

Accuracy Assessment of Monthly Rainfall Predictions using Seasonal ARIMA and Long Short-Term Memory (LSTM)

Ahmad Aldizar Akbar¹⁾, Yahya Darmawan^{1,2)}*, Arief Wibowo³⁾, Hayatul Khairul Rahmat¹⁾

¹⁾ Disaster Management Department, Faculty of Management and Business, University of Budi Luhur, Jakarta, Indonesia

²⁾ Climatology Department, Indonesia College of Meteorology Climatology and Geophysics (STMKG), Tangerang, Indonesia

³⁾ Information System Department, Faculty of Information Technology, University of Budi Luhur, Jakarta, Indonesia

¹⁾ 2234500060@student.budiluhur.ac.id ; ^{1,2)} yahya.darmawan@bmkgo.id* ; ³⁾ arief.wibowo@budiluhur.ac.id ;

¹⁾ hayatul.khairulrahmat@budiluhur.ac.id

¹⁾ Corresponding author

ARTICLE INFO

Article History:

Received January 1, 2020

Revised January 21, 2020

Accepted February 12, 2020

Keywords:

SARIMA

LSTM

Machine Learning

Rainfall Data

Accuracy

Correspondence:

Email: yahya.darmawan@bmkgo.id

ABSTRACT

Hydro meteorological disasters are common in Indonesia. Rainfall predictions can help mitigate the impact of these disasters. This research aims to compare the accuracy of monthly rainfall prediction models using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) methods. The input data consists of monthly rainfall records from four locations: Sampali, Kualanamu, Belawan, and Tuntungan, located around Medan, North Sumatra. The data spans from 2000 to 2020, with training data from 2000 to 2018 and test data from 2019 to 2020. SARIMA models effectively capture general trends and seasonality in linear time series data with clear patterns but struggle with extreme changes or sharp fluctuations due to their reliance on linear relationships. In contrast, LSTMs are effective at modeling complex, non-linear relationships, making them suitable for capturing general trends, seasonal patterns, and more complicated variations in the data. Understanding the characteristics of the data is crucial before applying SARIMA or LSTM models.

1. Introduction

According to data from the National Disaster Management Agency (BNPB) in 2023, Indonesia experienced a significant increase in hydro-meteorological disasters. Hydro-meteorological disasters include floods, landslides, and tornadoes, which often occur in various regions of Indonesia [1]. The main causes of these disasters are high rainfall, climate change, and environmental damage due to deforestation and unplanned urbanization. Considering Indonesia's geographical and tropical climate conditions, the likelihood of hydro-meteorological disasters is very high. Regions with hilly topography and numerous river systems, such as West Java, Sumatra, and Kalimantan, are particularly vulnerable to these disasters. Additionally, global climate change, which increases the frequency and intensity of extreme weather, further exacerbates the risk of hydro-meteorological disasters in Indonesia in the future. Therefore, better mitigation and adaptation efforts are needed to reduce the impact of these disasters [2].

Information about rainfall is crucial for mitigation and adaptation needs, especially concerning hydro-meteorological disasters such as floods and landslides [2]. Predicted rainfall data for short periods, such as 1 to 2 years ahead, is essential for effective planning and decision-making. These predictions can be

made using statistical or machine learning methods, each of which has its advantages in processing historical weather data and climate trends [3]. Statistical methods can provide a general overview based on past rainfall patterns, while machine learning can process more complex data and produce more accurate predictions by considering various related variables. With accurate rainfall prediction information, governments and communities can prepare mitigation measures such as flood control infrastructure construction, drainage system improvements, and better spatial planning management. Additionally, this information also aids in long-term adaptation by designing strategies to reduce the risk and impact of hydro-meteorological disasters in the future.

Weather forecasting is one solution that can be used to address the issue of data provision. A good forecast should have high accuracy and precision. There are several methods that can be used to predict future rainfall using statistical approaches such as SARIMA and the LSTM algorithm. The Autoregressive Integrated Moving Average (ARIMA) model completely disregards independent variables in making forecasts. ARIMA utilizes past and present values of the dependent variable to generate accurate short-term forecasts. However, ARIMA's long-term forecasting accuracy is less reliable. Forecasts tend to become flat (constant) over sufficiently long periods. SARIMA model is an extension of the ARIMA model with the addition of seasonal components [4]. Long Short-Term Memory (LSTM) is an algorithm for analyzing time series data, such as rainfall, and has proven to be reliable for predictions in recent years [5]. LSTM has emerged as a competitive model for handling seasonal data like rainfall. One of the significant advantages of this algorithm is its ability to remember long-term sequences, a challenge for traditional feature techniques. Moreover, LSTM can handle larger data sets by utilizing all the information within the data as its input [6].

Monthly rainfall predictions using LSTM and SARIMA methods have been widely applied to various types of data, including rainfall data. However, the prediction accuracy of monthly rainfall data for both methods can vary depending on the characteristics of the data [6]. Therefore, this study aims to analyze the accuracy of monthly rainfall predictions using the SARIMA and LSTM models in the Medan area and its surroundings. The better prediction model, based on accuracy and minimal error, is expected to serve as a reference for various applications that require high-quality monthly rainfall data. Accurate monthly rainfall information can support multiple sectors that depend on weather forecast information.

2. Method

2.1. Data used

The study uses monthly rainfall data from four rainfall stations around the Medan area in North Sumatra province, covering the period from 2000 to 2020. The rainfall stations included in the study are Sampali, Belawan, Kualanamu, and Tuntungan. The data was obtained from BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) Regional I Medan, North Sumatra, Indonesia. For the purpose of analysis, the data has been divided into training and test datasets based on specific time intervals. The data from 2000 to 2018 is used as the training dataset, while the data from 2019 to 2020 is used as the test dataset. This division allows for the development and validation of predictive models based on historical rainfall patterns and for the assessment of the model's performance on recent data.

