Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190309

Development of GSTARIMA-ARCH Model for Rainfall Forecasting in Java Island using Big Data Analytics

Putri Monika[1], *Budi Nurani Ruchjana*[2,*], *Atje Setiawan Abdullah*[3] *and Rahmat Budiarto*[4]

[1]Doctoral Student at Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia
[3]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia
[4]College of Computing and Information, Al-Baha University, Alaqiq 65779-7738, Saudi Arabia

Received: 2 Dec. 2024, Revised: 24 Jan. 2025, Accepted: 10 Feb. 2025
Published online: 1 May 2025

Abstract: Spatio-Temporal (ST) model is developed with the combination of ARCH to overcome non-constant error variance through data analytics life cycle method. The subsequent model developed is Generalized Space-Time Autoregressive (GSTAR), which simultaneously considers spatial and temporal dependence in rainfall data. Following this process, GSTAR is combined with ARCH to overcome the assumption of heteroscedasticity in rainfall. Therefore, this research aimed to develop a combined GSTARIMA-ARCH to forecast rainfall on Java Island, which is characterized by high rainfall intensity. The methodology used in this research was analysis and modeling of GSTARIMA-ARCH in line with data analytics life cycle, particularly designed for handling Big Data in rainfall analysis. Consequently, the results showed that forecast produced by GSTARIMA-ARCH was more accurate than other conventional models. Moreover, this model could be used as a tool for decision-making by relevant agencies, as well as in the field of meteorology for exploring weather and rainfall. Finally, the model is also applicable in water resources management and natural disaster mitigation in Java Island.

Keywords: GSTARIMA, ARCH, Heteroscedasticity, Big Data Analytics, Forecasting, Rainfall

1 Introduction

Time series data from different locations are often considered independent of each other [1,2], allowing for separate analysis and forecasting of each data as univariate time series. Typically, a widely used model for analyzing univariate time series data is Autoregressive (AR). In the context of multiple locations that share the same time sequence, AR can be combined and analyzed simultaneously as a multivariate time series. For example, a first-order AR model for rainfall forecasting in Cities A, B, and C can be analyzed collectively using Vector Autoregressive (VAR). However, both univariate and multivariate time series have significant limitations, as they do not account for spatial relationship or influences between observation location. The model generally assumes that the current observation at a location is mainly a function of the previous, disregarding potential spatial dependencies.

Spatio-Temporal (ST) model is an extension of time series that includes data sorted by location and time [3]. This model considers both the influence of location and time of observation during analysis and forecasting. It is crucial to be aware that the method for analyzing ST was in line with the framework developed by George E. P. Box and Gwilym M. Jenkins, commonly known as Box-Jenkins method [4]. This method consists of three iterative stages, namely identification, parameter estimation, and diagnostic testing. The stages provide a structural guide for explorers and practitioners to conduct forecasting using ST model.

Space-Time Autoregressive (STAR) is an extension of Box-Jenkins method that includes both spatial and temporal relationship in data analysis. This model assumes homogeneous characteristics across locations and is applied to stationary data. Furthermore, model uses location weights to differentiate spatial influences and also adapts the same AR and ST parameters for each

* Corresponding author e-mail: budi.nurani@unpad.ac.id

location [5,6].However, STAR model's assumptions of homogeneity often show limitations, particularly when applied to real-world data. Research has found that the characteristics between locations are heterogeneous, with variations in factors such as distance, relief, population, and economic condition. Relating to this research, Ruchjana (2002) improved the assumptions of model by developing Generalized Space-Time Autoregressive (GSTAR) which assumes heterogeneous characteristics between locations and stationary data [7]. As opposed to STAR, GSTAR allows AR and ST parameters to vary by location, thereby making the model more flexible and potentially more accurate for diverse spatial data. For example, GSTAR has been used to forecast production in three petroleum wells of Jatibarang Field volcanic layer.

GSTAR is rapidly gaining popularity among practitioners and explorers in the fields of stochastic modeling, mathematics, and statistics. Since its initial development by Ruchjana (2002), the model has been modified and improved based on findings from real data analysis, leading to improved performance and wider applicability. For instance, GSTAR has been successfully used for short-time forecasting of PM2.5 in Beijing-Tianjin-Hebei [8]. Yundari et al. further improved the model by incorporating kernel function method, which has proven effective in forecasting COVID-19 cases [9,10,11,12]. In the context of non-stationary data, Generalized Space-Time Autoregressive Integrated (GSTARI) was developed. The model is further enhanced by including MA component, resulting in GSTARMA. It is crucial to be informed that Diagicinto first introduced GSTARMA in 2006 for crime rate forecasting, which was later refined by Min et al. in 2010 [13,14].Subsequently, Akbar et al. applied this model to forecast air pollution [15], while Andayani et al. developed GSTARIMA and GSTARIMA-X models with exogenous elements through transfer functions [16,17]. Most recently, GSTARIMA has been extended to higher orders, specifically developing third-order model for rainfall forecast [18].

