value for each data point using the Weibull method (p=m/(N+1))

2. The cumulative probability for each month's data series is classified into three classes, namely with probability intervals as follows:

- Class 0 = dry

p = 0 – 1/3

- Class 1 = normal

p = 1/3– 2/3

- Class 2 = wet

p = 2/3 – 1

3. From the data range (dry, normal, wet), look for the average of the three results, namely dry, normal, and wet discharge.

4. Next, a three-class first-order monthly transition matrix is created for dry, normal and wet monthly discharge. example of a monthly transition matrix for the month of February, Table 3.

Table 3.

Class 3rd order monthly Markov Chain transition matrix (dry, normal and wet) Jan to Feb.

Monthly discharge Jan	Monthly discharge Feb				
	0	1	2		
0	0.56	0.22	0.22	1.00	P0N
1	0.22	0.22	0.56	1.00	PIN
2	0.22	0.56	0.22	1.00	P2N
	1.00	1.00	1.00	3.00	PNN
	P0N	P1N	P2N	PNN	

5. Next, perform monthly discharge data prediction based on the monthly probability of 3 classified categories that have been calculated.

The value of the coefficient correlation between the historical discharge comparison and the estimated discharge simulation of the Markov Chain model is 63% (Figure 5). Based on the correlation value, the correlation value of the SARIMA forecast discharge with the generation discharge is a moderate relationship. It shows that this model is capable of predicting an increase or decrease in discharge of up to 63%. Based on the simulation of Markov Chain models, the MAPE values are 26%. The MAPE value in the SARIMA model is 25%, which shows reasonable or moderate forecasting accuracy. This shows that the model is good at predicting monthly discharge [3-5, 34, 35].

The Markov Chain model can predict monthly average discharge because the model predicts discharge using the probability method of possible past discharges, which often appear to predict the future. Moreover, the discharge is classified into three classes: dry, normal, and wet, which makes the model effective in predicting data with uniform patterns but has limitations in forecasting extreme data. Extreme data can disrupt normal distribution, and the Markov Chain model is most effective for forecasting discharge when the data approximates a normal and uniform distribution [37].

3.2.2. Seasonal Autoregressive Integrated Moving Average (SARIMA)

ARIMA is a way to estimate time series data stochastically. The ARIMA model represents three models, namely from Moeeni et al. [13], Yang et al. [14] and Kenyi and Yamamoto [15]:

1. Autoregressive model (AR)
2. Moving averages (MA)
3. Autoregressive and moving average model (ARMA)

SARIMA method, which was originally a development of the Box-Jenkins (ARIMA) method. The SARIMA model can overcome the seasonal pattern of a time period. Based on Figure 7 illustrates that the discharge data typically decreases in January each year and increases in the subsequent month, and this pattern repeats annually, reflecting the impact of seasonality, therefore, the ARIMA model is developed into the SARIMA model. Based on Figure 7, the Karian discharge data from 1997-2023 is stationary (Figure 7 and Table 3), therefore, there is no need for differencing. If the data is stationary in the first difference, then modeling using ARMA (p,q) d shows the degree of stationarity. Based on plotting data and stationary tests, stationary discharge data at the origin level, the model chosen is the ARMA model [26-28]:

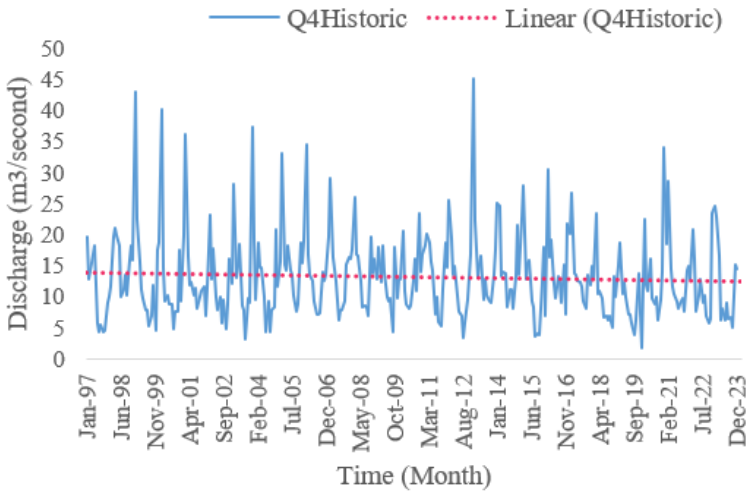

Figure 7.
Karian input discharge plots (1997-2023).