2.2. Software used

The research was conducted using several tools for modeling SARIMA and LSTM. The tools used in this study consist of freely downloadable software:

- a. Anaconda Navigator is a popular and comprehensive Python distribution application. Anaconda provides many pre-installed data science libraries, such as Numpy, Pandas, Matplotlib, and others, making it an excellent choice for accessing various environments used in Python programming.
- b. Spyder is an Integrated Development Environment (IDE) used for developing Python applications. As an IDE, this application is specifically designed for numerical data analysis and provides various useful tools and features for data scientists and researchers.
- c. Microsoft Excel is spreadsheet software used for data analysis and visualization. It can be run on Microsoft Windows and Mac OS.

2.2.1. SARIMA

The SARIMA model comprises several components: Autoregressive (AR) term: The AR component indicates the dependency of the current value on previous values in the time series. The number of AR terms is denoted by p , for example, AR(p). Integrated (I) term: The I component indicates the number of differentiations needed to make the data stationary. Differentiation is performed to eliminate trends or seasonal patterns in the data. The number of I terms is denoted by d , for example, I(d). Moving Average (MA) term: The MA component indicates the dependency between the current value and the residual values (the difference between the current value and the predicted value) in previous periods. The number of MA terms is denoted by q , for example, MA(q).

The SARIMA model adds a seasonal component to each ARIMA component. The ARIMA model is constructed with orders (p, d, q), which is a mixed model of autoregressive (AR) with order p followed by moving average (MA) with order q that undergoes differencing (d) times. Data series plots should assume stationarity. If the data to be processed is not stationary, differencing needs to be performed, and initial values of the d order can be estimated. However, differencing is unnecessary if the data to be processed is already stationary [7].

The ARIMA model equation (p,d,q) can be written as follows:

$$Y_t = \xi + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \dots + \phi_p Y_{t-p} + \varepsilon_t$$

Where :

Y_t = stationary series value

ξ = model mean value

Q_j = self-regression model parameters ($j = 1, 2, \dots, p$)

ε_t = normally distributed random error with mean zero and variance σ^2

2.2.2. Long Short Term Memory (LSTM)

LSTM is an architecture of RNN (Recurrent Neural Network). LSTM can be used to process sequential data, thus it can be employed for predicting time series data. LSTM can detect data to be stored and data not to be used for pruning, as LSTM consists of 4 neuron layers commonly referred to as gates to regulate memory in each neuron. Research is conducted by predicting weather data such as rainfall and temperature using the Long Short Term Memory (LSTM) method. The predicted model results can provide suggestions for implementing a good weather prediction model through predictions using the LSTM method. [8].

The mathematical calculations in the LSTM model carried out at each stage are given as follows:

$$\begin{aligned}
 \text{forget gate :} & \quad \mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{h}_{t-1} + \mathbf{b}_f) & = \sigma(\bar{\mathbf{f}}_t) \\
 \text{input gate :} & \quad \mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{h}_{t-1} + \mathbf{b}_i) & = \sigma(\bar{\mathbf{i}}_t) \\
 \text{candidate state :} & \quad \mathbf{z}_t = \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{R}_c \mathbf{h}_{t-1} + \mathbf{b}_c) & = \tanh(\bar{\mathbf{z}}_t) \\
 \text{cell state :} & \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{z}_t \\
 \text{cell gate :} & \quad \mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{h}_{t-1} + \mathbf{b}_o) & = \sigma(\bar{\mathbf{o}}_t) \\
 \text{output :} & \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
 \end{aligned}$$

\mathbf{x}_t is the input vector at time t , \mathbf{W} and \mathbf{R} are weight matrices. \mathbf{b} is the bias vector. σ and \tanh are activation/transfer functions, and \odot denotes element wise multiplication. [9] provides the detailed algorithm.

2.3. Accuracy Assessment

2.3.1. Root Mean Square Error

Root Mean Square Error (RMSE) is one method used to evaluate forecasting predictions, employed to measure the accuracy level of a model's forecasts. RMSE is the square root of the average of the squared errors and also represents the magnitude of errors produced by a forecasting model [10]. RMSE can be expressed with the formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}$$

Where:

RMSE = Root Mean Square Error

y = bulk Rain observation

x = bulk rain model output

n = Number of data

2.3.2. Mean Absolute Error

Mean Absolute Error (MAE) is one method used to measure the accuracy level of a forecasting model. The MAE value indicates the average absolute error between the forecasted/predicted values and the actual values [11]. It is a commonly used measurement for predicting errors in time series analysis, where the term Mean Absolute Deviation (MAD) is sometimes used interchangeably, referring to the Mean Absolute Error [12]. MAE can be explained as follows:

$$MAE = \frac{1}{n} \sum_{i=0}^n |f_i - y_i|$$

Where:

MSE = Mean Absolute Error

f_i = value of forecasting results

y_i = actual value

n = amount of data

2.3.3. Temporal Pattern Analysis

The objectives of the Temporal Pattern Analysis line graph are multifaceted, aiming to provide a comprehensive understanding of rainfall trends and model performance. Firstly, the graph visualizes temporal trends, helping to identify patterns, seasonality, and anomalies in the rainfall data over time. This allows for a clear illustration of how rainfall fluctuates in each location. Additionally, the graph serves to compare model predictions by plotting the actual rainfall data against the predicted values from both LSTM and SARIMA models. This comparison highlights the accuracy and efficacy of each model in capturing the temporal dynamics of rainfall. Furthermore, the graph aids in evaluating model performance by examining how closely the predicted values align with the actual rainfall over different periods, thus identifying periods where the models underperform or overperform. Anomaly detection is another critical objective, as the graph helps in spotting significant deviations between predicted and actual values, which may indicate model limitations or unique climatic events. Lastly, the graph supports decision-making by offering visual evidence to stakeholders, enabling them to determine which model is more suitable for specific locations based on temporal prediction performance. Overall, the Temporal Pattern Analysis line graph is a valuable tool for understanding, comparing, and enhancing rainfall prediction models.

