Forecasting the amount of rainfall in West Kalimantan using Generalized Space-time Autoregressive model

R Utami*, N Nurhayati, S Maryani
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Jenderal Soedirman University

*E-mail: riani.utami@mhs.unsoed.ac.id

Abstract. The generalized space-time autoregressive or GSTAR is a space-time model which can be used to analyze time-series data in several locations considered to be correlated. In this research, the GSTAR model is applied to forecast the amount of rainfall in West Kalimantan, especially at Sintang Station, Melawi Station, and Ketapang Station. The data used for modeling is data of the amount of rainfall for the period January 2013-December 2017, while the data used for model validation is data for the period January 2018-December 2018. The spatial weights used are uniform weights, inverse distance weights, and normalized cross-correlation weights, and the estimation method used is the ordinary least square or OLS estimation. The best model is selected based on the smallest RMSE (root mean square error). The results showed that all spatial weights gave the same good GSTAR(1:1) model because they had almost the same RMSE value. Thus, this model can be used to forecast the amount of rainfall for the period January 2019-December 2019. The forecast results show that for Sintang Station and Melawi Station, the highest amount of rainfall is estimated to occur in March 2019 and the lowest will occur in August 2019. Meanwhile, for Ketapang Station, the highest amount of rainfall is estimated to occur in January 2019 and the lowest will occur in September 2019.

1. Introduction
Rain is part of the water cycle and is the main water source that supplies water to the earth’s surface. The volume of rain that occurs at a certain time is called the amount of rainfall. The amount of rainfall in an area is usually influenced by the amount of rainfall in other areas. This makes the amount of rainfall in a location with other nearby locations tend to have almost the same amount of rainfall. Apart from being influenced by other regions, rainfall is also influenced by previous times. Therefore, rainfall data are considered as spatio-temporal data.

The amount of rainfall at a location can be predicted using a space-time model, one of which is the Generalized Space-time Autoregressive (GSTAR) model. The GSTAR model was chosen because this model can be used to predict a spatial phenomenon or the phenomenon of the occurrence of an area that is influenced by the events of other regions. The GSTAR model is a generalization of the Space-Time Autoregressive (STAR) model. In the STAR model, the parameters are assumed to be the same for all locations, while GSTAR is assumed to be different. Therefore, the GSTAR model is considered more realistic than the STAR model [1]. In the GSTAR model, the spatial relationship is represented...
by the spatial weight matrix $W = (w_{ij})$ measuring $N \times N$, with $w_{ij}$ representing the spatial relationship between location $i$ and location $j$ [1].

The GSTAR model with time order $p$ and spatial order $\lambda_1, \lambda_2, \ldots, \lambda_p$ denoted $\langle p; \lambda_1, \lambda_2, \ldots, \lambda_p \rangle$ can be written as follows.

$$ z(t) = \sum_{k=1}^{p} [\Phi_{k0} + \sum_{\ell=1}^{\lambda_k} \Phi_{k\ell} W^{(\ell)}] z(t-k) + \varepsilon(t), \quad (1) $$

where $z(t) = [z_1(t), z_2(t), \ldots, z_N(t)]$ for $z(t) = [z_1(t), z_2(t), \ldots, z_N(t)]$ is the observation vector with dimension $N$ at time $t$. Furthermore, $\Phi_{k0} = \text{diag}(\varnothing^{(1)}_{k0}, \ldots, \varnothing^{(N)}_{k0})$ is a diagonal matrix with the element $\varnothing^{(i)}_{k0}$ denotes the autoregressive parameter at lag time $k$ in location $i$, and $\Phi_{k\ell} = \text{diag}(\varnothing^{(1)}_{k\ell}, \ldots, \varnothing^{(N)}_{k\ell})$ is a diagonal matrix with $\varnothing^{(i)}_{k\ell}$ denotes the space-time parameter at lag time $k$ and spatial order $\ell$ in location $i$. Meanwhile, $W^{(\ell)}$ is a spatial weight matrix of order $\ell$ with the elements $0 \leq w_{ij} \leq 1$ and $\sum_{j} w_{ij} = 1$. $\varepsilon(t) = (\varepsilon_1(t), \varepsilon_2(t), \ldots, \varepsilon_N(t))$ is the error vector with $N$ dimension at time $t$ with $\varepsilon_i(t)$ represents the model error at location $i$ at time $t$ which is following the white noise assumption (constant mean and variance, and uncorrelated) and normally distributed.

