ARIMAX model for rainfall forecasting in Pangkalpinang, Indonesia

R Amelia*, D Y Dalimunthe, E Kustiawan and I Sulistiana

Department of Mathematics, Faculty of Engineering, Universitas Bangka Belitung, Balunijuk Sub-district, Bangka Regency, Bangka Belitung Archipelago Province, Indonesia

*E-mail: ririn-amelia@ubb.ac.id

Abstract. In recent years, the weather and climate are unpredictable and the most visible is the rotation of the rainy season and the dry season. The extreme changes in rainfall can cause disasters and losses for the community. For that we need to predict the rainfall to anticipate the worst events. Rainfall is included in the periodic series data, so the forecasting method that can be used is the ARIMAX model which is ARIMA model expanded by adding the exogen variable. The aim of this research is to predict the rainfall data in Pangkalpinang City, Indonesia. The best model for each rainfall is ARIMAX (0,1,3) for monthly rainfall data and ARIMAX (0,1,2) for maximum daily rainfall. This research shows that there is an influence maximum wind speed variable to monthly rainfall and maximum daily rainfall in the Pangkalpinang City. Nevertheless, when viewed from the ARIMA and ARIMAX models based on the obtained AIC value, the ARIMAX value is still better than ARIMA. However, the prediction value using ARIMAX needs to increase again to take into account seasonal data rainfall. Then, possible to add other exogeneous factors besides maximum wind speed.

1. Introduction

Indonesia is a tropical country with high rainfall. The climate outlook for Indonesia in 2020/2021 shows that 85 percent of the monsoon zones in Indonesia have entered the rainy season since late December 2020. Monthly rainfall in most parts of Indonesia is predicted to increase compared to last year. Some areas have the potential for high rainfall from January to April 2021 and experience an increase in rainfall of 40 percent to 80 percent higher than last year rainfall [1]. In 2018, the rainfall rate in Indonesia could reach 620 mm per month. Based on these predictions, there is a possibility of extreme weather conditions in 2021 [2].

Many people know that extreme changes in rainfall can cause damage and even disaster or loss to the community. For this reason, information about future rainfall is needed that can be used to predict the worst events that can occur and be anticipated early on. Information relating to conditions in the future cannot be determined with certainty but can be predicted. Rainfall is included in the category of periodic series data, so the forecasting method that can be used is the time series method.

Time series models have advantages in certain situations. They can be used more easily for forecasting purposes because the historical sequences of observations upon study variables are readily available at equally spaced intervals over discrete point of time. These successive observations are statistically dependent and the time series modelling is concerned with techniques for the analysis of
such dependence [3]. However, regardless of the predictability performance, easy to use models have always been welcomed over the complex and ambiguous alternatives [4].

One of the time series models that can be used to predict rainfall is the Autoregressive Integrated Moving Average (ARIMA) method. The characteristic of the ARIMA method is each cascading is stationary (has a mean and constant variance also covariance lag that does not depend on where the calculation is done) [5]. The method is also called the Box-Jenkins method as developed by George Box and Gwilym Jenkins in 1976 [6]. However, this model only involves one variable without involving its dependence on other variables. The other factors can affect rainfall, such as humidity, temperature, and wind speed. Therefore, in this research the ARIMA model will be used by involving an eXogen variable. So, the model used is the ARIMAX model.

The ARIMAX models as dynamic regression where it combined linear regression and ARIMA process into one model. The application of ARIMAX model has been used in agriculture, engineering, and economics as ARIMAX model provides better forecast performance than ARIMA model [7, 8, 9].

This research intends to predict rainfall in Pangkalpinang City, Bangka Belitung Islands Province, Indonesia. These predictions involve other factors that affect it, namely wind speed. High wind speed can destroys the formation of convective clouds or rain-forming clouds that causing low rainfall. This research aims to find out the best model and can be used to predict rainfall data in Pangkalpinang City.