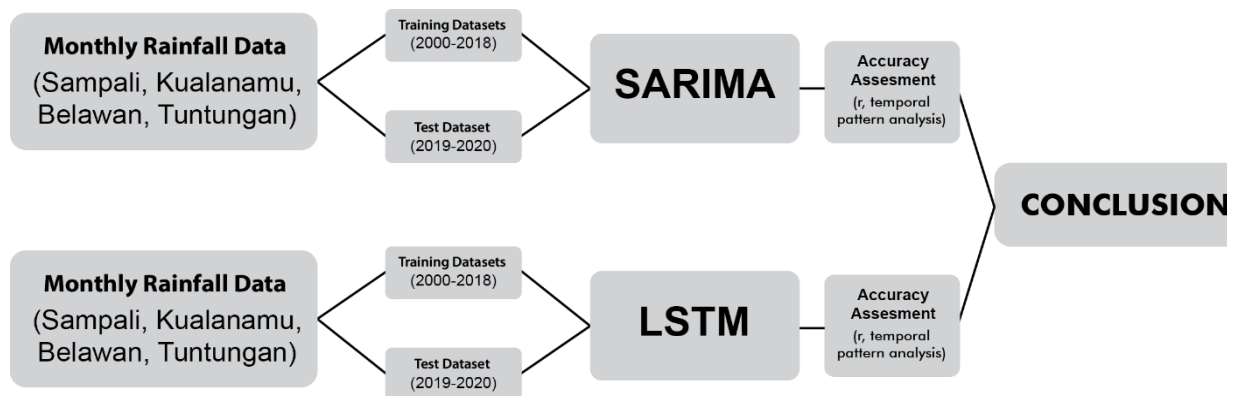


Figure 1. Schematic flow of research.

3. Results and Discussion

3.1 SARIMA

SARIMA determines the model order for each Region automatically. The programming in Python employs the "auto_arma" system for model order selection. Non-seasonal model orders are expressed in lowercase (p,d,q), while seasonal model orders are expressed in uppercase (P,D,Q). The seasonal component is denoted by 's', where 12 represents the number of months in one year. The use of the "auto_arma" function automates the complex task of model selection by testing various combinations of p, d, q, P, D, Q, and s parameters to find the best fit for the given data. This is particularly useful for time series data with strong seasonal patterns, such as monthly rainfall data, where the model needs to account for both short-term fluctuations and annual seasonal effects. Table 1 would provide an example

of how these parameters are applied to the rainfall data for each region, demonstrating the specific model orders selected by "auto_arima" for both non-seasonal and seasonal components.

Table 1. Non-seasonal model orders and seasonal model orders of SARIMA for all locations

SARIMA model (p,d,q) (P,D,Q,s)	
Sampali	(1, 0, 0)(2, 0, 0, 12)
Kualanamu	(2, 1, 0)(1, 0, 1, 12)
Belawan	(2, 0, 0)(1, 0, 2, 12)
Tuntungan	(1, 0, 0)(2, 0, 0, 12)

3.2 LSTM

The number and size (number of neurons) of hidden layers in the LSTM algorithm significantly influence the model's ability to capture patterns in sequential data. In this study, the hidden layer structure in the LSTM model is a crucial part of the artificial neural network (ANN) model creation process. The model comprises two hidden layers, each using the Rectified Linear Activation (ReLU) function. ReLU is a popular activation function that deactivates neurons with negative input values and leaves positive input values unchanged, helping to avoid issues like the vanishing gradient problem. The first hidden layer consists of 64 units and uses ReLU as the activation function, with an input shape specified by the look_back parameter. It is implemented in the script as `model.add(Dense(64, activation='relu', input_shape=(look_back,)))`. The second hidden layer consists of 32 units and also uses the ReLU activation function, implemented as `model.add(Dense(32, activation='relu'))`. These hidden layers process the input data with appropriate dimensions to capture underlying patterns without overfitting or underfitting. The Sequential approach is used to build the model, where layers are added one after the other, making the neural network architecture straightforward to construct and understand. By specifying the hidden layers and their respective parameters, the LSTM model is better equipped to process sequential data and uncover intricate patterns in the rainfall data, ultimately leading to more accurate predictions.

3.3 Accuracy Assessment

3.3.1. RMSE

The RMSE comparison graph is plotted on an actual scale (mm). The largest RMSE values for both models are found in Belawan, with 27.68 mm for LSTM and 28.36 mm for SARIMA. In contrast, the smallest RMSE values for LSTM are in Sampali, with 14.45 mm, while the smallest RMSE values for SARIMA are in Tuntungan, with 2.61 mm. This indicates that while LSTM performs better in some locations, SARIMA performs significantly better in others, particularly in Tuntungan (Figure 2).