The stationary test uses the Dickey-Fuller method with the following hypothesis and results [38]:

H0: data is not stationary

H1: stationary data

P-value < 0.05 → reject H0 →Conclusion: data is stationary Table 4.

Table 4.
Stationary Test.

Level	Dickey-Fuller	P-Value	Conclusion
Origin	-10.709	0.01	Stationary

The next stage after the stationary test is to estimate the model. This will determine the initial estimated values for the parameters of the ARMA model based on the autocorrelation correlogram (ACF) and partial autocorrelation correlogram (PACF) patterns. From the results of the model selection, the best model was then re-selected, using statistical tests [26-28].

From the ACF, PACF, and statistical test results, the best model was SARIMA (2,0,2) (2,1,0) with a drift model. For this reason, the Karian input discharge forecast uses the SARIMA (2,0,2) (2,1,0) with a drift model. After obtaining the best model, a prediction of Karian's input discharge from 1997 to 2023 was carried out, and then calibration was performed using the historic data (Figure 6) [26-28].

Based on the calibration results of the SARIMA input discharge forecast and historic data, the correlation value is 65%. Based on the correlation value, the correlation of the SARIMA forecast discharge with generation discharge has a moderate relationship. This shows that this model is capable of predicting an increase or decrease in discharge of up to 65%. The MAPE value in the SARIMA model is 25%, which shows reasonable or moderate forecasting accuracy. This indicates that the model is good at predicting monthly discharge [3-5, 34, 35].

A SARIMA model works well for predicting discharge data like river flow because it captures both short-term fluctuations and seasonal patterns that naturally occur in this kind of data. Discharge often follows a seasonal rhythm; for example, rivers might swell during rainy seasons and then slow down in drier months. SARIMA models are built to handle this by including seasonal components (P, D, Q, s), which recognize and adjust for these repeating cycles, whether they happen daily, monthly, or yearly. By combining trend detection, seasonal behavior, and random variations, SARIMA becomes a powerful tool for understanding and forecasting discharge patterns, even when the data is noisy or unpredictable [13-15, 26-28].

3.3. Multiple Linear Regression (MLR) Model

In the Multiple Linear Regression (MLR) model, there are various model configurations. The selection of the forecasting discharge model is conducted using several variations of the quaternary MLR model. This approach is based on previous studies showing that the four-variable (quaternary) variation produces a higher correlation value compared to the two-variable (binary) and three-variable (ternary) variations. Moreover, using more than four variables tends to yield correlation values that are not significantly different from those produced by the four-variable variation [3-5, 17, 18].

For the quaternary MLR variations in this study, three variations were created: MLR-1 variation of Q4 (Q4t-1 P1t-1 P4t-1); MLR-2 variation of Q4 (Q4t-1 Q4t-2 P1t-1). MLR-3 variation of Q4 (Q4t-1 Q4t-2 P4t-1).

The selection of these variations is based on the hydrological components present in the upstream area of the Karian Reservoir, specifically P1, P4, and Q4 (Figure 1). This approach was chosen because using hydrological components from the downstream area of the Karian Reservoir would not align with the characteristics of Karian's input discharge. This discrepancy occurs due to changes in hydrological characteristics downstream of the Karian Reservoir, where the flow has been influenced by the reservoir itself. As a result, the downstream hydrological characteristics differ from those found in the upstream components of the Karian Reservoir.

Additionally, the model variations are also based on differences in the time periods of rainfall and discharge recordings. There are time step variations of t-1, which predict the value for period t, and t-2 time steps, which predict the discharge value for period t. This approach aims to evaluate how strongly the time factor influences future prediction results [3-5].