The application of GSTAR and GSTARIMA in real data analysis raises the assumption of heteroscedasticity in model errors, signifying that errors have a non-constant variance [19]. To address this issue, research has developed various extensions of GSTAR model. A significant approach was the combination of GSTAR with ARCH model, which is designed to handle non-constant variance in time series data [20,21]. Relating to this discussion, Nainggolan et al. developed GSTAR-ARCH, and Bonar et al. proposed GSTARI-ARCH [19,22]. Additionally, GSTARI-X-ARCH was introduced as an extension that includes exogenous variables useful for rainfall forecasting [23]. These models aim to improve the accuracy of ST forecasting by accounting for heteroscedasticity. In a related research, Monika et al. conducted literature review on the development of GSTARIMA that specifically addresses heteroscedasticity assumptions [24].

Based on previous research, explorers have identified an opportunity to develop new, more comprehensive models, and procedures. Specifically, the discovery has led to the development of a high-order GSTARIMA, called GSTARIMA (3,1,1), and it is combined with ARCH to overcome the non-constant error variance. This new model is called GSTARIMA (3,1,1)-ARCH and represents a significant advancement in the field. In the context of this research, the application of this model focuses on forecasting rainfall in Java Island, Indonesia. This process is achieved by using data obtained from NASA POWER big data website [25,26]. To conduct the analysis and modeling, data analytics life cycle methodology is used [27]. Furthermore, this methodology includes discovery, data preparation, model planning, model building, communicating results, and operationalization. The exploration provides benefits to science through model development and application to real data, especially in mathematical statistics, including stochastic modeling. Moreover, rainfall forecasting results can be used as valuable recommendations for various agencies in Indonesia. For example, the results can be applied when making policies such as disaster mitigation, weather forecasts by meteorological, climatological, and geophysical agencies (BMKG), as well as planning for planting season.

2 Materials and Methods

2.1 Space-Time Model with Box-Jenkins Procedure

A stochastic process was a series of random variables $Z(\omega,t)$, with ω representing the sample space and t was the time index. A set of random variables $\{Z_{t_1},Z_{t_2},Z_{t_3},\ldots,Z_{t_n}\}$ of a stochastic process $\{Z_t(\omega,t):t=0,\pm1,\pm2,\ldots\}$ were given [1]. Moreover, in this context, time series were considered as a stochastic process, $\{Z_t(\omega,t):t=0,\pm1,\pm2,\ldots\}$, where the random variables Z were indexed by time t in the sample space ω [1].

The main purpose of analyzing time series data was to produce predictions or forecasts of phenomena that occurred in the future by using the Box-Jenkins method. Furthermore, time series method with the Box-Jenkins consisted of three stages of procedures, namely model identification, parameter estimation, and diagnostic checking. In this research, model identification was conducted by checking data stationarity and plotting the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) [1,2]. Additionally, parameter estimation was performed by determining the correct value of the parameter coefficients to verify the estimation to be included in the model. Diagnostic checking included a white noise and normality examination to ensure that the selected model was suitable enough to be used in forecasting [4].

Research experts have developed ST model based on the principles and procedures of Box-Jenkins method [23]. Specifically, ST is a mathematical model for analyzing relationship between space and time in various phenomena and systems. In ST modeling, weight matrix, which is a square matrix with elements consisting of corresponding location weights was calculated. Additionally, weight matrix could be calculated based on the actual distance between locations. It is crucial to be aware that weight matrix is also known as inverse distance weight and can be calculated using the following formula.

$$w_{ij} = \frac{1}{d_{ij}}, \tag{1}$$

where w_{ij} represented the inverse distance weight matrix elements at location i and j while d_{ij} was the distance between locations i and j and as the two locations were closer, the weight given became greater. Furthermore, standardization was performed on the above matrix in the form of w_{ij} to obtain $\sum_{i\neq j} w_{ij}^{(l)} = 1$ that the inverse distance weight was,

$$\mathbf{W} = \begin{bmatrix} 0 & \frac{w_{12}}{w_{12}+w_{13}+w_{14}} & \frac{w_{13}}{w_{12}+w_{13}+w_{14}} & \frac{w_{14}}{w_{12}+w_{13}+w_{14}} \\ \frac{w_{21}}{w_{21}+w_{23}+w_{24}} & 0 & \frac{w_{23}}{w_{21}+w_{23}+w_{24}} & \frac{w_{24}}{w_{21}+w_{23}+w_{24}} \\ \frac{w_{31}}{w_{31}+w_{32}+w_{34}} & \frac{w_{32}}{w_{31}+w_{32}+w_{34}} & 0 & \frac{w_{34}}{w_{31}+w_{32}+w_{34}} \\ \frac{w_{41}}{w_{41}+w_{42}+w_{43}} & \frac{w_{42}}{w_{41}+w_{42}+w_{43}} & \frac{w_{43}}{w_{41}+w_{42}+w_{43}} & 0 \end{bmatrix} \tag{2}$$