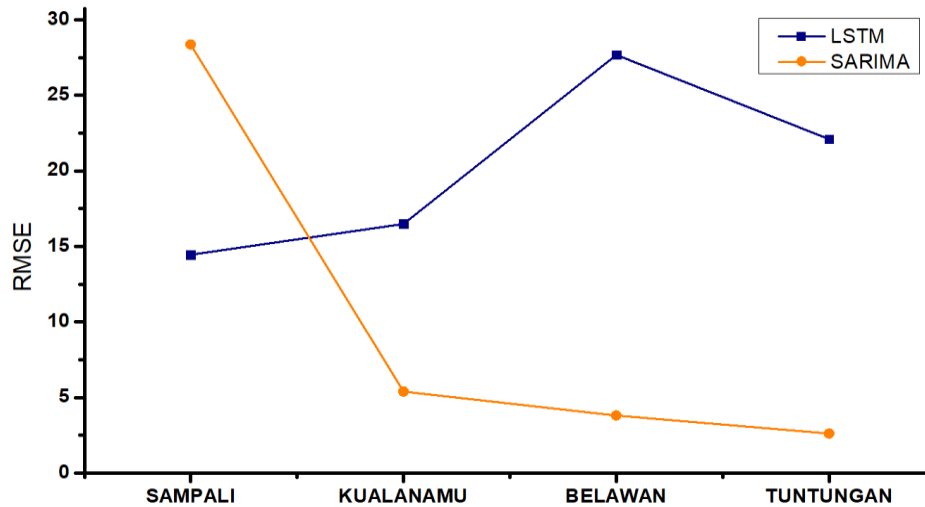


Figure 2. The RMSE comparison graph shows the RMSE values for LSTM and SARIMA models across four locations: Sampali, Kualanamu, Belawan, and Tuntungan.

The analysis of the provided RMSE data unveils insights into the performance of LSTM and SARIMA models across different regions. Notably, the comparison highlights distinct strengths and weaknesses of each model. In Belawan, both models exhibit higher RMSE values, indicating challenges in accurately predicting environmental variables in this area. However, in Sampali, the LSTM model showcases better predictive accuracy with a smaller RMSE, suggesting its proficiency in capturing temporal dependencies. Conversely, SARIMA outperforms LSTM in Tuntungan, indicating its effectiveness in capturing seasonal and autoregressive components of the data. These findings underscore the importance of understanding regional variations and model-specific characteristics for improving predictive accuracy. Further investigation and refinement of the models, considering the unique dynamics of each region, could lead to enhanced forecasting capabilities in hydrological or environmental applications.

3.3.2. MAE

The MAE comparison graph is presented on an actual scale (mm), showcasing the performance of both LSTM and SARIMA models across the four locations. Notably, the highest MAE values are recorded in Belawan, with LSTM and SARIMA yielding 5.65 mm and 5.79 mm respectively. Conversely, the smallest MAE values are observed in Sampali, with LSTM achieving 2.95 mm and SARIMA outperforming with a mere 0.53 mm. Additionally, in Tuntungan, SARIMA demonstrates its superior performance with a notably low MAE compared to LSTM. This analysis highlights the varying efficacy of both models across different locations, emphasizing the importance of considering regional factors in rainfall prediction (Figure 3). Overall, SARIMA shows better performance than LSTM in three out of the four locations (Kualanamu, Belawan, and Tuntungan), demonstrating significantly lower forecasting errors. LSTM only performs better in Sampali. This analysis suggests that SARIMA generally provides more accurate forecasts than LSTM for the locations considered, with Tuntungan showing the most significant improvement when using SARIMA over LSTM.

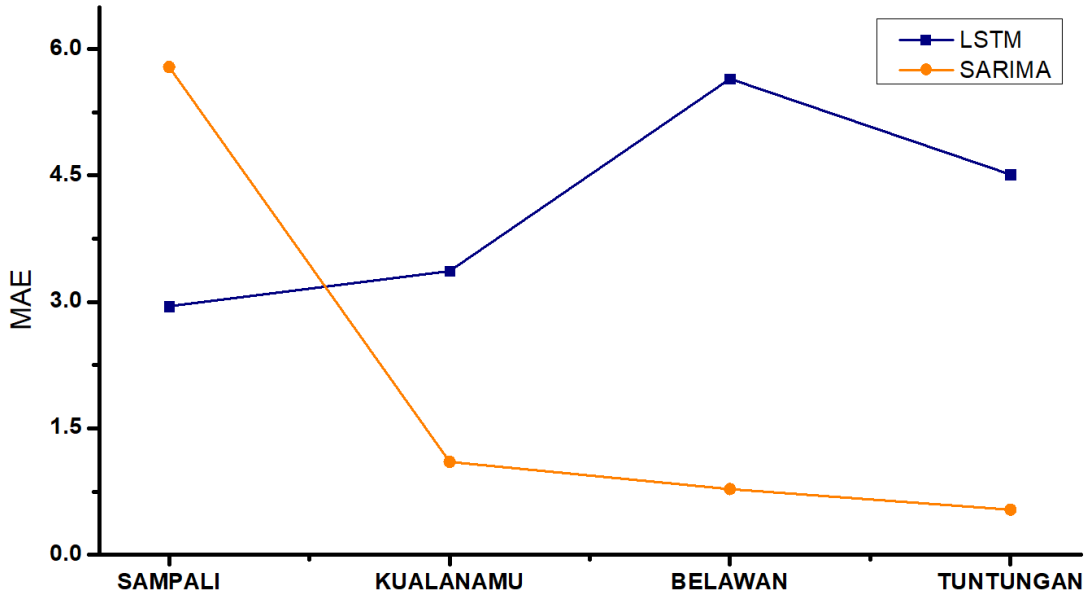


Figure 3. The MAE comparison graph shows the MAE values for LSTM and SARIMA models across four locations: Sampali, Kualanamu, Belawan, and Tuntungan.

3.3.3. Correlation Coefficient (r)

The correlation comparison graph is plotted on an actual scale (mm). The highest correlation values for both models are in Sampali, with values of 0.65 mm for LSTM and 0.56 mm for SARIMA in Belawan, and the smallest values are in Tuntungan, with 0.49 mm for LSTM and 0.47 mm for SARIMA in Kualanamu (Figure 4).

