IPC

Indonesian Physics Communication

Department of Physics, Universitas Riau, Pekanbaru, Indonesia
p-ISSN 1412-2960 | e-ISSN 2579-521X

Vol. 22 | No. 1
March 2025

Web: https://kfi.ejournal.unri.ac.id
E-mail: kfi@ejournal.unri.ac.id

Evaluation of the Arima-Kalman model in predicting rainfall in Medan City in 2023 using observation data from 2013 – 2022

Alva Josia Lumbantoruan, Yahya Darmawan*, Munawar Munawar, Nardi Nardi, Fendy Arifianto, Ervan Ferdiansyah
Indonesia College of Meteorology, Climatology and Geophysics (STMKG), Tangerang 15221, Indonesia

*Corresponding author: yahya.darmawan@bmkg.go.id

ABSTRACT

This paper aims to evaluate the ARIMA-Kalman model in predicting rainfall in Medan City for the year 2023. The data used are historical observation data of rainfall from 2013 to 2022 that have been tested for stationary and homogeneity, which proved not to require additional correction. The analysis results show that the ARIMA-Kalman model can capture the general pattern of rainfall well, and shows superiority in producing predictions that are closer to the actual data, with a mean absolute error (MAE) value of 54.11, which is lower than the MAE of the ARIMA model which reaches 55.66. Although the ARIMA model has a smaller root mean square error (RMSE) (66.67 compared to 69.75 for ARIMA-Kalman), the ARIMA-Kalman model shows better consistency, especially in capturing significant fluctuations, such as the peak rainfall that occurred in July 2023. Therefore, ARIMA-Kalman is proven to be more accurate and reliable for predicting rainfall in Medan city, making it a better choice to support water resources planning and management.

Keywords: ARIMA; Kalman filter; model; prediction; rainfall

Received 27-01-2025 | Revised 08-03-2025 | Accepted 24-03-2025 | Published 31-03-2025

INTRODUCTION

Rainfall is one of the variables determining climatic conditions which is directly related to the success of various fields, including agriculture and plantations. In Indonesia, rainfall is the main factor influencing the climate in tropical areas. However, the level of rainfall is often difficult to predict, so many people still lack strategies for dealing with fluctuations in rainfall [1].

Medan City, as one of the big cities in Indonesia, is not immune to the impact of changes in rainfall [2]. Most of the activities of the Medan City population depend on rainfall, from farmers, fishermen, transportation users, to food production. Therefore, accurate rainfall predictions are needed to help various sectors prepare for weather changes [1].

In an effort to improve the accuracy of rainfall predictions, various forecasting methods have been developed. One of the commonly used methods is the autoregressive integrated moving average (ARIMA) model [3]. The ARIMA model is effective for time series data and can produce accurate short-term forecasts [4]. However, to improve prediction accuracy, several studies have combined the ARIMA model with other methods, such as the Kalman Filter.

The Kalman filter is a model of part of the state space that can be applied in forecasting models. This model uses a recursive technique to integrate the latest observation data into the model, correct previous predictions, and make further predictions optimally based on past information and current data [5]. The combination of ARIMA and Kalman filter has shown promising results in improving the accuracy of rainfall predictions [6].

Evaluation of the ARIMA-Kalman model in predicting rainfall in Medan City in 2023 is important to do. This aims to determine the extent to which the model can provide accurate and reliable predictions. The results of this evaluation are expected to help parties affected by rainfall in establishing strategies to deal with the ups and downs of rainfall in the future.

LITERATURE REVIEW

Time Series Analysis

Time series analysis is a statistical method used to analyze data collected over a period of time. It is essential in various fields, including meteorology, economics, and engineering, as it allows researchers to understand patterns and trends in time-related data. In the context of rainfall, time series analysis helps in identifying seasonal patterns, long-term trends, and fluctuations that can affect water resource planning and management [4].

A time series consists of a series of observations taken at regular time intervals. One of the main objectives of time series analysis is to forecast future values based on historical data. To achieve this goal, it is important to identify whether the data is stationary or not [7]. Data is said to be stationary if its mean and variance are constant over time [8]. If the data is not stationary, then a transformation, such as differencing, needs to be performed to make it stationary. Stationarity tests can be performed using the autocorrelation function (ACF) and partial autocorrelation function (PACF), which help in determining the relationship between the current observation and the previous observations.