Research about GSTAR has been applied in various fields, such as GDP data modeling in West European Countries [2], and forecasting oil production data at Volcanic Layer Jatibarang, West Java, Indonesia [3]. Motivated by this result, this research will focus on forecasting the amount of rainfall in West Kalimantan using the GSTAR model using three spatial weights, uniform weight, inverse distance weight, and normalized cross-correlation weight. In 2017, [4] also predicted the amount of rainfall in West Kalimantan but different in location and the type of spatial weights.

### 2. Method

The data used in this research is rainfall data at three stations in West Kalimantan, that is Sintang Station, Melawi Station, and Ketapang Station. Data obtained from Badan Pusat Statistik (BPS), West Kalimantan Province from January 2013 to December 2018. The variables used in this research are presented in Table 1.

| Variable | Description | Unit |
|----------|-------------|------|
| $z_1(t)$ | The amount of rainfall at Sintang Station at time $t$ after is subtracted by the average value | millimeter |
| $z_2(t)$ | The amount of rainfall at Melawi Station at time $t$ after is subtracted by the average value | millimeter |
| $z_3(t)$ | The amount of rainfall at Ketapang Station at time $t$ after is subtracted by the average value | millimeter |

To analyze the data, we use free statistical software, R-4.0.3. The procedure of this research is

1. Describe the data.
2. Divide the data into two groups, namely:
   a. In-sample data: January 2013 - December 2017 (60 observations)
   b. Out-sample data: January 2018 - December 2018 (12 observations)
3. Check the data stationarity
4. Determine the weight matrix in the GSTAR model, namely uniform weight, inverse distance weight, normalized cross-correlation weight.
5. Choose the GSTAR model’s order by looking at the ACF plot and the PACF plot.
6. Estimate model parameters using the OLS (ordinary least square) method.
7. Check residual assumptions.
8. Calculate the RMSE value for each model.
9. Select the model based on the smallest RMSE value.
10. Forecast the amount of rainfall in the next period.

3. Result and Discussion
This section discusses the research results included data description, stationarity testing, spatial weight matrix, GSTAR model identification, parameter estimation, checking the residual assumptions, calculating the RMSE value, and forecasting the amount of rainfall in the next period.

3.1. Data Description
The statistics of the rainfall data at three stations in Kalimantan Barat are presented in Table 2. From the table, we found that Melawi Station has the highest average rainfall, while Sintang Station has the lowest average rainfall. The range between the minimum and maximum values is quite large. Comparing to other stations, the standard deviation in Melawi is also quite large. It shows that data distribution occurred in Melawi Station is spread widely.

| Statistic      | Sintang Station | Melawi Station | Ketapang Station |
|----------------|-----------------|----------------|------------------|
| Average        | 255.83          | 324.60         | 231.50           |
| Standard Deviation | 121.99         | 145.45         | 141.29           |
| Minimum        | 26.40           | 61.00          | 19.50            |
| Maximum        | 633.50          | 715.80         | 556.00           |

3.2 Stationarity Test
Data is stationary if the mean and variance are constant [5]. Data is “stationary in the mean” if the data is stable fluctuating around the average. Data is “stationary in the variance” if the data fluctuates from time to time, but the average does not need to be constant. The stationarity can be checked by examining the time series data plot and ADF test. The data plot can be seen in Figure 1.

Figure 1. Plot of the rainfall data at Sintang Station, Melawi Station, and Ketapang Station.

Figure 1 shows that the fluctuation of the plot is around the average value and the fluctuation is stable, it can be concluded that the rainfall data at three locations are stationary in the mean and variance.