GSTAR introduced by Ruchjana assumed that the characteristics of each location were heterogeneous. Additionally, the GSTAR(p,λ_k) model had a time order p and spatial order λ_k which was expressed in matrix form through Equation (3) [28],

$$\mathbf{z}(t) = \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\left[\boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{z}(t-k)\right] + \mathbf{e}(t), \tag{3}$$

where

- $\mathbf{z}(t)$: a vector of variables of size $(N\times 1)$ at time t,
- $\mathbf{z}(t-k)$: vector of variables of size $(N\times 1)$ at time $(t-k)$,
- λ_k : spatial order in the k-th autoregressive,
- v_k : spatial order of the k-th moving average,
- $\boldsymbol{\Phi}_{kl}$: autoregressive and space time parameters at time order k and spatial order l of size $(N\times N)$ in the form of diagonal matrix $\left(\Phi_{kl}^{(1)},\Phi_{kl}^{(2)},\Phi_{kl}^{(3)},\ldots,\Phi_{kl}^{(N)}\right)$,
- $\mathbf{W}^{(l)}$: weight matrix of size $(N\times N)$ at spatial order $l, l = 0,1,2,\ldots,\lambda_k$ containing $w_{ii}=0$ and $\sum_{i\neq j} w_{ij}=1$,
- $\mathbf{e}(t)$: error vector of size $(N\times 1)$ at time t, assuming $\mathbf{e}(t)^{iid}N\left(\mathbf{0},\sigma^2\mathbf{I}\right)$.

GSTAR that passed through the differencing process was transformed into GSTARI model. In addition, the general form of GSTARI (p,d,λ_k) with differencing order d in Equation (4) [23],

$$\mathbf{y}(t) = \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\left[\boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-k)\right] + \mathbf{e}(t), \tag{4}$$

where

$$\begin{aligned}\mathbf{y}(t) &= \mathbf{z}(t)-\mathbf{z}(t-1), \mathbf{y}(t-1)\\ &= \mathbf{z}(t-1)-\mathbf{z}(t-2),\ldots,\mathbf{y}(t-k)\\ &= \mathbf{z}(t-k)-\mathbf{z}(t-k-1),\end{aligned} \tag{5}$$

Digiacinto first introduced GSTARMA in discussing a new method to model regional unemployment in Italy. GSTARMA (p,d,λ_k) was an extension of GSTAR model with the addition of MA error elements. Moreover, model was applied to stationary data and expressed in Equation (6) [14].

$$\begin{aligned}\mathbf{z}(t) = &\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\left[\boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{z}(t-k)\right]\\ &-\sum_{k=1}^{q}\sum_{l=0}^{v_k}\left[\boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-k)\right]+\mathbf{e}(t),\end{aligned} \tag{6}$$

- $\boldsymbol{\Theta}_{kl}$: MA parameters at time order k and spatial order l of size $(N\times N)$ in the form of diagonal matrix $\left(\Theta_{kl}^{(1)},\Theta_{kl}^{(2)},\Theta_{kl}^{(3)},\ldots,\Theta_{kl}^{(N)}\right)$,
- $\mathbf{e}(t)$: error vector of size $(N\times 1)$ at time t, assuming $\mathbf{e}(t)^{iid}N\left(\mathbf{0},\sigma^2\mathbf{I}\right)$.

GSTARMA model developed on non-stationary data was called GSTARIMA, which was first introduced by Min. The developed model was used in urban traffic networks to forecast short-term traffic flow. GSTARIMA $(p_{\lambda_k},d,q_{v_k})$, d was the differencing order expressed in Equation (7) [13].

$$\begin{aligned}\mathbf{y}(t) = &\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\left[\boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-k)\right]-\sum_{k=1}^{q}\sum_{l=0}^{v_k}\left[\boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-k)\right]\\ &+\mathbf{e}(t),\end{aligned} \tag{7}$$

where

$$\begin{aligned}\mathbf{y}(t) &= \mathbf{z}(t)-\mathbf{z}(t-1), \mathbf{y}(t-1)\\ &= \mathbf{z}(t-1)-\mathbf{z}(t-2),\ldots,\mathbf{y}(t-k)\\ &= \mathbf{z}(t-k)-\mathbf{z}(t-k-1),\end{aligned} \tag{8}$$

2.1.1 Space-Time Model Identification

Pfeifer and Deutsch used Space-Time Autocorrelation Function (STACF) and Space-Time Partial Autocorrelation Function (STPACF) to identify STARMA [2,6]. In the calculation of STACF and STPACF, a spatial weight matrix was used to calculate the

autocovariance between two vectors of observations that had passed through a differencing process with a time lag m. Furthermore, the row of observation vectors $\mathbf{z}(1)$, $\mathbf{z}(2)$, $\mathbf{z}(3)$, ..., $\mathbf{z}(t)$ included observations at all locations $i = 1,2,3,...,N$ and time $t = 1,2,3,...,T$.

$$\Gamma(m) = \mathbb{E}[\mathbf{z}(t)(\mathbf{z}(t+m))'], \tag{9}$$

STACF function was obtained by standardizing the autocovariance function of the time lag m for observations with spatial lags k and l, namely $\gamma_{lk}(m)$. Generally, the variance of STACF had a fixed value for each spatial lag, which was explained using Equation (10).

$$\rho_{lk}(m) = \frac{\gamma_{lk}(m)}{\sqrt{\gamma_{ll}(0)\gamma_{kk}(0)}} \tag{10}$$

The order of GSTARIMA model was identified by observing the lags cut on the STPACF plot. In addition, the lags were defined as the last coefficient of ϕ_{lk} $=$ $(l = 0,1,2,...,\lambda$ and $k = 1,2,3,...)$ in the Yule-Walker Equation on the STPACF for spatial order λ.