ARIMA Model

The ARIMA model is one of the methods most commonly used in time series analysis. This model was introduced by Box and Jenkins and is a generalization of the autoregressive moving average (ARMA) model with the addition of an integration component to deal with non-stationary data. ARIMA combines three main components: autoregressive (AR), integrated (I), and moving average (MA) [9].

Autoregressive

AR is a component that shows that current values are influenced by previous values. This reflects the relationship between current observations and past observations. In the context of rainfall, this means that rainfall in a particular month can be influenced by rainfall in previous months.

Integrated

I is a process that involves differentiating the data to remove trends and make the data stationary. This is important because many time series analysis methods, including ARIMA, require stationary data to generate accurate results. The differentiation process can be carried out one or more times, depending on the level of data non-stationarity.

Moving Average

MA is a component that shows that the current value is influenced by the prediction error of previous values. This helps in reducing the variability in the data. In the context of rainfall, this means that the error in the rainfall prediction in the previous month can affect the rainfall prediction in the current month.

The ARIMA model is expressed in the notation ARIMA (p, d, q), where p is the order of the autoregressive component, d is the amount of differentencing performed to make the data stationary, and q is the order of the MA component. The selection of parameters p, d, and q is done through ACF and PACF analysis, as well as parameter significance tests [10]. The ARIMA model has proven effective in capturing temporal patterns in time series data, so it is often used in rainfall forecasting.

Kalman Filter

The Kalman filter is a method used to estimate the state of a dynamic system from a series of measurements containing noise. This method is very useful in situations where the available data has uncertainty, such as in rainfall measurements which are often influenced by external factors. The Kalman filter works by combining previous estimates with the latest observation data to produce a more accurate estimate.

The basic concept of the Kalman Filter involves two main steps: prediction and updating. In the prediction step, the model is used to estimate the state of the system at a later time, while in the updating step, the estimate is updated based on the latest observation data. This process is recursive, meaning that the Kalman Filter can continue to update the estimate as new data is added.

The main advantage of the Kalman Filter is its ability to provide more accurate estimates by minimizing prediction errors. This method can also be adapted for various applications, including system control, signal processing, and time series analysis. In the context of rainfall forecasting, the Kalman filter can be used to improve the accuracy of the ARIMA model by correcting the prediction based on the latest observation data.

Application of the ARIMA-Kalman Model in Rainfall Forecasting

In this study, the ARIMA model is used to analyze rainfall data in Medan City, while the Kalman filter is used to improve the accuracy of ARIMA model predictions [11]. The rainfall data used covers a fairly long period to a long time, allowing for in-depth analysis of existing patterns and trends.

The results of this study are expected to provide a significant contribution to water resources management and agricultural planning in Medan City, especially in facing the challenges of climate change and increasingly frequent extreme weather. By using the ARIMA-Kalman model, it is expected to reduce uncertainty in rainfall predictions, so that decision making in water resources management can be done better.

RESEARCH METHODS

Research Location

This research is located in Medan City, which is the capital of North Sumatra Province, Indonesia. Medan City is located in the northern part of Sumatra Island and is one of the largest cities in Indonesia with unique rainfall characteristics. The city is geographically located at coordinates 3°35′N (north latitude) and 98°40′E (east longitude).

Figure 1. Research location in Medan City.

Research Data

The data used in this study are divided into two parts, namely training data and test data.

Training Data

The training data consists of monthly rainfall observation data for Medan City from 2013 to 2022. This data was obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG). The data collected will include the total rainfall for each month during that period. This Training Data will be used to build ARIMA and Kalman filter models.

Test Data

The test data consists of monthly rainfall observation data for Medan City for 2023. This data will be used to test the accuracy of the model that has been built using the training data.

Research Tools

In this study, several tools and software used for data analysis and processing are as follows:

R Studio

R Studio is an integrated development environment (IDE) for the R programming language used for statistical analysis and data modeling. In this study, R Studio will be used to build ARIMA and Kalman filter models, as well as to perform the necessary statistical analysis.

Microsoft Word

Microsoft Word is used to compile research journals, including text processing, and compiling tables and graphs.

ArcGIS

ArcGIS is used to map rainfall data and analyze rainfall distribution patterns in the Medan City area.

Mendeley

Mendeley is used to manage and organize references and citations in research.

Research Stages

Stationary Test of Rainfall Data

Before building the ARIMA model, a data stationarity test is carried out. Time series data is said to be stationary if its statistics do not change over time, meaning there is no clear trend or seasonal pattern. A common test used to test stationarity is the augmented Dickey-fuller (ADF) test. The ADF Test formula is as follows:

$$\Delta y_t = \alpha + \beta t + \varphi y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \varepsilon_t \qquad (1)$$

If the p-value of the ADF test is less than the significance level (for example, 0.05), then the null hypothesis (non-stationary data) is rejected, and the data is considered stationary.