Stationary identification by looking only at a plot of the data pattern is often subjective. Alternatively, data stationarity can be identified using the ADF test. The ADF test was carried out using tseries package on the R-4.0.3 software. The ADF test results are shown in Table 3. Table 3 shows that each variable has a p-value <0.05, so it can be concluded that the data is stationary in the mean and variance.

| No | Variable | p-value | Decision     |
|----|----------|---------|--------------|
| 1. | $z_1(t)$ | 0.040   | Stationary   |
| 2. | $z_2(t)$ | 0.010   | Stationary   |
| 3. | $z_3(t)$ | 0.010   | Stationary   |
3.3 Spatial Weight Matrix

The weight matrix of the GSTAR model used in this research is the uniform weight matrix, the inverse distance weight matrix, and the normalized cross-correlation weight matrix.

3.3.1. Uniform weight matrix. The element of a uniform weight matrix in the GSTAR model is calculated based on the number of nearby locations. According to [6], uniform weight is calculated by the formula

$$ w_{ij} = \frac{1}{n_i} $$

(2)

where $n_i$ is the number of the nearby location to $i$ location in the first-order spatial lag.

| Location | Nearby location       | Number of the nearby location |
|----------|-----------------------|------------------------------|
| Sintang  | Melawi                | 1                            |
| Melawi   | Sintang and Ketapang  | 2                            |
| Ketapang | Melawi                | 1                            |

From Table 4 and equation (2), we obtain the following uniform weight matrix

$$ W = \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix} $$

3.3.2. Inverse distance matrix. The element of an inverse distance weight matrix is calculated based on the actual distance. The formula is

$$ w_{ij} = \frac{\frac{1}{d_{ij}}}{\sum_{j=i}^{n} \frac{1}{d_{ij}}} $$

(3)

and $w_{ij} = 0$ for $j = i$. In equation (3), $d_{ij}$ is the distance from location $i$ to location $j$ [7].

| No  | Label | Definition                                      | Value   |
|-----|-------|------------------------------------------------|---------|
| 1.  | $d_{AB}$ | the distance from Sintang Station to Melawi Station | 72 km   |
| 2.  | $d_{AC}$ | the distance from Sintang Station to Ketapang Station | 573 km  |
| 3.  | $d_{BC}$ | the distance from Melawi Station to Ketapang Station | 621 km  |

From Table 5 and equation (3), we get the following inverse distance weight matrix

$$ W = \begin{bmatrix} 0 & 0.888 & 0.112 \\ 0.896 & 0 & 0.104 \\ 0.520 & 0.480 & 0 \end{bmatrix} $$

3.3.3. Normalized cross-correlation weight matrix. The element of a cross-correlation weight matrix between two locations is calculated for each time lag $k$. According to [8], the normalized cross-correlation weight is calculated by the formula (4).

$$ w_{ij} = \frac{r_{ij}(k)}{\sum_{k \neq i} |r_{ik}(k)|} $$

(4)

where
\[ r_{ij}(k) = \frac{\sum_{t=k+1}^{T} (z_i(t) - \bar{z}_i)(z_j(t-k) - \bar{z}_j)}{\sqrt{\sum_{t=1}^{T} (z_i(t) - \bar{z}_i)^2 \sum_{t=1}^{T} (z_j(t) - \bar{z}_j)^2}} \]

is the cross-correlation between location \(i\) and \(j\). The cross-correlation value for the first lag time is presented in Table 6.

**Table 6. Cross correlation values for \(k = 1\)**

| \(r_{AB}(k)\) | Description | \(k = 1\)  |
|----------------|-------------|------------|
|                | Cross correlation value between Sintang Station and Melawi Station | 0.236      |
|                | Cross correlation value between Sintang Station and Ketapang Station | 0.140      |
|                | Cross correlation value between Melawi Station and Sintang Station | 0.316      |
|                | Cross correlation value between Melawi Station and Ketapang Station | 0.107      |
|                | Cross correlation value between Ketapang Station and Sintang Station | 0.235      |
|                | Cross correlation value between Ketapang Station and Melawi Station | 0.244      |

From equation (4) and Table 6, we get the normalized cross-correlation weight as follows:

\[ W = \begin{bmatrix} 0 & 0.628 & 0.372 \\ 0.747 & 0 & 0.253 \\ 0.491 & 0.509 & 0 \end{bmatrix} \]

### 3.4 Model Identification

The order of the GSTAR model can be identified based on the ACF plot and the PACF plot for each location. ACF plot and PACF plot can be seen in Figure 2 and Figure 3.