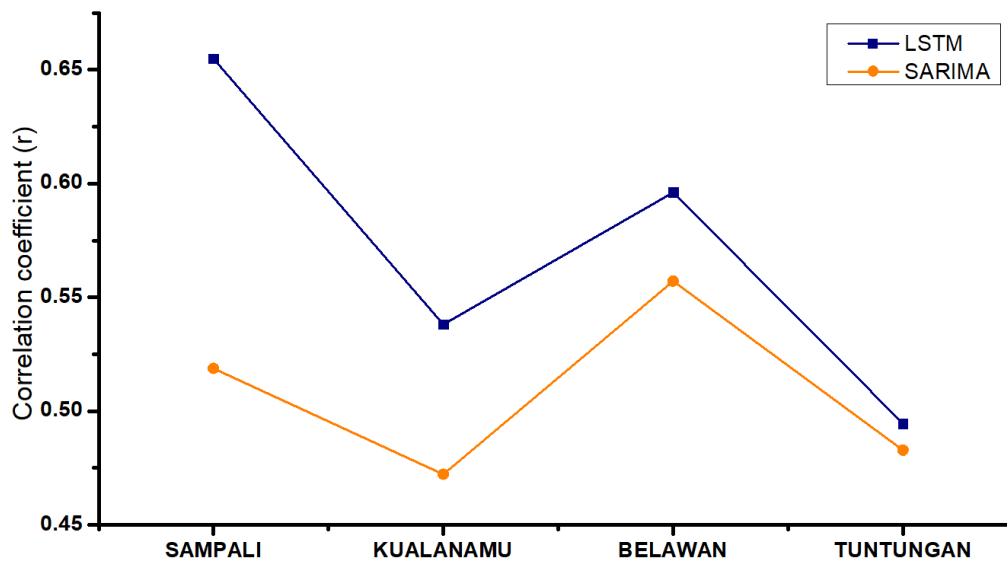


Figure 4. The Correlation Coefficient (r) comparison graph shows the MAE values for LSTM and SARIMA models across four locations: Sampali, Kualanamu, Belawan, and Tuntungan.

The correlation comparison graph, plotted on an actual scale (mm), provides insights into the relationship between predicted and actual rainfall values for both LSTM and SARIMA models across four locations.

In Sampali, both models achieve their highest correlation values, with LSTM at 0.65 mm and SARIMA at 0.56 mm, indicating a strong relationship between predicted and observed values in this location. Conversely, in Belawan, the correlation values are notably lower, with SARIMA showing a correlation of 0.56 mm. The smallest correlation values are observed in Tuntungan for LSTM (0.49 mm) and in Kualanamu for SARIMA (0.47 mm), suggesting weaker predictive accuracy in these areas. Overall, the analysis reveals that while both models perform well in Sampali, their effectiveness varies in other locations, with SARIMA generally providing more stable and reliable predictions across the board except in areas where LSTM shows higher correlations. This highlights the importance of location-specific model selection for optimal rainfall prediction accuracy.

3.4 Temporal Pattern Analysis

Temporal Pattern Analysis for every location can be found below:

3.4.1. Sampali

In Sampali, there are differences in the values generated by the SARIMA model, which are quite different from the observed values, with only a few closely approaching the observed values. The graph in the LSTM (Long Short-Term Memory) method with monthly data shows two curves. The gray line (Observed) indicates the actual rainfall values with significant variations at some points. There are sharp peaks and drastic drops in this data. The blue line (LSTM) shows the prediction results from the LSTM model. These predictions tend to be smoother and follow the general trend of the observed data, although they do not always accurately capture all peaks and valleys. These results are obtained using a batch size of 64 with a correlation value of 0.65 for the LSTM model (Figure 5).

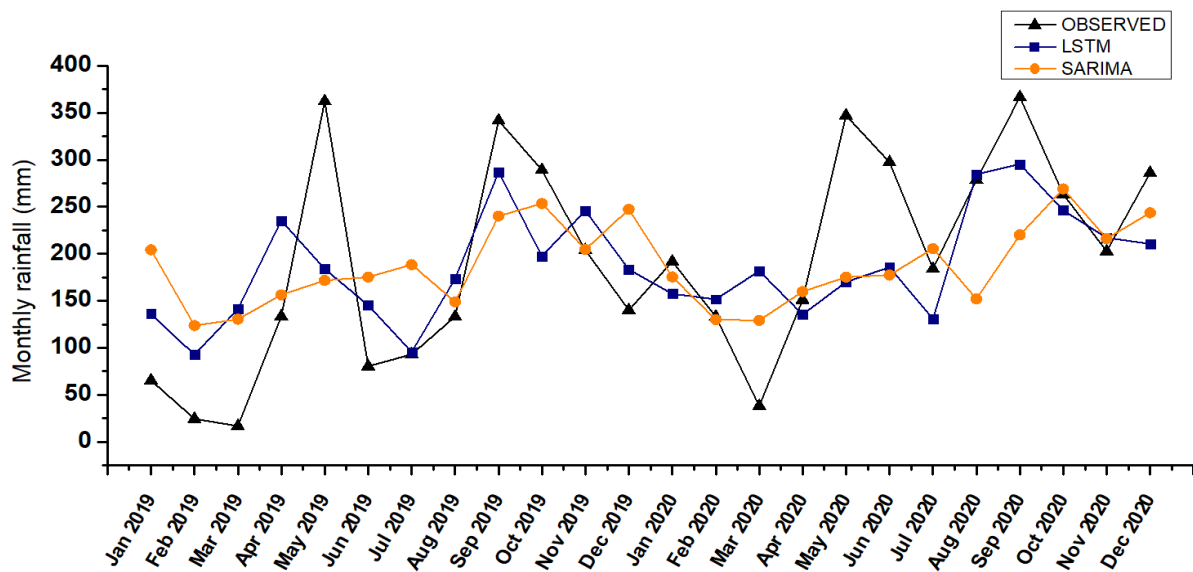


Figure 5. Temporal patterns of predicted monthly rainfall using SARIMA and LSTM models compared to observed data in Sampali from January 2019 to December 2020.