Rainfall Data Homogeneity Test

The homogeneity test aims to ensure that the rainfall data used comes from the same population and is not affected by systematic changes. One common method used to test homogeneity is the Pettitt Test. The Pettitt Test calculates the K statistic to determine whether there is a significant change in the data. The Pettitt test formula is as follows:

$$K = \sum_{i=1}^{n} \sum_{j=1+1}^{n} sgn(y_j - y_i) \qquad (2)$$

If the K value exceeds the specified critical value, then the data is considered non-homogeneous.

ARIMA Modeling

The ARIMA model is one of the common methods used in time series analysis for forecasting. This model combines three main components, namely AR, I, and MA.

Autoregressive

AR is an ARIMA component that shows that the current value is influenced by previous values. The AR model can be expressed by the formula:

$$y_t = c + \sum_{i=1}^{p} \theta_j y_{t-j} + \varepsilon_t \qquad (3)$$

Integrated

I is the differentencing process used to make data stationary, namely eliminating trends and seasonality. The differentencing process can be expressed as:

$$y_t{}' = y_t - y_{t-1} \qquad (4)$$

Moving Average

MA is a component that shows that the current value is influenced by prediction errors from previous values. The MA model can be expressed by the formula:

$$y_t = c + \sum_{i=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (5)$$

From equations (3), (4) and (5), the overall ARIMA model can be expressed as:

$$y_t = c + \sum_{i=1}^{p} \theta_j y_{t-j} + \sum_{i=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (6)$$

ARIMA-Kalman Filter Prediction

According to Kalman (1960), the Kalman filter can be used to predict the state of a system that changes over time by minimizing estimation errors. This process involves two main steps: prediction and updating.

Prediction Steps

State Estimation at Time

To estimate the state at time, you can use the following formula:

$$\hat{x}_t^- = F\hat{x}_{t-1} + Bu_t \quad (7)$$

Error Covariance Estimation

To estimate the error covariance, you can use the following formula:

$$P_t^- = FP_{t-1}F^T + Q \quad (8)$$

Update Steps

Calculating Kalman Gain

To calculate the Kalman gain, you can use the following formula:

$$K_t = P_t^- H^T (HP_t^- H^T + R)^{-1} \quad (9)$$

Updating State Estimates

To estimate the state, you can use the following formula:

$$\hat{x}_t = \hat{x}_t^- + K_t(z_t - H\hat{x}_t^-) \quad (10)$$

Updating Error Covariance

To update the error covariance, you can use the following formula:

$$P_t = (I - K_t H)P_t^- \quad (11)$$

Model Evaluation

The three evaluation metrics that will be used are mean absolute error (MAE) and root mean squared error (RMSE).

Mean Absolute Error

MAE measures the average absolute error between the predicted and actual values, giving a clear picture of how much the prediction is wrong in the same units as the data.

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t| \quad (12)$$

Root Mean Squared Error

RMSE calculates the square root of the average of the squared errors between the actual and predicted values. RMSE imposes a larger penalty for larger errors, making it more sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|} \quad (13)$$

RESULTS AND DISCUSSION

Stationary Test of Rainfall Data

The results of the Augmented Dickey-Fuller (ADF) test in Figure 2 on the time series data of rainfall in Medan City in 2013 – 2023 show that the statistical value of the ADF test is -6.5853 with a lag order of 5 and a p-value of 0.01. Because the p-value is smaller than the significance level of 0.05, the null hypothesis (H0), which states that the data is not stationary, is rejected. Thus, the data can be considered stationary. This means that the time series data does not have a trend or seasonal pattern that significantly affects the mean, variance, or autocorrelation, making it suitable for further analysis to carry out a prediction model.

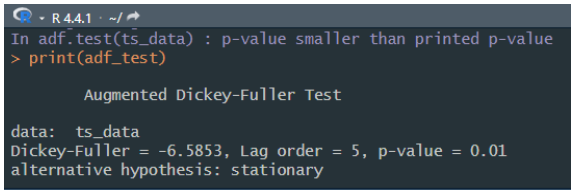

Figure 2. Stationary test using ADF.