The MLR-1 model (Figure 7) has a correlation value of 76%, it shows that the relationship between the Q4 MLR-1 forecast discharge and historic data has a high correlation. Meanwhile, the MAPE value in the MLR-2 is 19%, which shows good forecasting accuracy. This shows that the model is good at predicting monthly discharge [3-5, 34, 35].

The MLR-1 model is effective at predicting discharge; this is because the MLR-1 model Q4 (Q4t-1P1t-1P4t-1) uses t-1 month rainfall data and t-1 month discharge data. The variables t-1 month rainfall and t-1 month discharge have a strong correlation with t month historic discharge data. This is because the rainfall in the previous month (t-1) becomes the discharge in month t, and the relationship between discharge t and discharge t-1 is also strong. The strong correlation between a river's discharge in the current month (Q_t) and its discharge in the previous month (Q_{t-1}) can be attributed to the inherent persistence in hydrological systems. This persistence arises because various hydrological processes—such as groundwater contributions, soil moisture retention, and basin storage effects—exhibit temporal continuity, causing river discharge to be autocorrelated over time.

A study examining river discharge variability across major global basins highlighted that the variations in river discharge are influenced by climate and the nature of the Earth's surface and subsurface. These factors contribute to the temporal persistence observed in discharge records, leading to significant correlations between consecutive monthly discharges. These factors result in a good discharge forecast [2, 26].

The MLR-2 model Figure 7 it has a correlation value of 75%, which shows that the relationship between the Q4 MLR-2 forecast discharge and historic data has a high correlation. Meanwhile, the MAPE value in the MLR-2 is 22%, indicating reasonable or moderate forecasting accuracy. This demonstrates that the model is effective at predicting monthly discharge [3-5, 34, 35].

The MLR-2 is good at predicting monthly average discharge, this is because the Q4 MLR-2 uses variables Karian previous month discharge time step t-1(Qt- 1) and t-2(Qt-2), also using Variable rainfall the previous month time step t-1

(P1t-1). The result of the model correlation of historic discharge shows a strong correlation with the predicted discharge. Because the rainfall in the previous month t-1 becomes the discharge in month t, apart from that, there is a strong relationship between the discharge in Qt-2 and the predicted discharge Qt, but not as strong as the relationship between Q_{t-1} and Q_t. Furthermore, the position also determines the strength of the relationship. For rainfall P1, the position is further away than with P4 on the Karian reservoir, so that the value of MAPE with P1 is a bit different from the MAPE with P4 [3-5].

The MLR-3 model (Figure 7) has a correlation value of 75%. It can be seen that the relationship between the Q4 MLR-3 forecast discharge and historic data has a strong correlation. The MAPE value in the Q4 MLR-3 it shows reasonable or moderate forecasting accuracy. This indicates that the model is good at predicting monthly discharge [3-5, 34, 35].

The MLR-2 is good at predicting monthly average discharge, this is because the Q4 MLR-2 uses variables Karian's previous month discharge time step t-1(Qt- 1) and t-2(Qt-2), also using Variable rainfall the previous month time step t-1 (P4t-1). The result of the model correlation of historic discharge shows a strong correlation with the predicted discharge. Because the rainfall in the previous month t-1 becomes the discharge in month t, apart from that, there is a strong relationship between the discharge in Qt-2 and the predicted discharge Qt, but not as strong as the relationship between Q_{t-1} and Q_t. Furthermore, the position also determines the strength of the relationship. For rainfall P4, the position is closer to Karian Reservoir than P4 so the value of MAPE with P4 is a bit different than MAPE with P1 [3-5].

4. Model Comparison

To determine the selected forecast discharge, a statistical test comparison of the Markov Chain forecast model, SARIMA, and the quaternary MLR model is carried out in Table 5.

Table 5.
Comparison of Karian input discharge forecast models (Q4) Year (1997-2023).