2.1.2 Parameter Estimation of Space-Time Model

Parameter Estimation of GSTARIMA (p,d,λ_k) was conducted by estimating GSTARI and GSTIMA with Ordinary Least Square (OLS) and Maximum Likelihood Estimation (MLE) methods, respectively [29]. Additionally, parameter estimation of GSTARI (3,1,1) using OLS was obtained as follows.

$$\begin{aligned}\mathbf{y}^{(i)}(t) = {} & \boldsymbol{\Phi}_{10}^{(i)}\mathbf{y}^{(i)}(t-1) + \boldsymbol{\Phi}_{11}^{(i)}\mathbf{W}^{(l)}\mathbf{y}^{(i)}(t-1) \\ & + \boldsymbol{\Phi}_{21}^{(i)}\mathbf{W}^{(l)}\mathbf{y}^{(i)}(t-2) \\ & + \boldsymbol{\Phi}_{31}^{(i)}\mathbf{W}^{(l)}\mathbf{y}^{(i)}(t-3) + \mathbf{e}^{(i)}(t)\end{aligned} \tag{11}$$

GSTAR (3,1,1) for 4 locations monitored linear model equation as follows.

$$\mathbf{Y} = \begin{pmatrix} y^{(1)}(t) \\ y^{(2)}(t) \\ y^{(3)}(t) \\ y^{(4)}(t) \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \Phi_{10}^{(1)} \\ \Phi_{10}^{(2)} \\ \vdots \\ \Phi_{31}^{(4)} \end{pmatrix}; \mathbf{e} = \begin{pmatrix} e^{(1)}(t) \\ e^{(2)}(t) \\ e^{(3)}(t) \\ e^{(4)}(t) \end{pmatrix};$$
$$\mathbf{X} = \begin{pmatrix} y^{(1)}(t-1) & 0 & \cdots & 0 & 0 \\ 0 & y^{(2)}(t-1) & \cdots & 0 & 0 \\ 0 & 0 & \cdots & v^{(3)}(t-3) & 0 \\ 0 & 0 & \cdots & 0 & v^{(4)}(t-3) \end{pmatrix} \tag{12}$$

Estimated parameter values $\hat{\boldsymbol{\beta}}$ for GSTARI (1,1,1) and GSTARI (3,1,1) was calculated using OLS method in Equation (13).

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{13}$$

Estimated parameter values $\hat{\boldsymbol{\beta}}$ in GSTARI were used to calculate data forecasting. In addition, the errors from model were re-modeled with GSTIMA following Equation (14).

$$\mathbf{y}(t) = \mathbf{e}(t) - \sum_{k=1}^{q}\sum_{l=0}^{v_k} \boldsymbol{\theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-k) \tag{14}$$

In this research, the parameters of GSTIMA (1,1,1) were estimated. Assuming the error was white noise, the distribution of the error became,

$$\mathbf{e} = \begin{pmatrix} e^{(1)}(t) \\ e^{(2)}(t) \\ e^{(3)}(t) \\ e^{(4)}(t) \end{pmatrix}, \tag{15}$$

Multivariate normal with zero mean and constant variance $\sigma^2\mathbf{I}_{N\times T}$. Relating to this discussion, the probability function was obtained as follows.

$$f(\mathbf{e}|\boldsymbol{\Phi},\boldsymbol{\Theta},\sigma^2) = (2\pi)^{\frac{TN}{2}}|\sigma^2\mathbf{I}_{N\times T}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\mathbf{e}'\mathbf{e}\right) \tag{16}$$

with, $S(\boldsymbol{\Phi},\boldsymbol{\Theta}) = \mathbf{e}'\mathbf{e}$. Conditional Likelihood Function of $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$ and σ^2 was,

$$L(\mathbf{e}|\boldsymbol{\Phi},\boldsymbol{\Theta},\sigma^2) = (2\pi)^{-\frac{TN}{2}}(\sigma^2)^{-\frac{TN}{2}}\exp\left(-\frac{S(\boldsymbol{\Phi},\boldsymbol{\Theta})}{2\sigma^2}\right) \tag{17}$$

with $S(\boldsymbol{\Phi},\boldsymbol{\Theta})$ as the conditional sum of squares function.

$$S(\boldsymbol{\Phi},\boldsymbol{\Theta}) = \hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{18}$$

where, the vector $\hat{\mathbf{e}}$ represented the error vector of GSTARI model. Conditional Maximum Likelihood Estimation of σ^2, $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$ was,

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\Phi}},\hat{\boldsymbol{\Theta}})}{TN} \tag{19}$$

with $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Theta}}$ which minimizes $S(\boldsymbol{\Phi},\boldsymbol{\Theta})$.