Homogeneity Test of Rainfall Data

The results of the homogeneity test using Pettitt's Test in Figure 3 on the time series data of rainfall in Medan City in 2013 – 2023 show that the U* statistic value is 1305 with a p-value of 0.02433, which is smaller than the significance level of 0.05. This indicates that the null hypothesis (H0), which states that there is no change point in the data, can be rejected, so that there is a significant change point in the distribution of rainfall. The change point was detected at time 63, which showed a significant change in the rainfall pattern at that time. Therefore, the data is homogeneous data.

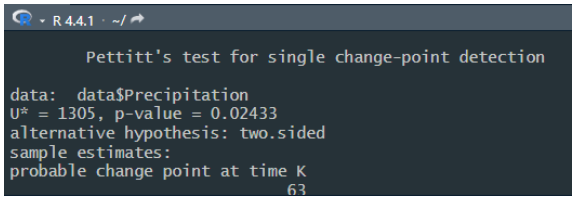

Figure 3. Homogeneity test using Pettitt's test.

ARIMA Modeling

Based on the output results from auto ARIMA, the models selected for the rainfall time series data are ARIMA (0,1,1) and (1,0,0). This model consists of a non-seasonal component (0,1,1), indicating no AR component, one differentiation to achieve stationarity, and one MA component. In addition, there is a seasonal component (1,0,0), reflecting one seasonal AR component, without seasonal differentiation, and without a seasonal MA component, with an annual seasonal pattern of 12 periods/month. The model will be used in ARIMA prediction and filtered by the Kalman filter method.

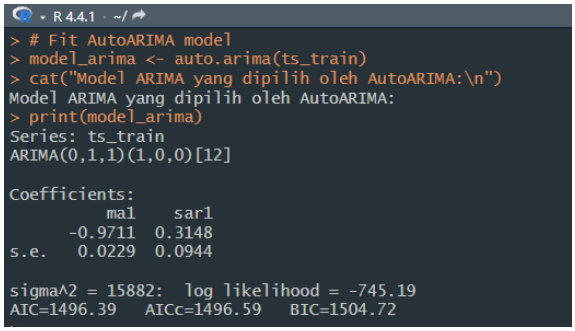

Figure 4. ARIMA modeling using AutoARIMA.

ARIMA-Kalman Prediction

Based on the graphs in Figure 5 and Table 1, the prediction results using the ARIMA and ARIMA-Kalman models show that both models can follow the actual rainfall pattern, but ARIMA-Kalman is more accurate in capturing fluctuations. In the graph, the green line (ARIMA-Kalman) is closer to the blue line (actual) than the red line (ARIMA), especially in periods with significant changes in rainfall. This can also be seen in the table, where the ARIMA-Kalman prediction value is closer to the actual data in almost every month. For example, in July 2023, the actual rainfall of 302.8 mm is predicted to be 241.6 mm by ARIMA and 265.4 mm by ARIMA-Kalman, indicating that ARIMA-Kalman provides closer results to reality.

Overall, the combination of ARIMA with the Kalman algorithm has been proven to be able to improve prediction accuracy compared to ARIMA alone. In months such as December 2023, where actual rainfall reaches 326.5 mm, ARIMA predicts 295.4 mm, while ARIMA-Kalman provides a closer prediction, namely 292.3 mm. By using historical data on Medan city rainfall from 2013 – 2022 as a basis, ARIMA-Kalman successfully captures complex rainfall patterns, making it a superior method for prediction.

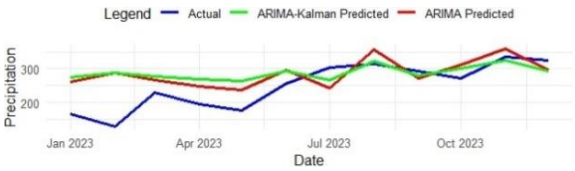

Figure 5. Plot of actual rainfall vs ARIMA vs ARIMA-Kalman.

Table 1. Prediction results using ARIMA and ARIMA-Kalman.