2. Methodology

This research uses rainfall data from Pangkalpinang City, Bangka Belitung Islands Province, Indonesia from January 2009 to June 2021. The data was obtained from the Meteorological, Climatological, and Geophysical Agency (BMKG) of the Bangka Belitung Islands Province. Then the data to be analyzed is monthly rainfall data and maximum daily rainfall data.

For ARIMA model description, almost the same methodology is used as George Box and Jenkins (1976). These models consist of three parts. The first part represents the Autoregressive model and is coded by AR (p). The other part represents the Moving Average model is denoted by MA (q). The third part represents the degree of integration (Integrated) and is denoted by the symbol (d) and is defined as the differences that must be made on the time series to be stable. The order of integration defines the order of differencing to remove the unit root from a time series to become stationary. For the aim of making a time series weak stationery (stationary in mean and variance), which Box-Jenkins method needs most, in addition to differencing, we usually need to transform the series by the natural log or square root to stabilize the variance. The mathematical formula of the model ARIMA (p, d, q) are [8, 9, 10]:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

(1)

Where \( y_t = \nabla^d Y_t \) dan \( \nabla^d \) represent the displacement factor (the differences \( d \) for time series \( Y_t \)). That is \( y_t = \nabla Y_t = Y_t - Y_{t-1} \). The ARIMA model can be rewritten as a function of the posterior displacement coefficient (B) as follows:

$$\phi(B)y_t = c + \theta(B)\epsilon_t$$

(2)

Where \( \phi(B), \theta(B) \), the polynomials are of degree \( p \) and \( q \), respectively, and are as follows:

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

(3)

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

(4)

When the stationarity has been addressed, then we need to identify the order (the \( p \) and \( q \)) of the autoregressive and moving average terms. The tools used are the autocorrelation function (ACF) and the partial autocorrelation function (PACF) [11, 12, 13]. Using the ARIMA model along with explanatory variables (ARIMAX) is belongs to the class of dynamic regressions, which encompasses a wide variety of models, including classic multiple regression models where input variables have an instantaneous effect on the output variable. For representing the ARIMAX model, we use equation as [8, 9, 10]:

...
\[ \phi(B) y_t = c + \beta X_t + \theta(B) \varepsilon_t \]  

(5)

While here is an explanatory variable at time \( t \), it is possible to have more than one explanatory variables. Also, along with the instantaneous effect, the effect of explanatory variables on the dependent variable might be delayed for several lags in some situations [10].

Then, the residual tests are carried out using the Akaike Information Criterion (AIC) test and the Ljung-Box test or also known as Q statistics. The AIC test formula is calculated as [8]):

\[ AIC(p, q) = n \log(RSS / n) + 2(p + q) \]

(6)

where \( n \) is the number of data points (observations) and RSS is the residual sums of squares. The model with smaller AIC values is considered to be the best model. Meanwhile the Ljung–Box test the magnitudes of the residuals autocorrelation for significance [8]:

- \( H_0 \): The data are independently distributed with no serial correlation
- \( H_1 \): The data are not independently distributed which they exhibit serial correlation.

The maximum likelihood method was applied for the estimation of the parameters. To assess if the residual series was white noise, the Ljung-Box Q test was performed [14]. The statistical tool used in this paper is R version 3.6.2 to analyse the time series data and building the ARIMA and ARIMAX model. The best model was chosen based on the smallest value of AIC.

3. Result and Discussion

The descriptive statistics for each data rainfall will be shown in Table 1. The data of monthly rainfall and maximum daily rainfall have high variance. That indicates that they have high variability in the period. When viewed from the maximum value of each data, the highest monthly rainfall in the period January 2009 to June 2021 reached 602 mm with maximum daily rainfall is 183 mm.

| Table 1. Descriptive Statistics of the monthly rainfall and maximum daily rainfall. |
|----------------|----------------|
|                | Monthly Rainfall | Maximum Daily Rainfall |
| N              | 150             | 150                   |
| Mean           | 189.19          | 47.58                 |
| Median         | 184.60          | 42.50                 |
| Mode           | 0.00            | 0.00                  |
| Std. Deviation | 121.85          | 30.59                 |
| Variance       | 14846.72        | 935.82                |
| Skewness       | 0.438           | 1.120                 |
| Kurtosis       | -0.036          | 2.346                 |
| Range          | 602.00          | 183.90                |
| Minimum        | 0.00            | 0.00                  |
| Maximum        | 602.00          | 183.00                |

If analyzed from the results of the data plot in Figure 1, it can be seen that in 2016 there was a fairly high spike in rainfall. Based on the figure, each data set is not stationary. To progress into the forecasting process, we can perform data transformations or use differencing data.