		CHAIN-MARKOV	SARIMA	MLR		
				1	2	3
Accuracy	R2	0.63	0.66	0.76	0.75	0.75
Error	MAPE	0.26	0.25	0.19	0.22	0.21

Based on Table 5, it is found that the quaternary MLR-1 model Q4(Q4t-1P1t-1P4t-1) is the most representative model in forecasting Karian input discharge. This is represented by the correlation value with historic data, which is 76% and has the smallest model error value. It indicates that the MLR-1 model is a statistically nuanced model that is suitable for random and stochastic changes in hydrological components [3-5, 34, 35].

5. Conclusion

Based on the results, the Markov Chain, SARIMA, and MLR models all have an accuracy level of probability in increasing or decreasing discharge ranging from above 63% to 76%. This is also evidenced by the results of the correlation obtained using the three models, which produced a value in the range of 0.6 to 1, indicating a significant relationship between each model. Therefore, these models are suitable alternatives for forecasting discharge. The simulation results from the calibration of the water source model show that the largest correlation, ranging from 75% to 76%, was obtained using the MLR model, followed by the SARIMA and Markov Chain models with correlation values of 66% and 63%, respectively. The least average error rate (MAPE) between 19% and 22% was obtained using the MLR model, while the other models produced values of 25% and 26%. Therefore, the most representative water source model for discharge forecasts, and the closest to the historical series of discharges, is the MLR, followed by the SARIMA model and the Markov Chain model. In comparison to other models, the MLR-1 model produced the largest coefficient correlation of 0.76 or 76%. This shows that this model is capable of forecasting an increase or decrease in discharge of up to 76%. Therefore, the MLR-1 model is the best model that can serve as the basis for modeling the optimal management of water sources in the Ciberang-Ciujung watershed [3-5, 34, 35].

This study strengthens the argument that the comparison of the three models (Markov Chain, SARIMA, and MLR) with a statistical approach (calibration criteria decomposition) can help improve our understanding of model performance. This approach can assist in the design of a diagnostically robust evaluation strategy that supports the proper identification of hydrologically consistent models. With the comparison of the three models, it is hoped that the performance of the most representative model can be used as a basis for modeling analysis, especially to predict discharge data according to the conditions of the research area. The most representative alternative model in this study will be used as the basis for modeling discharge predictions and utilized to achieve optimal reservoir conceptualization.

References

[1] F. Li, G. Ma, C. Ju, S. Chen, and W. Huang, "Data-driven forecasting framework for daily reservoir inflow time series considering the flood peaks based on multi-head attention mechanism," *Journal of Hydrology*, vol. 645, p. 132197, 2024. https://doi.org/10.1016/j.jhydrol.2024.132197

[2] A. Gupta and A. Kumar, "Two-step daily reservoir inflow prediction using ARIMA-machine learning and ensemble models," *Journal of Hydro-environment Research*, vol. 45, pp. 39–52, 2022. https://doi.org/10.1016/j.jher.2022.05.001

[3] A. Bashir, M. A. Shehzad, I. Hussain, M. I. A. Rehmani, and S. H. Bhatti, "Reservoir inflow prediction by ensembling wavelet and bootstrap techniques to multiple linear regression model," *Water Resources Management*, vol. 33, no. 15, pp. 5121-5136, 2019.