The parameters of GSTARIMA were non-linear, in which case the parameters were obtained using an iteration producer through the Marquardt algorithm minimizing the sum of squares error.

$$\mathbf{e}(t) = \mathbf{y}(t) - \boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-1) + \boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-1) \tag{20}$$

$$\mathbf{e}(t-1) = \mathbf{y}(t-1) - \boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-2) + \boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-2) \tag{21}$$

$$\mathbf{e}(t-2) = \mathbf{y}(t-2) - \boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-3) + \boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-3) \tag{22}$$

$$\vdots$$

$$\mathbf{e}(t-n+1) = \mathbf{y}(t-n+1) - \boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-n) + \boldsymbol{\Theta}_{kl}W^{(l)}\mathbf{e}(t-n) \tag{23}$$

Equation (21) was substituted into Equation (20) to obtain the parameters Θ_{kl}^{l}, Equations (21) and Equation (22) were substituted into Equation (20) to obtain the parameter Θ_{kl}^{l}, when Equations (20) through Equation (23) were substituted into Equation (20), the parameters obtained $\hat{\theta}_{kl}^{(t-n)}$. Moreover, the iteration procedure in estimating model parameters was necessary to overcome the non-linearity. The non-linear nature of GSTARIMA constrained the determination of parameter confidence intervals. In addition, the sum of square was optimized by estimating the least square estimate as follows.

$$S(\boldsymbol{\Phi},\boldsymbol{\Theta}) = S(\boldsymbol{\delta}) \approx S(\hat{\boldsymbol{\delta}}) + (\boldsymbol{\delta}-\hat{\boldsymbol{\delta}})'\mathbf{Q}(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}) \tag{24}$$

with, $\boldsymbol{\delta}' = (\boldsymbol{\Phi}',\boldsymbol{\Theta}')$ and $\mathbf{Q} = \frac{1}{2}\left[\frac{\partial S(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_j}\right]$ for $i = 1,2,3,\ldots,K$ and $j = 1,2,3,\ldots,K$.

$$S(\boldsymbol{\delta}) = \sum_{t=1}^{T} \mathbf{e}(t)'\mathbf{e}(t) \tag{25}$$

$$\frac{\partial S(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}_i} = \sum_{t=1}^{T} 2\mathbf{e}(t)' \left.\frac{\partial \mathbf{e}(t)}{\partial \boldsymbol{\delta}_i}\right|_{\hat{\boldsymbol{\delta}}} = 0 \tag{26}$$

$$\frac{1}{2}\left.\frac{\partial^2 S(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_j}\right|_{\hat{\boldsymbol{\delta}}} = \sum_{t=1}^{T} \left.\frac{\partial^2 S(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_j}\right|_{\hat{\boldsymbol{\delta}}} + \sum_{t=1}^{T} \left.\frac{\partial \mathbf{e}(t)'}{\partial \boldsymbol{\delta}_i}\frac{\partial \mathbf{e}(t)}{\partial \boldsymbol{\delta}_j}\right|_{\hat{\boldsymbol{\delta}}} \left.\frac{\partial^2 \mathbf{e}(t)}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_j}\right|_{\hat{\boldsymbol{\delta}}} \tag{27}$$

$\left.\frac{\partial^2 \mathbf{e}(t)}{\partial \boldsymbol{\delta}_i \partial \boldsymbol{\delta}_j}\right|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}$ being a function of $\mathbf{e}(t)$ that occurred before time t, and under the condition that the model was fit, $E[\mathbf{e}(t)'\mathbf{e}(t-k)] = 0$ for $k \geq 1$ was ignored. The matrix $\mathbf{Q} = \mathbf{X}'\mathbf{X}$ with,

$$\mathbf{X} = \begin{bmatrix} \left.\frac{\partial \mathbf{e}(1)}{\partial \boldsymbol{\delta}_1}\right|_{\hat{\boldsymbol{\delta}}} & \left.\frac{\partial \mathbf{e}(1)}{\partial \boldsymbol{\delta}_2}\right|_{\hat{\boldsymbol{\delta}}} & \cdots & \left.\frac{\partial \mathbf{e}(1)}{\partial \boldsymbol{\delta}_k}\right|_{\hat{\boldsymbol{\delta}}} \\ \left.\frac{\partial \mathbf{e}(2)}{\partial \boldsymbol{\delta}_1}\right|_{\hat{\boldsymbol{\delta}}} & \left.\frac{\partial \mathbf{e}(2)}{\partial \boldsymbol{\delta}_2}\right|_{\hat{\boldsymbol{\delta}}} & \cdots & \left.\frac{\partial \mathbf{e}(2)}{\partial \boldsymbol{\delta}_k}\right|_{\hat{\boldsymbol{\delta}}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial \mathbf{e}(T)}{\partial \boldsymbol{\delta}_1}\right|_{\hat{\boldsymbol{\delta}}} & \left.\frac{\partial \mathbf{e}(T)}{\partial \boldsymbol{\delta}_2}\right|_{\hat{\boldsymbol{\delta}}} & \cdots & \left.\frac{\partial \mathbf{e}(T)}{\partial \boldsymbol{\delta}_k}\right|_{\hat{\boldsymbol{\delta}}} \end{bmatrix} \tag{28}$$

In this research, the sum of squares function was approximated by Equation (29):

$$S(\boldsymbol{\delta}) = S(\hat{\boldsymbol{\delta}}) + (\boldsymbol{\delta}-\hat{\boldsymbol{\delta}})'\mathbf{Q}(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}}) \tag{29}$$