![Figure 1. Plot data monthly rainfall and maximum daily rainfall in period January 2009 – June 2021.](image-url)
In Figure 2 it can be seen that the monthly rainfall data and the maximum daily rainfall data can be assumed to be stationary after differencing one time according to the existing data conditions. Then, the degree of integration (d) which is known from the results of the differencing process will be used in formula (1).

![Figure 2](image)

**Figure 2.** Time series plot from January 2009 – June 2021 (rainfall data plots that have been made different once).

After that, we compare the estimate autocorrelation functions (ACF) and partial autocorrelation functions (PACF) with various theoretical ACF and PACF to find match it. In choosing models ARIMA, we want the model that fits the realization with the smallest number of the estimated parameters. Based on the ACF and PACF plots in Figure 3, there are several ARIMA models indicated.

![Figure 3](image)

**Figure 2.** (a) ACF and (b) PACF for monthly rainfall data; (c) ACF and (d) PACF for maximum daily rainfall data.

The models are then selected based on the smallest AIC value (see Table 2). Based on the ACF and PACF plots in Figure 3, there are several ARIMA models. The models are then selected based on the smallest AIC value (Table 2). ARIMA models that have the smallest AIC for monthly rainfall data and maximum daily rainfall are ARIMA (0,1,3) and ARIMA (0,1,2).

Then the residual assumption test is carried out on the ARIMA model that has been select. Based on the QQ plot (see Figure 3) and a p-value of more than 5%, we can be conclude that the ARIMA model chosen for both monthly and maximum daily rainfall has met the residual assumptions. It means the residuals are independent. So, after obtaining the best ARIMA model, it will ARIMAX modeling is carried out by simultaneously combining the ARIMA model that has been obtained with the eXogen
variable, namely the maximum wind speed. The maximum wind speed increase is particular problem for small tropical island countries.

Table 2. ARIMA models of the monthly rainfall and maximum daily rainfall.

| Models       | AIC     | Models       | AIC     |
|--------------|---------|--------------|---------|
| ARIMA (0,1,2) | 1856.65 | ARIMA (0,1,1) | 1519.54 |
| **ARIMA (0,1,3)** | **1844.43** | **ARIMA (0,1,2)** | **1455.71** |
| ARIMA (3,1,2) | 1844.99 | ARIMA (3,1,1) | 1475.45 |
| ARIMA (3,1,3) | 1846.64 | ARIMA (3,1,2) | 1456.86 |

Table 3. The Ljung–Box test the magnitudes of the residuals autocorrelation.

| Best Model       | Data                   | p-value |
|------------------|------------------------|---------|
| ARIMA (0,1,3)    | Monthly Rainfall       | 0.3267  |
| ARIMA (0,1,2)    | Maximum Daily Rainfall | 0.1842  |

Figure 3. QQ-plots, sample ACF and PACF of the residuals (a) monthly rainfall and (b) maximum daily rainfall data.

Then the residual assumption test can be seen in Table 4. When we viewed the AIC value, it can be seen that the ARIMAX model has an AIC value that is smaller than the ARIMA model. This indicates that the ARIMAX model is still better than the ARIMA model in modeling monthly rainfall and maximum daily rainfall.