**Figure 2.** ACF plot for \(z_1(t)\), \(z_2(t)\), and \(z_3(t)\).

**Figure 3.** PACF plot for \(z_1(t)\), \(z_2(t)\), and \(z_3(t)\).

Based on Figure 2 and Figure 3, it can be seen that the ACF plot and the PACF plot cut off at the first lag so that the possible order of the GSTAR model is only the GSTAR(1:1) model.

### 3.5 Model Estimation

The OLS method is the best estimation method that is unbiased, linear, and the best (BLUE). According to [1], GSTAR model for each location \(i\) can be written as

\[ z_i = X_i \phi_i + \epsilon_i \] (5)

where,

\[ z_i = \begin{bmatrix} z_i(p) \\ z_i(p+1) \\ \vdots \\ z_i(T) \end{bmatrix}, \quad X_i = \begin{bmatrix} v_i^{(0)}(p-1) & \ldots & v_i^{(2)}(p-1) & \ldots & v_i^{(0)}(0) & \ldots & v_i^{(2)}(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_i^{(0)}(T-1) & \ldots & v_i^{(2)}(T-1) & \ldots & v_i^{(0)}(T-p) & \ldots & v_i^{(2)}(T-p) \end{bmatrix} \]
\[ \phi_i' = (\phi^{(i)}_{10}, \ldots, \phi^{(i)}_{p_1}, \phi^{(i)}_{20}, \ldots, \phi^{(i)}_{p_2}, \ldots, \phi^{(i)}_{2p}, \ldots, \phi^{(i)}_{p_p}), \text{ and } \epsilon_i = \begin{bmatrix} \epsilon_i(p) \\ \vdots \\ \epsilon_i(T) \end{bmatrix}. \]

Therefore, the GSTAR model for all locations can be presented as the following linear model
\[ z = X\phi + \epsilon, \] (6)

where \( z = (z_1', \ldots, z_N') \), \( X = \text{diag}(X_1, \ldots, X_N) \), \( \phi = (\phi_1', \ldots, \phi_N') \), and \( \epsilon = (\epsilon_1', \ldots, \epsilon_N') \).

Estimation of parameters using the OLS method can be obtained by minimizing the number of squares of errors defined as
\[ X'X\phi = X'z \] (7)

The estimation results of the GSTAR model parameters using uniform weights, inverse distance weights, and normalized cross-correlation weights can be seen in Table 7.

| Parameter | Uniform weight | Inversion distance weight | Normalized cross-correlation weight |
|-----------|----------------|---------------------------|-----------------------------------|
| \( \phi^{(1)}_{10} \) | 0.306 | 0.299 | 0.292 |
| \( \phi^{(2)}_{10} \) | 0.235 | 0.205 | 0.206 |
| \( \phi^{(1)}_{11} \) | 0.522 | 0.544 | 0.542 |
| \( \phi^{(2)}_{11} \) | 0.098 | 0.111 | 0.127 |
| \( \phi^{(1)}_{12} \) | 0.245 | 0.289 | 0.294 |
| \( \phi^{(2)}_{12} \) | 0.089 | 0.043 | 0.048 |

Based on equation 1, the GSTAR model equation for each location can be calculated by substituting the estimated value in Table 7 to the matrix \( \Phi_{k0} \) and matrix \( \Phi_{k1} \). The GSTAR model equation for each location can be written as follows

1. **GSTAR(1:1) for uniform weight**
\[ z_1(t) = [0.306 z_1(t-1) + 0.098 z_2(t-1)] + \epsilon_1(t), \]
\[ z_2(t) = [0.235 z_2(t-1) + 0.123 z_1(t-1) + 0.122 z_3(t-1)] + \epsilon_2(t), \]
\[ z_3(t) = [0.522 z_3(t-1) + 0.089 z_2(t-1)] + \epsilon_3(t). \]

2. **GSTAR(1:1) model for inverse distance weight**
\[ z_1(t) = [0.299 z_1(t-1) + 0.098 z_2(t-1) + 0.012 z_3(t-1)] + \epsilon_1(t), \]
\[ z_2(t) = [0.205 z_2(t-1) + 0.259 z_1(t-1) + 0.030 z_3(t-1)] + \epsilon_2(t), \]
\[ z_3(t) = [0.544 z_3(t-1) + 0.022 z_1(t-1) + 0.021 z_2(t-1)] + \epsilon_3(t). \]