[4] A. M. Mihel, N. Krvavica, and J. Lerga, "Regression-based machine learning approaches for estimating discharge from water levels in microtidal rivers," *Journal of Hydrology*, vol. 646, p. 132276, 2025.
[5] T. Jannah, Diyanti, and B. Santosa, "Flood discharge prediction by multiple linear regression method: Case study: Ciliwung watershed – cisadane," *Journal of Multidisciplinary Engineering Science Studies*, vol. 10, no. 4, pp. 569–578, 2024.
[6] M. C. Maniquiz, S. Lee, and L.-H. Kim, "Multiple linear regression models of urban runoff pollutant load and event mean concentration considering rainfall variables," *Journal of Environmental Sciences*, vol. 22, no. 6, pp. 946-952, 2010. https://doi.org/10.1016/S1001-0742(09)60203-5
[7] T. R. Lathrop, A. R. Bunch, and M. S. Downhour, "Regression models for estimating sediment and nutrient concentrations and loads at the Kankakee River, Shelby, Indiana, December 2015 through may 2018," US Geological Survey. https://doi.org/10.3133/sir20195005, 2328-0328, 2019.
[8] B. Ermias and V. Vishal, "Application of artificial intelligence for prediction of swelling potential of clay-rich soils," *Geotechnical and Geological Engineering*, vol. 38, pp. 6189-6205, 2020. https://doi.org/10.1007/s10706-020-01427-x
[9] M. M. Dorafshan, M. H. Golmohammadi, K. Asghari, and C. De Michele, "A novel fuzzified markov chain approach to model monthly river discharge," *Water Resources Management*, pp. 1-21, 2024.
[10] M. Jayanti, A. Sabar, H. D. Ariesyady, M. Marselina, and M. Qadafi, "A comparison of three water discharge forecasting models for monsoon climate region: A case study in cimanuk-jatigede watershed Indonesia," *Water Cycle*, vol. 4, pp. 17-25, 2023. https://doi.org/10.1016/j.watcyc.2023.01.002
[11] M. Jayanti, D. Marganingrum, H. Santoso, A. Sabar, H. D. Ariesyady, and M. Mariana, "The operation optimization of multipurpose reservoir between arima, continuous, and chain markov model on jatigede reservoir, Indonesia," *Continuous, and Chain Markov Model on Jatigede Reservoir, Indonesia*, vol. 5, pp. 30–45, 2024. https://doi.org/10.1016/j.watcyc.2024.01.003
[12] H. Bonakdari, A. H. Zaji, A. D. Binns, and B. Gharabaghi, "Integrated Markov chains and uncertainty analysis techniques to more accurately forecast floods using satellite signals," *Journal of Hydrology*, vol. 572, pp. 75-95, 2019. https://doi.org/10.1016/j.jhydrol.2019.02.027
[13] H. Moeeni, H. Bonakdari, and I. Ebtehaj, "Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach," *Journal of Earth System Science*, vol. 126, pp. 1-13, 2017. https://doi.org/10.1007/s12040-017-0798-y
[14] Z. Yang, D. Dong, Y. Chen, and R. Wang, "Water inflow forecasting based on visual modflow and gs-sarima-lstm methods," *Water*, vol. 16, no. 19, p. 2749, 2024. https://doi.org/10.3390/w16192749
[15] M. G. S. Kenyi and K. Yamamoto, "A hybrid SARIMA-Prophet model for predicting historical streamflow time-series of the Sobat River in South Sudan," *Discover Applied Sciences*, vol. 6, no. 9, p. 457, 2024. https://doi.org/10.1007/s42452-024-06083-x
[16] Korean International Cooperation Agency (KOICA), "Design report Korea rural community & agriculture corporation in association with korea water resources corporation republic of Korea Korea international cooperation agency," *Feasibility Study and Detailed Design of the Karian Dam Project*, 2016.
[17] S. Kamwaga, D. M. Mulungu, and P. Valimba, "Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 106, pp. 17-28, 2018.
[18] L. V. Duarte, K. T. M. Formiga, and V. A. F. Costa, "Comparison of methods for filling daily and monthly rainfall missing data: Statistical models or imputation of satellite retrievals?," *Water*, vol. 14, no. 19, p. 3144, 2022.
[19] A. O. Eruola, G. C. Ufoegbune, A. O. Eruola, J. A. Awomeso, and S. A. Abhulimen, "Determination of areal rainfall using estimation methods in a tropical wet and dry climate," *Journal of Hydrology*, vol. 37, pp. 079-082, 2015.