Table 4. The Ljung–Box test the magnitudes of the residuals autocorrelation model ARIMAX

| Best Model       | Data                   | log likelihood | AIC     |
|------------------|------------------------|----------------|---------|
| ARIMAX (0,1,3)   | Monthly Rainfall       | -914.51        | 1839.02 |
| ARIMAX (0,1,2)   | Maximum Daily Rainfall | -720.28        | 1448.56 |

The ARIMAX model will be used to forecast monthly rainfall data and maximum daily rainfall in the city of Pangkalpinang for the period July 2021 until July 2022 with the results in Table 5. Based on the prediction results of the ARIMAX model, it can show that monthly rainfall has a fluctuating trend. Until July 2022 the predicted value of monthly rainfall only reaches 90.1 mm. however, when compared between the AIC model ARIMA (table 3) and ARIMAX (model E), the ARIMAX model is more suitable for predicting rainfall data. The ARIMAX model does not include seasonal factors in the data processing. Therefore, further analysis is needed to combine exogenous and seasonal factors in the analysis of rainfall data.
Table 5. The forecast monthly rainfall data and maximum daily rainfall

| Month       | Monthly Rainfall (mm) | Maximum Daily Rainfall (mm) |
|-------------|-----------------------|----------------------------|
| July 2021   | 69.85                 | 27.12                      |
| August 2021 | 77.36                 | 30.76                      |
| September 2021 | 72.11            | 30.30                      |
| Oktober 2021 | 72.64               | 32.50                      |
| November 2021 | 79.25              | 31.04                      |
| Decembe 2021 | 79.21               | 29.47                      |
| January 2021  | 77.61               | 30.52                      |
| February 2021 | 72.41              | 29.88                      |
| March 2022    | 78.65               | 33.91                      |
| April 2022    | 88.98               | 30.05                      |
| May 2022      | 85.49                | 28.84                      |
| June 2022     | 59.17                | 29.84                      |
| July 2022     | 91.10                | 32.28                      |

4. Conclusion
This research shows that there is an influence maximum wind speed variable to monthly rainfall and maximum daily rainfall in the Pangkalpinang City area, Indonesia. Nevertheless, when viewed from the ARIMA and ARIMAX models based on the obtained AIC value, the ARIMAX value is still better than ARIMA. However, the prediction value using ARIMAX needs to increase again to take into account seasonal data rainfall. Then, possible to add other eXogenous factors besides maximum wind speed.

References
[1] Dry season and the rainy season in Indonesia, http://www.bmkg.go.id. Retrieved April 1, 2021.
[2] Extreme weather is predicted to hit Bangka Belitung for the next three days, https://regional.kompas.com. Retrieved April 1, 2021
[3] Sirohi, Sanjeev, and Urmil V 2016 International Journal of Agricultural and Statistical Sciences 12 327
[4] Islam F, and Imteaz M 2020 Hydrology 7 52
[5] Gujarati DN 2009 Essential of Econometrics (New York: McGraw-Hill Co.)
[6] Lutkepohl H 2005 New Introduction to Multiple Time Series Analysis (Berlin: Springer Science + Business Media, Inc.,)
[7] Cui H, and Peng X 2015 Mathematical Problems in Engineering 2015
[8] Ling SCA., Ramba H., Mohd. JAK, Kim PC, and Chong MH 2016 AIP Conference Proceedings 1739 020077
[9] Abd HT, Essa AK, and Jassim FM 2021 International Journal on Advanced Science, Engineering and Information Technology 11 465
[10] Ahmadi AK 2020 The International Journal of Business Management and Technology 4 5
[11] Box GE, and Jenkins GM, 1976 Time Series Analysis: Forecasting and Control Revised Ed.Holden-Day (San Francisco CA USA)
[12] Cryer JD, Chan K 2008 Time Series Analysis With Applications in R. (Springer)
[13] Kendeya T, Kelemu N, Zergaw G, and Kebede T 2018 International Journal of Computing Science and Applied Mathematics 4 42
[14] Hossain S, Ahmed S, and Uddin MJ 2020 Science of The Total Environment 10

Acknowledgments
We gratefully acknowledge the funding from Universitas Bangka Belitung through the RKAKL FT for the publication of this paper.