The confidence interval estimate for $[\boldsymbol{\Phi},\boldsymbol{\Theta}]' = \boldsymbol{\delta}$ was obtained from $S(\boldsymbol{\delta}) = S(\hat{\boldsymbol{\delta}}) + \frac{K}{TN-K}S(\hat{\boldsymbol{\delta}}) \sim F_{K,TN-K,\alpha}$ with the quadratic equation representing $S(\boldsymbol{\delta})$ in Equation (29). The matrix $\mathbf{Q}$ was the numerical estimate used to prepare the confidence interval construction. Moreover, the exact sum of squares function $S(\boldsymbol{\delta})$ was replaced with the conditional sum of squares $S_*(\boldsymbol{\delta})$ when using conditional maximum likelihood. The matrix $\mathbf{Q}$ was used to calculate the moment matrix in the linearized estimation method, where the confidence interval on σ^2 was calculated in the case of a linear model with $(\sigma^2|Z(1),Z(2),\ldots,Z(T)) \sim S_*(\hat{\boldsymbol{\delta}})\chi^2_{TN-K}$.

2.1.3 Diagnostic Checking Model ST

1.Chi-Squared QQ plots
Chi-squared QQ plots were used to check the normality assumption of the model [30]. When the sample came from a normal distribution, the QQ plot showed a straight-line pattern. In addition to testing whether the errors of the model were multivariate normal, the Chi-squared QQ plot test was used, which included the values of d_j^2.

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1,2,\ldots,n \tag{30}$$

Here, d_j^2 represents the test statistic, $\mathbf{x}_j$ is the jth observation, $\bar{\mathbf{x}}$ represents the mean of the observations, and $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix. Moreover, when $d_j^2 \leq \chi_p^2(\alpha)$ or p-value $> \alpha$, it means that the model errors were multivariate normally distributed.

2.Lagrange Multiplier Test
The ARCH-Lagrange Multiplier (ARCH-LM) test examines the assumption of heteroscedasticity in model errors. To conduct the ARCH-LM test, a regression was performed on the square of the model errors using the following equation [31,32]:

$$\mathbf{e}_t^2 = \alpha_0 + \alpha_1 \mathbf{e}_{(t-1)}^2 + \cdots + \alpha_q \mathbf{e}_{(t-m)}^2 \tag{31}$$

where n is the number of observations, R^2 represents the coefficient of determination of the regression model on e_t^2, and m is the number of time lags. Additionally, the Lagrange Multiplier (LM) statistic is given by:

$$\text{LM} = n \times R^2 \tag{32}$$

When the value of LM is greater than the value of χ_m^2, this indicates an ARCH effect on the error or heteroscedasticity in the model.

2.2 Data Analytics Life Cycle

Data analytics life cycle was a methodology used to analyze big data allowing useful information to be obtained [33,34]. In addition, the data analytics life cycle consisted of six stages, including [27],

- •Discovery → problems were identified by the research, understood data sources, and formulated initial hypotheses.
- •Data Preparation → data were collected for exploration from predetermined sources, cleaning it of missing values and noise. The results of data cleaning were transformed from daily data to monthly or according to the criteria and requirements of explorers.

- •Model Planning → model used was planned by the research to identify data, determine models, methods, workflows, and evaluate criteria to test the accuracy of model.
- •Model Building → the exploration divided the data into training and examining data. At this stage, the data was modeled with ST following the Box-Jenkins procedure.
- •Communicate Results → the results obtained from modeling and data processing were interpreted by the research. Furthermore, the explorations then conducted trials and verifications to determine whether the research was successful.
- •Operationalize → the research provided a final report and recommendations for relevant agencies and applied the model to appropriate environments.

3 Results and Discussions

3.1 Development of GSTARIMA-ARCH Model

GSTARIMA errors that did not meet the assumption of homoskedasticity or had non-constant variance were estimated through the ARCH method. Additionally, parameter estimation of GSTARIMA-ARCH was divided into estimation by Maximum Likelihood Estimation (MLE) and Generalized Least Squares (GLS) methods.

The model equations are:

$$\mathbf{y}(t) = \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\left[\boldsymbol{\Phi}_{kl}\mathbf{W}^{(l)}\mathbf{y}(t-k)\right] - \sum_{k=1}^{q}\sum_{l=0}^{v_k}\left[\boldsymbol{\Theta}_{kl}\mathbf{W}^{(l)}\mathbf{e}(t-k)\right] + \mathbf{e}(t) \tag{33}$$

$$\mathbf{e}(t) = \mathbf{D}_t\boldsymbol{\eta}_t \tag{34}$$

$$\mathbf{e}_t \mid F_{t-1} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_t) \tag{35}$$

where $\mathbf{y}(t) = \mathbf{z}(t) - \mathbf{z}(t-1)$, $\mathbf{y}(t-1) = \mathbf{z}(t-1) - \mathbf{z}(t-2)$, …, $\mathbf{y} - \mathbf{k}(t) = \mathbf{z}(t-k) - \mathbf{z}(t-k-1)$.