The provided graph compares the monthly rainfall from January 2019 to December 2020 as observed (blue line) and predicted by two models: LSTM (Long Short-Term Memory, blue line) and SARIMA (Seasonal Autoregressive Integrated Moving Average, orange line). The observed rainfall data shows significant variability, with notable peaks around May 2019, August 2019, and September 2020. The

LSTM model generally follows the observed trends but shows more fluctuations and is sometimes closer to the observed values. In contrast, the SARIMA model provides a smoother prediction, capturing the overall seasonal trends but missing some of the sharper peaks and troughs present in the observed data. Both models show a similar trend to the observed data but differ in their response to short-term fluctuations.

3.4.2. Kulanamu

The SARIMA model used in Kulanamu seems to provide a rough representation of the actual data, with weaknesses in capturing extreme changes. This may require parameter adjustments or other model approaches that can be more sensitive to sharp fluctuations in the data. Overall, the predictions generated by the LSTM model in Kulanamu follow the general trend of the observed data. The LSTM model is quite adept at capturing seasonal patterns and cycles in the data. Some major seasonal peaks and valleys are well followed, although the prediction amplitudes are often lower or higher than the actual data. These results are obtained using a batch size of 64 with a correlation value of 0.54 for the LSTM model (Figure 6).

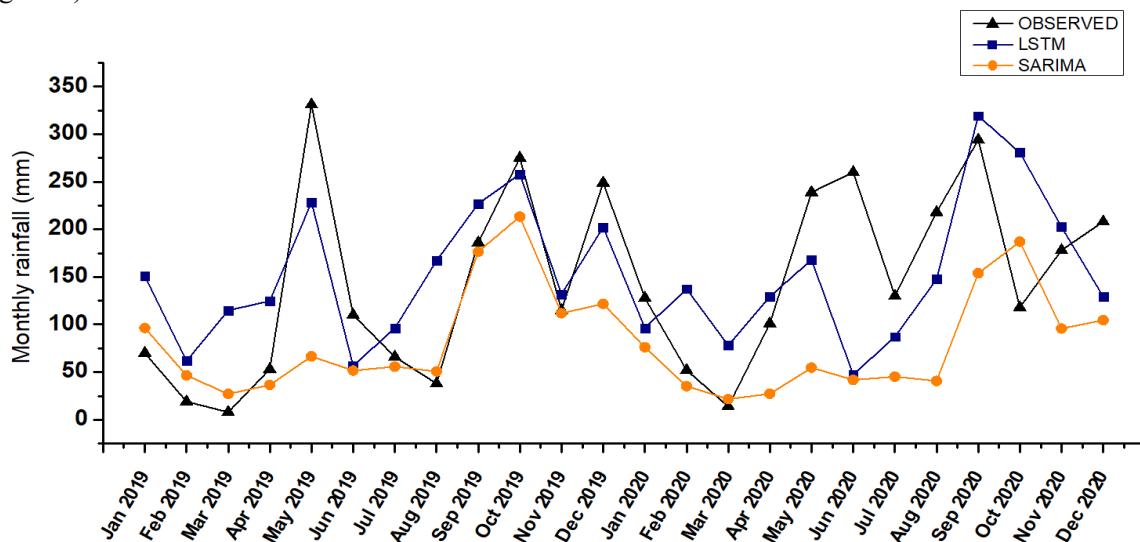


Figure 6. Temporal patterns of predicted monthly rainfall using SARIMA and LSTM models compared to observed data in Kulanamu from January 2019 to December 2020.

The graph presents a comparison of observed monthly rainfall (blue line) against predictions from LSTM (blue line) and SARIMA (orange line) models from January 2019 to December 2020. The observed rainfall data exhibits high variability with significant peaks around May 2019, September 2019, and September 2020. The LSTM model captures the general pattern of the observed rainfall, closely following the sharp increases and decreases, but sometimes exaggerates the magnitude of these fluctuations. Conversely, the SARIMA model provides a smoother prediction, successfully identifying the overall trend but often underestimating the magnitude of the rainfall peaks and failing to capture some short-term variations. Overall, while both models reflect the seasonal patterns of rainfall, the LSTM model is more responsive to sudden changes, whereas the SARIMA model offers a steadier, less variable forecast.

3.4.3. Belawan

In Belawan, the graph illustrates how the SARIMA model attempts to predict future values from time series data based on its historical patterns. Although this model's predictions are quite good at following the general trend, there are limitations in capturing extreme fluctuations, which are a common challenge in time series modeling. The orange line indicates predictions that tend to be smoother and follow the general trend of the observed data, although they do not accurately capture all peaks and valleys. On the other hand, the LSTM model demonstrates good capability in capturing the general trends and seasonal patterns of monthly rainfall data. Although it does not always accurately capture all peaks and valleys. These results are obtained using a batch size of 32 with a correlation value of 0.60 for the LSTM model (Figure 7).