3. **GSTAR(1:1) model for normalized cross-correlation weight**
\[ z_1(t) = [0.292 z_1(t-1) + 0.079 z_2(t-1) + 0.047 z_3(t-1)] + \epsilon_1(t), \]
\[ z_2(t) = [0.206 z_2(t-1) + 0.219 z_1(t-1) + 0.074 z_3(t-1)] + \epsilon_2(t), \]
\[ z_3(t) = [0.542 z_3(t-1) + 0.024 z_1(t-1) + 0.024 z_2(t-1)] + \epsilon_3(t). \]

The GSTAR model equation with a negative coefficient interprets that the amount of rainfall in an area in the previous negatively affected the amount of rainfall in the current period. Conversely, a positive coefficient value shows that the amount of rainfall in an area in the previous period has a positive effect on the amount of rainfall in the current period. For example, the GSTAR(1:1) model with uniform weight at Ketapang Station can be interpreted, if the amount of rainfall last month at Ketapang Station increased by 1 mm, while at other stations and other times it was constant, the rainfall at the Ketapang Station in the next period will decrease by 0.522 mm.
3.6 Checking Model Assumption

The model assumptions in this research are white noise assumption and normality assumption. White noise assumption can be inferred from the stationary residual around zero, and plot of residuals ACF and residuals PACF which lie on their confidence interval line. Normality assumption can be deduced from the symmetric histogram linear Q-Q plot of residuals. The results of checking GSTAR(1:1) model assumption for each spatial weight is presented in Table 8.

Table 8 shows that GSTAR(1:1) model assumption for each spatial weight matrix has been fulfilled. It means that the model is suitable to forecast the amount of rainfall in West Kalimantan.

3.7 Calculation of RMSE Value

The criterion for selecting the best model is determined by the RMSE value. The best model has the smallest RMSE value. The RMSE value can be calculated by the following formula

\[
RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (z_t - \hat{z}_t)^2}
\]

where \(z_t\) is the observed value at time \(t\), and \(\hat{z}_t\) is the predicted value at time \(t\). The criteria for selecting the best model are determined by taking into account the RMSE value based on the in-sample residual. The results of forecasting accuracy can be known based on the value based on the out-sample residual. The best model is the model that has the smallest RMSE value.

Table 9 shows that the in-sample RMSE and the out-sample RMSE for each spatial weight gave almost the same RMSE value. This proves that all the models produced are equally good, so that in predicting the amount of rainfall in the next period it is carried out using the GSTAR(1:1) model with three spatial weights, namely uniform weights, inverse distance weight, and normalized cross-correlation weight.

3.8 Forecasting The Next Period

A good forecast is a forecast that produces a forecast value that is not far from the true value. Forecasting the amount of rainfall in the next period is obtained from the GSTAR(1:1) model with three spatial weights there are uniform weight, inverse distance weight, and normalized cross-correlation weight. Forecasting the amount of rainfall for the next periods is presented in Table 10.

Based on Table 10, it can be seen that the results of forecasting the amount of rainfall in January 2019 to December 2019 for Sintang Station, Melawi Station, and Ketapang Station are fluctuating results. The forecast results for Sintang Station and Melawi Station, the highest amount of rainfall is estimated to occur in March 2019 and the lowest will occur in August 2019. As for Ketapang Station,
The highest amount of rainfall is estimated to occur in January 2019 and the lowest will occur in September 2019.