Month	Year	Actual	ARIMA prediction	ARIMA-Kalman prediction
1	2023	164.5	261.4100122	275.303027
2	2023	128.4	287.5098831	288.3529624
3	2023	228.3	267.4548556	278.3254487
4	2023	196.1	247.5887295	268.3923856
5	2023	175.1	236.6324508	262.9142463
6	2023	256.4	294.7825854	291.9893136
7	2023	302.8	241.6383368	265.4171893
8	2023	315.7	355.9866253	322.5913335
9	2023	293.4	270.6976623	279.946852
10	2023	271	312.7596979	300.9778699
11	2023	335.5	359.8591032	324.5275725
12	2023	326.5	295.4122566	292.3041492

Prediction Result Validation

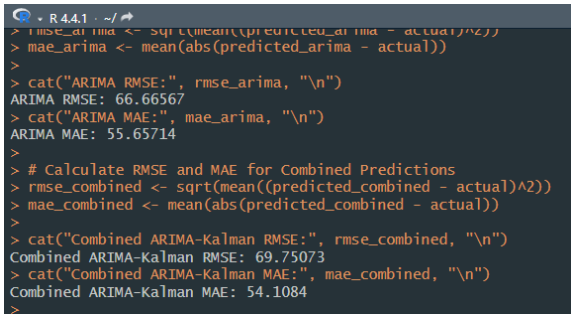

Figure 6. Validation of ARIMA and ARIMA-Kalman prediction results.

The validation results in Figure 6 show that the RMSE and MAE values for the ARIMA model are 66.67 and 55.66, respectively, while for the ARIMA-Kalman model they are 69.75 and 54.11, respectively. Although ARIMA has a lower RMSE than ARIMA-Kalman, the MAE value of ARIMA-Kalman is smaller, indicating that the average prediction of ARIMA-Kalman is closer to the actual value. The higher RMSE of ARIMA-Kalman is likely due to its sensitivity to large prediction errors at some data points.

Overall, ARIMA-Kalman can be considered superior because the smaller MAE value indicates that this model is more consistent and provides prediction results that are closer to the actual data, making it a better choice for rainfall prediction in Medan City.

CONCLUSION

Based on the research discussion, it can be concluded that the monthly rainfall data from BMKG for 2013 – 2023 is stationary and homogeneous. The models used in this ARIMA-Kalman study are ARIMA (0,1,1) and (1,0,0). The ARIMA-Kalman prediction model is superior to the ARIMA model alone.

REFERENCES

1. Siregar, N. A. (2022). Peramalan curah hujan di Kota Medan menggunakan

metode support vector regression. *J. Informatics Data Sci.*, **1**(1), 7–9.
2. Anggraini, N., Pangaribuan, B., Siregar, A. P., Sintampalam, G., Muhammad, A., Damanik, M. R. S., & Rahmadi, M. T. (2021). Analisis pemetaan daerah rawan banjir di Kota Medan Tahun 2020. *Jurnal Samudra Geografi*, **4**(2), 27–33.
3. Selamat, I. K. & Setyawan, Y. (2023). Prediksi curah hujan perbulan di Kota Yogyakarta Periode 2015 – 2019 menggunakan metode *autoregressive integrated moving average* (ARIMA) dan Kalman filter (Studi kasus : Data curah hujan Tahun 2015 – 2019). *J. Stat. Ind. dan Komputasi*, **8**(1), 15–31.
4. Aprilia, M. & Desviona, N. (2021). The implementation of a filter Kalman method forecasting rainfall obtained through model ARIMA in Kota Jambi. *Nucleus*, **2**(2), 69
5. Zulfi, M., Hasan, M., & Purnomo, K. D. (2018). The development rainfall forecasting using Kalman filter. *Journal of Physics: Conference Series*, **1008**(1), 012006.
6. Ananda, E. Y. P. & Wahyuni, M. S. (2021). Rainfall forecasting model using ARIMA and Kalman filter in Makassar, Indonesia. *Journal of Physics: Conference Series*, 2123(1), 012044.
7. Rusdi, R. (2011). Uji Akar-Akar Unit dalam Model Runtun Waktu Autoregresif. *Statistika*, **11**(2), 67–78.
8. Ruhiat, D., Andiani, D., & Kamilah, W. N. (2020). Forecasting data runtun waktu musiman menggunakan metode singular spectrum analysis (SSA). *Teorema: Teori dan Riset Matematika*, **5**(1), 47–60.
9. Samsiah, D. N. (2008). Analisis data runtun waktu menggunakan model ARIMA (p, d, q). *Skripsi UIN Sunan Kalijaga, Yogyakarta*.
10. Laamena, N. S. (2019). Prakiraan harga rumah di Kota Semarang dengan model deret waktu. *Jurnal Satya Informatika*, **4**(01), 41–52.
11. Rianto, M. & Yunis, R. (2021). Analisis runtun waktu untuk memprediksi jumlah mahasiswa baru dengan model random forest. *Paradigma*, **23**(1), 70–74.