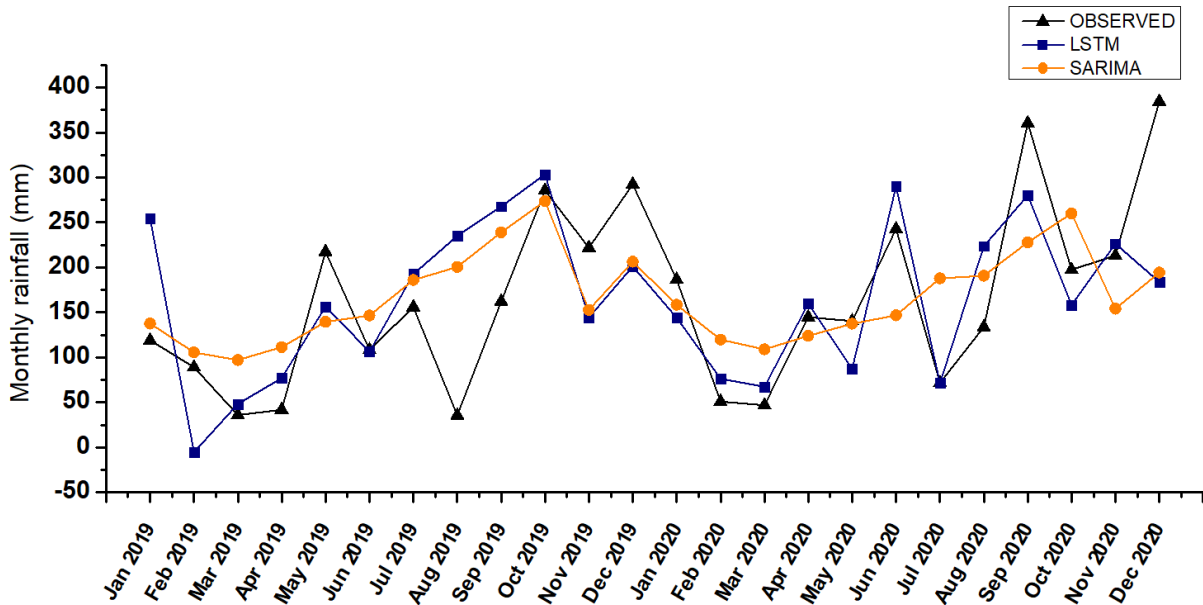


Figure 7. Temporal patterns of predicted monthly rainfall using SARIMA and LSTM models compared to observed data in Belawan from January 2019 to December 2020.

The graph depicts a comparative analysis of observed monthly rainfall (black line) and the predictions generated by the LSTM (Long Short-Term Memory, blue line) and SARIMA (Seasonal Autoregressive Integrated Moving Average, orange line) models from January 2019 to December 2020. The observed data exhibits considerable fluctuations, with notable peaks in May 2019, September 2019, and December 2020. The LSTM model closely follows the observed data's trends and captures the peaks and troughs more accurately, though it sometimes overestimates the magnitude of rainfall. This indicates the LSTM model's strength in handling complex patterns and short-term variations. In contrast, the SARIMA model provides a smoother and more consistent prediction, effectively capturing the overall seasonal trend but often underestimating the intensity of rainfall peaks and missing some short-term variability. The SARIMA model's predictions are less volatile, offering a more stable but less detailed forecast. Therefore, while the LSTM model demonstrates better performance in tracking the rapid changes in rainfall, the SARIMA model excels in offering a generalized view of seasonal trends.

3.4.4. Tuntungan

The graph in Tuntungan appears to be less responsive to extreme fluctuations. For example, the high peaks around points 237-239 and sharp declines around point 241 are not fully reflected in the prediction

line. The blue line (Observed) shows observed data with significant fluctuations at some points. There are several sharp peaks and drastic declines in this data. The orange line (Predicted) indicates the prediction results from the SARIMA model. These predictions tend to be smoother and do not always capture sharp peaks and valleys in the observed data. In contrast to the prediction data in the previous city, Tuntungan shows quite different results from the observational data, although there are still some points that approximate. In this city, there are still difficulties in predicting and capturing extreme fluctuations in the data. These results are obtained using a batch size of 64 with a correlation value of 0.49 for the LSTM model (Figure 8).

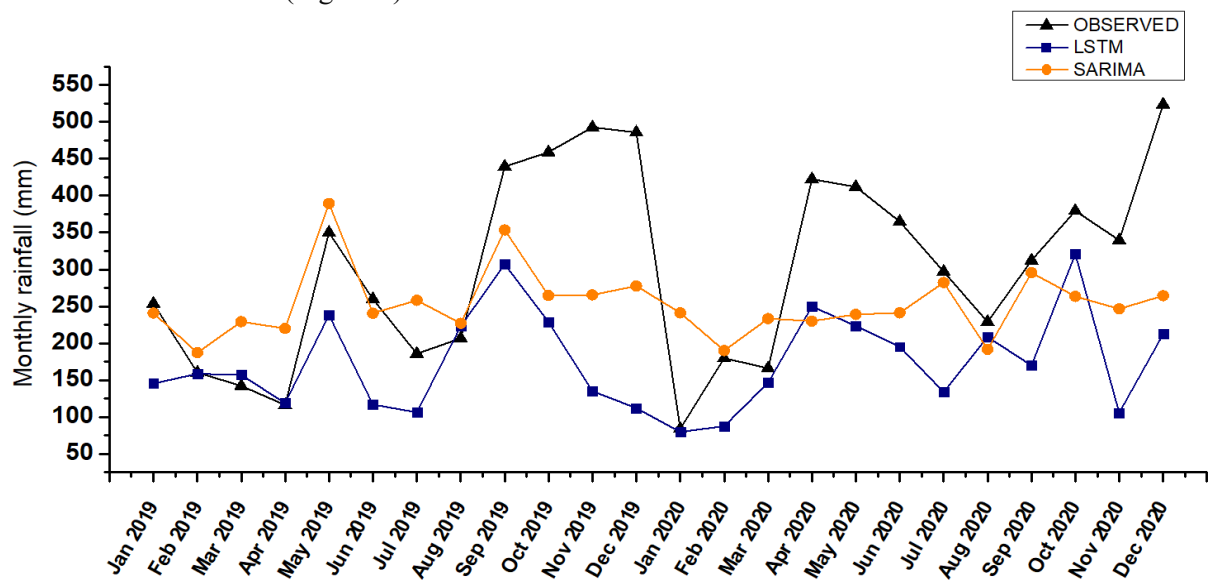


Figure 8. Temporal patterns of predicted monthly rainfall using SARIMA and LSTM models compared to observed data in Tuntungan from January 2019 to December 2020.