[20] J. B. Granato, "Development of regression equations to estimate flow durations and low-flow frequency statistics in New Hampshire streams," U.S. Geological Survey, Water-Resources Investigations Report 02-4298, 2003.
[21] J. H. Cho and J. H. Lee, "Multiple linear regression models for predicting nonpoint-source pollutant discharge from a highland agricultural region," *Water*, vol. 10, no. 9, p. 1156, 2018. https://doi.org/10.3390/w10091156
[22] A. Sabar, "Directorate of water and irrigation," *Global Trends in Sustainable Water Resources Infrastructure Development in the Context of Expert Discussion on the Formulation of Indonesia's Eco-Efficient Water Infrastructure Policy*, 2002.
[23] M. Marselina, A. Sabar, I. Rachmatiah Siti Salami., and M. D., "Water discharge forecast model in the context of optimizing the management of the saguling-kaskade citarum reservoir," *Theoretical Journal*, vol. 24, no. 1, 2017.
[24] L. A. Dar, "Rainfall-runoff modeling using multiple linear regression technique," *International Journal for Research in Applied Sciences, Engineering and Technology*, vol. 5, no. 7, pp. 214-218, 2017.
[25] S. Patel, M. Hardaha, M. K. Seetpal, and K. Madankar, "Multiple linear regression model for stream flow estimation of Wainganga River," *American Journal of Water Science and Engineering*, vol. 2, no. 1, pp. 1-5, 2016.
[26] A. Eshragh, B. Ganim, T. Perkins, and K. Bandara, "The importance of environmental factors in forecasting australian power demand," *Environmental Modeling & Assessment*, vol. 27, no. 1, pp. 1-11, 2022.
[27] U. H. Perez-Guerra *et al.*, "Seasonal autoregressive integrated moving average (SARIMA) time-series model for milk production forecasting in pasture-based dairy cows in the Andean highlands," *Plos one*, vol. 18, no. 11, p. e0288849, 2023.
[28] S. Lee and H. K. Kim, "Adsas: Comprehensive real-time anomaly detection system in international workshop on information security applications." Cham: Springer International Publishing, 2018, pp. 29-41.
[29] X. Chang, M. Gao, Y. Wang, and X. Hou, "Seasonal autoregressive integrated moving average (SARIMA) model for precipitation time series," *Journal of Mathematics and Statistics*, vol. 8, no. 4, pp. 500–505, 2012.
[30] H. A. Mombeni, S. Rezaei, S. Nadarajah, and M. Emami, "Estimation of water demand in Iran based on SARIMA models," *Environmental Modeling & Assessment*, vol. 18, pp. 559-565, 2013.
[31] P. Narayanan, A. Basistha, S. Sarkar, and S. Kamna, "Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India," *Comptes Rendus Geoscience*, vol. 345, no. 1, pp. 22-27, 2013. https://doi.org/10.1016/j.crte.2012.12.001
[32] H. Du, Z. Zhao, and H. Xue, "ARIMA-M: A new model for daily water consumption prediction based on the autoregressive integrated moving average model and the Markov chain error correction," *Water*, vol. 12, no. 3, p. 760, 2020. https://doi:10.3390/w12030760
[33] W. Wang, Y. Du, K. Chau, H. Chen, C. Liu, and Q. Ma, "A comparison of BPNN, GMDH, and ARIMA for monthly rainfall forecasting based on wavelet packet decomposition," *Water*, vol. 13, no. 20, p. 2871, 2021.

[34] S. Prion and K. A. Haerling, "Making sense of methods and measurement: Pearson product-moment correlation coefficient," *Clinical Simulation in Nursing*, vol. 10, no. 11, pp. 587-588, 2014.
[35] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38-48, 2016.
[36] D. Suryadarma, A. Suryahadi, and S. Sumarto, "Sectoral growth and job creation: Evidence from Indonesia," *Journal of International Development*, vol. 25, no. 4, pp. 549-561, 2013.
[37] M. Goyal and S. K. Bhagat, "A novel fuzzified markov chain approach to model monthly river discharge," *Water Resources Management*, vol. 38, no. 2, pp. 511–527, 2024.
[38] Z. Guo, "Research on the augmented dickey-fuller test for predicting stock prices and returns," in *Proceedings of the 2023 International Conference on Advanced Education, Management, and Social Science (AEMSS), pp. 123-130*, 2023.