A vector of observation data $z(0), z(1), z(2), \ldots, z(T)$, which consists of T observations at N locations, was used. The ARCH-Regression model for parameter estimation at location k with data $t = 1,2,3,\ldots,T$ is expressed as follows:

$$\mathbf{y}_k(t) = \mathbf{X}_k'(t)\boldsymbol{\beta}_k + \mathbf{e}_k(t) \tag{36}$$

$$\mathbf{e}_k(t) = \sqrt{h_k(t)}\,\eta_k(t) \tag{37}$$

$$h_k(t) = \alpha_{0k} + \alpha_{1k}e_k^2(t-1) \tag{38}$$

In simple terms, the ARCH-Regression model at location $k = 1$ is expressed as:

$$\mathbf{y}_t = \boldsymbol{\Phi}_0\mathbf{z}_{t-1} + \boldsymbol{\Phi}_1\mathbf{V}_{t-1} - \boldsymbol{\Theta}_0\mathbf{e}_{t-1} - \boldsymbol{\Theta}_1\mathbf{U}_{t-1} + \mathbf{e}_t = \mathbf{X}_t'\boldsymbol{\beta} + \mathbf{e}_t \tag{39}$$

where the conditional variance of the error $\mathbf{e}_t$ is given by:

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2 \tag{40}$$

For all unknown parameters, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$ with:

$$\boldsymbol{\beta} = \begin{bmatrix}\Phi_0\\ \Phi_1\\ \Theta_0\\ \Theta_1\end{bmatrix} \tag{41}$$

$$\boldsymbol{\alpha} = \begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix} \tag{42}$$

Additionally, the procedure for estimating the variance parameters $\theta = (\beta, \alpha)'$ is described as follows.

1. The procedure for estimating the variance parameter $\boldsymbol{\alpha} = \begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix}$ was as follows.
   - •Regressed $\mathbf{Y}$ against $\mathbf{X}$ with the OLS method, the estimated regression coefficients were obtained as $\mathbf{b} = \begin{pmatrix}b_1\\ b_2\end{pmatrix}$.
   - •Calculated the error $\mathbf{e}_t = \mathbf{y}_t - \mathbf{X}_t'\mathbf{b}$ with $t = 1,2,3,\ldots,T$. Initial estimated values for the parameters α were obtained from the autoregression coefficients $\mathbf{e}_t^2$ and $\mathbf{e}_{t-1}^2$, i.e., for example, $\boldsymbol{\alpha} = \begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix}$.
   - •Calculated the variance for each t with values a: $h_t = \alpha_0 + \alpha_1 e_{t-1}^2$. Then regress $\frac{e_t^2}{h_t} - 1$ against $\frac{1}{h_t}$ and $\frac{e_{t-1}^2}{h_t}$, thereby obtaining the regression coefficient $\mathbf{d}_a = \begin{pmatrix}\tilde{a}_0\\ \tilde{a}_1\end{pmatrix}$.
   - •The estimated parameters from steps 2 and 3 were obtained as follows:

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix}\hat{a}_0\\ \hat{a}_1\end{pmatrix}' = \boldsymbol{\alpha} + \mathbf{d}_\alpha = \begin{pmatrix}\alpha_0\\ \alpha_1\end{pmatrix}' + \begin{pmatrix}\tilde{\alpha}_0\\ \tilde{\alpha}_1\end{pmatrix}' = \begin{pmatrix}\alpha_0 + \tilde{\alpha}_0\\ \alpha_1 + \tilde{\alpha}_1\end{pmatrix}' \tag{43}$$

   Then $\mathbf{d}_\alpha$ for $t = 1,2,3,\ldots,T$ was calculated by: Suppose $\tilde{Z}_t = \left(\frac{1}{h_t}, \frac{e_t^2}{h_t}\right)$, $\tilde{Z}_t = (\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \ldots, \tilde{Z}_T)$, later $f_t = \frac{e_t^2}{h_t} - 1$, and $f' = (f_1, f_2, f_3, \ldots, f_T)$. Then $\mathbf{d}_\alpha = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'f$. Where $\tilde{Z}$ was the regressor matrix in the regression f_t against $\frac{1}{h_t}$ and $\frac{e_{t-1}^2}{h_t}$. The covariance matrix for $\hat{\boldsymbol{\alpha}}$ was $\mathrm{cov}(\hat{\boldsymbol{\alpha}}) = (\tilde{Z}'\tilde{Z})^{-1}$.

2.The procedure for estimating the regression parameters $\boldsymbol{\beta} = \begin{pmatrix} \Phi_0 \\ \Phi_1 \\ \Theta_0 \\ \Theta_1 \end{pmatrix}$ was as follows.

- •Calculated the variance parameter of location i with the steps in the procedure above, obtaining $\hat{\boldsymbol{\alpha}}_i = \begin{pmatrix} \hat{\alpha}_0^{(i)} \\ \hat{\alpha}_1^{(i)} \end{pmatrix}'$, for $i = 1,2,3,\ldots,N$.
- •For $t = 1,2,3,\ldots,T$, calculated the variance of each location i with,

$$\hat{h}_i(t) = \hat{\alpha}_0^{(i)} + \hat{\alpha}_1^{(i)} e_i^2(t). \tag{44}$$

- •Calculated the standardized error with:

$$\eta_i(t) = \frac{e_i^2(t)}{h_i(t)}, \tag{45}$$

for $i = 1,2,3,\ldots,N$ and $t = 1,2,3,\ldots,T$. Later, the correlation was calculated with,

$$\hat{\rho}_{ij} = \mathrm{Cor}(\eta_i, \eta_j). \tag{46}$$

- •Calculated the covariance matrix $\boldsymbol{\Sigma}$ using steps 2 and 3, and following the steps, the mean regression parameters were estimated using the GLS method.