The graph displays monthly rainfall data from January 2019 to December 2020, comparing observed rainfall (blue line) with predictions from two models: LSTM (blue line) and SARIMA (orange line). The observed rainfall exhibits significant variability, with peaks around mid-2019, early 2020, and mid to late 2020. Both models attempt to capture this variability but differ in their performance. The LSTM model shows greater fluctuations and often fails to align closely with the observed data, particularly in capturing the higher peaks and some of the lower troughs. This suggests that while LSTM can model complex patterns, it struggles with the high variability and extreme values present in the observed rainfall data. The SARIMA model, on the other hand, demonstrates a more smoothed prediction, capturing some of the general trends but missing the sharper peaks and valleys. SARIMA's performance indicates a better fit for the general trend but less sensitivity to abrupt changes in rainfall. Overall, neither model perfectly replicates the observed data, but the SARIMA model appears to provide a more stable prediction, albeit at the cost of missing some of the finer details captured by the LSTM model. This analysis highlights the challenge of accurately modeling highly variable time series data like monthly rainfall.

4. Conclusion

The SARIMA and LSTM models both demonstrate distinct capabilities and limitations in predicting monthly rainfall data. SARIMA excels in capturing general trends and seasonal patterns, providing a reliable overview of data movement, but it struggles with accurately predicting extreme changes and sharp fluctuations, leading to misalignments during high peaks and low valleys. In contrast, the LSTM model shows a stronger ability to follow the general direction and seasonal patterns of the data, sometimes closely matching observed values. However, it also faces challenges in accurately predicting very sharp peaks and valleys. Enhancing LSTM's prediction accuracy could involve optimizing hyper parameters, adding more training data, or integrating additional methods to handle high variability in the data. Overall, while SARIMA offers stability in trend prediction, LSTM's flexibility makes it more adept at capturing complex patterns, though both require further refinement for precise rainfall forecasting.

References

- [1] ES Adiningsih, "Overview of Drought Parameter Detection Methods Based on Remote Sensing Data," *Pros. Semin. Nas. Remote Sensing 2014* , pp. 210–220, 2014.
- [2] W. Rahmalina and Novreta, "Forecasting the Floating Drought Index Using the Sarima and SPI Methods," *Potential J. Civil Polytek.* , vol. 22, no. 1, pp. 64–75, 2020, doi: 10.35313/potensi.v22i1.1824.
- [3] S. Anwar, "Short Term Air Temperature Forecasting in Banda Aceh City Using the Autoregressive Integrated Moving Average (ARIMA) Method," *Malikussaleh J. Mech. Sci. Technol.* , vol. 5, no. 1, p. 6, 2017, doi: 10.29103/mjms.v5i1.10882.
- [4] LJ Sinay, FK Lembang, SN Aulele, and D. Mustamu, "Analysis of Monthly Rainfall in Ambon City Using the Heteroscedasticity Model: Sarima-Garch," *Media Stat.* , vol. 13, no. 1, pp. 68–79, 2020, doi: 10.14710/medstat.13.1.68-79.
- [5] SJ and SA, "A Comparative Analysis Of Web Information Extraction Techniques Deep Learning vs. Naïve Bayes vs. Back Propagation Neural Networks In Web Document Extraction," *ICTACT J. Soft Comput.* , vol. 06, no. 02, pp. 1123–1129, 2016, doi: 10.21917/ijsc.2016.0156.
- [6] D. Desmonda, T. Tursina, and MA Irwansyah, "Prediction of Rainfall Amount Using the Fuzzy Time Series Method," *J. Sist. and Technol. Inf.* , vol. 6, no. 4, p. 141, 2018, doi: 10.26418/justin.v6i4.27036.
- [7] AM Soetamto and Ulfah, *ARIMA Method, Training Module for Improving Seasonal Forecast Accuracy* . Jakarta: Meteorology and Geophysics Agency, 2006.
- [8] T. Lattifia, P. Wira Buana, and NKD Rusjyanthi, "Weather Prediction Model Using the LSTM Method," *JITTER J. Ilm. Technol. and Comput.* , vol. 3, no. 1, pp. 994–1000, 2022.
- [9] S. Hochreiter and J. Schmidhuber, "Long Short Term Memory. Neural Computation," *Neural Comput.* , vol. 9, no. 8, 1997.
- [10] JS Armstrong, "Evaluating Forecasting Methods," pp. 443–472, 2001, doi: 10.1007/978-0-306-47630-3_20.
- [11] AA Suryanto, "Application of the Mean Absolute Error (Mea) Method in the Linear Regression Algorithm for Predicting Rice Production," *Saintekbu* , vol. 11, no. 1, pp. 78–83, 2019, doi: 10.32764/saintekbu.v11i1.298.
- [12] AY Permatasari, "CHAPTER II Literature Review CHAPTER II LITERATURE REVIEW 2.1. 1–64," *Gastron. ecuatoriana y Tour. local.* , vol. 1, no. 69, pp. 5–24, 2023.

● **13% Overall Similarity**

Top sources found in the following databases:

- 11% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	icsejournal.com Internet	7%
2	mdpi.com Internet	2%
3	Ahmad Syarif. "Forecasting the Development of Islamic Bank in Indone... Crossref	<1%
4	University of Greenwich on 2023-12-20 Submitted works	<1%
5	Rizki Mahrivandi, Lienda Noviyanti, Gatot Riwi Setyanto. "Black-Litterm... Crossref	<1%
6	gssrr.org Internet	<1%
7	mdpi-res.com Internet	<1%
8	link.springer.com Internet	<1%

9	www2.mdpi.com Internet	<1%
10	mSPACE.lib.umanitoba.ca Internet	<1%
11	University of Wales Institute, Cardiff on 2024-03-17 Submitted works	<1%
12	etd.aau.edu.et Internet	<1%
13	Governors School for Science and Technology on 2023-04-17 Submitted works	<1%
14	Universitas Riau on 2021-09-10 Submitted works	<1%
15	appstate.edu Internet	<1%
16	naun.org Internet	<1%