**Table 10.** Forecasting the amount of rainfall for the next 12 months for each spatial weight

| Time       | Uniform weight | Inverse distance weight | Normalized cross-correlation |
|------------|----------------|-------------------------|------------------------------|
|            | Sintang Station | Melawi Station | Ketapang Station | Sintang Station | Melawi Station | Ketapang Station | Sintang Station | Melawi Station | Ketapang Station |
| 30-Jan-19  | 296.11          | 414.07                 | 406.41                      | 299.41          | 390.39         | 402.08          | 305.84          | 401.07         | 402.16          |
| 28-Feb-19  | 255.36          | 320.87                 | 245.98                      | 255.70          | 320.40         | 249.71          | 257.58          | 321.44         | 249.51          |
| 30-Mar-19  | 315.17          | 412.85                 | 353.39                      | 316.79          | 405.80         | 346.60          | 318.81          | 409.67         | 347.09          |
| 30-Apr-19  | 310.85          | 342.29                 | 204.06                      | 308.93          | 370.85         | 205.98          | 305.60          | 361.49         | 206.46          |
| 30-May-19  | 295.32          | 364.75                 | 361.26                      | 297.48          | 361.39         | 372.42          | 305.98          | 367.00         | 371.97          |
| 30-Jun-19  | 196.38          | 268.40                 | 157.32                      | 195.96          | 260.08         | 156.64          | 194.35          | 260.99         | 156.33          |
| 30-Jul-19  | 229.49          | 301.71                 | 254.23                      | 230.59          | 291.87         | 261.16          | 235.00          | 296.56         | 260.58          |
| 30-Aug-19  | 178.01          | 230.08                 | 116.19                      | 176.97          | 227.90         | 122.10          | 175.86          | 226.97         | 121.46          |
| 30-Sep-19  | 196.11          | 231.35                 | 99.46                       | 194.25          | 240.84         | 106.92          | 191.90          | 236.07         | 106.43          |
| 30-Oct-19  | 197.97          | 240.58                 | 102.85                      | 196.18          | 247.76         | 107.10          | 192.99          | 243.31         | 106.75          |
| 30-Nov-19  | 216.31          | 270.37                 | 175.56                      | 215.83          | 271.06         | 181.78          | 216.34          | 270.56         | 181.32          |
| 30-Dec-19  | 311.74          | 359.68                 | 251.46                      | 310.94          | 378.30         | 252.77          | 309.95          | 373.03         | 253.15          |

4. Conclusion
The GSTAR model suitable for rainfall forecasting in West Kalimantan is the GSTAR(1:1) model with uniform weight, inverse distance weight, or normalized cross-correlation weight. The forecast results show that for Sintang Station and Melawi Station, the highest amount of rainfall is estimated to occur in March 2019 and the lowest will occur in August 2019. As for Ketapang Station, the highest amount of rainfall is estimated to occur in January 2019 and the lowest will occur in September 2019.

Acknowledgment
The research is funded by the Directorate of Research and Community Service (DRPM) Deputy of Research and Development Strengthening, Kemenristek-BRIN Indonesia, according to Research Contract No. 176/SP2H/AMD/LT/DRPM/2020.

References
[1] Borovkova S, Ruchjana B N, and Lopuhaa H 2008 Consistency and Asymptotic Normality of Least Squares Estimators in Generalized Space-time Models Stat. Neerl. 62(4) 482-508.
[2] Nurhayati N, Pasaribu U S, and Neswan O 2012 Application of Generalized Space-Time Autoregressive Model on GDP Data in West European Countries J. Probab. Stat. 2012 1-16.
[3] Ruchjana B N, Borovkova S A, and Lopuhaa H P 2012 Least Squares Estimation of Generalized Space-Time Autoregressive (GSTAR) Model and Its Properties The 5th International Conference on Research and Education in Mathematics, American Institute of Physics pp. 61-64.
[4] Adan I, Kusnandar D, and Perdana H 2017 Application GSTAR(1,1) Model for Rainfall Data Buletin Ilmiah Matematika, Statistik, dan Terapannya (Bimaster).6(3) 159-166.
[5] Box G, Jenkins G M, Reinsel G, and Ljung G 2016 Time Series Analysis: Forecasting and Control, fifth edition (New Jersey).
[6] Yaffee R and McGee M 2000 Time Series Analysis and Forecasting (Academic Press, Inc. New York).
[7] Ruchjana B N 2002 Generalized Space-Time Autoregressive Model and Its Application to Oil Production Forum Statistika dan Komputasi (Bogor: IPB).
[8] Anggraeni D, Prahutama A, and Andari S 2013 Application of Generalized Space-time Autoregressive (GSTAR) for Volume of Vehicles Entering The Semarang Toll Road Media Statistika 6(2) 71-80.