Linear model equation with error assumption $(\mathbf{e}|X) \sim N(\mathbf{0}, \sigma^2\mathbf{U})$, where $\mathbf{U} = \begin{bmatrix} u_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & u_T \end{bmatrix}$, and the conditional variance of the error is expressed by

$$\mathbb{E}(\mathbf{e}'\mathbf{e} \mid \mathbf{X}) = \sigma^2\mathbf{U} = \sigma^2 \begin{bmatrix} u_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & u_T \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \sigma_T^2 \end{bmatrix}. \tag{47}$$

The linear model is heteroskedastic when $u_i \neq u_j$ for $i \neq j$, while the linear model is homoskedastic when $u_1 = u_2 = \cdots = u_T$. The GLS method was performed by a linear transformation of the model to obtain new data that met the assumptions of OLS method (homoskedastic). Moreover, suppose $\sigma_t^2 = h_{t-1} = \alpha_0 + \alpha_1 e_{t-1}^2$, the diagonal matrix $\mathbf{U}$ is symmetric and positive definite, showing that matrix $\mathbf{L}$ satisfies $\mathbf{U}^{-1} = \mathbf{L}'\mathbf{L}$. Following this calculation, transform the error using the matrix $\mathbf{L}$, so that $\tilde{\mathbf{e}} = \mathbf{L}\mathbf{e}$. The mean of $\tilde{\mathbf{e}}$ is $\mathbb{E}(\tilde{\mathbf{e}}) = \mathbb{E}(\mathbf{L}\mathbf{e}) = \mathbf{L}\mathbb{E}(\mathbf{e}) = 0$, and the conditional variance on X is

$$E(\tilde{\mathbf{e}}'\tilde{\mathbf{e}} \mid \mathbf{X}) = \mathbf{L}\sigma^2\mathbf{U}\mathbf{L}'. \tag{48}$$

From Equation (48), $\mathbf{U} = (\mathbf{L}'\mathbf{L})^{-1}$ is obtained. Therefore, the covariance matrix $\tilde{\mathbf{e}}$ conditional on X is written as

$$E(\tilde{\mathbf{e}}'\tilde{\mathbf{e}} \mid X) = \mathbf{L}\sigma^2\mathbf{U}\mathbf{L}' = \mathbf{L}\sigma^2(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}' = \sigma^2\mathbf{I}_T. \tag{49}$$

The regression model obtained by matrix transformation $\mathbf{L}$ is described as follows:

$$\mathbf{L}\mathbf{y} = \mathbf{L}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}\mathbf{e}. \tag{50}$$

For example, let $\tilde{\mathbf{y}} = \mathbf{L}\mathbf{y}$ and $\tilde{\mathbf{X}} = \mathbf{L}\mathbf{X}$, then we obtain:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{e}}. \tag{51}$$

The transformed model Equation (50) has mean $\mathbb{E}(\tilde{\mathbf{e}}) = 0$ and conditional variance $E(\tilde{\mathbf{e}}\tilde{\mathbf{e}}' \mid X) = \sigma^2\mathbf{I}_T$. Therefore, the transformed results fulfill the assumption of error $(\mathbf{e}|\mathbf{X}) \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$. The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{gls}}$ is calculated using the formula:

$$\hat{\boldsymbol{\beta}}_{\mathrm{gls}} = \left(\mathbf{X}'\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}. \tag{52}$$

3.2 Modeling Procedure

This research uses data analytics life cycle model (DALCM) for modeling and analysis, a method developed for big data analysis and development of Knowledge Discovery in Databases (KDD) Data Mining stage. Even though KDD data mining consisted of three stages, namely pre-processing, data mining, and post-processing, DALCM provided a more detailed approach with six distinct stages. These stages included discovery, data preparation, model planning, model building, communicating results, and operationalization. The primary difference between DALCM and KDD was in the analysis and modeling phase.

Figure 1 showed the stages of modeling GSTARIMA-ARCH analysis on rainfall data using DALCM. At the discovery stage, the research problem was identified, relevant literature was reviewed to serve as reference material, data sources for modeling were determined, and initial hypotheses were verified.

At the data preparation stage, data from sources that were determined at the previous stage was retrieved. The data for this research originated from NASA POWER website, which was an open source. In addition to the data preparation stage, the location was selected by inputting latitude and longitude coordinates, observation time, and data interval. The data collected was used as input for data cleaning from missing values and noisy data. Furthermore, daily data was transformed into monthly data based on the amount of rainfall each month and then filtered based on the season in Indonesia. In this research, data used in December, January, and February (DJF) was called the wet season which showed high intensity of rainfall.

GSTARIMA-ARCH modeling stage was conducted at planning and building stages while forecasting results were interpreted at result communication stage. Finally, operationalize stage contained output in the form of recommendations for related environments, scientific articles, and code programming that was used and

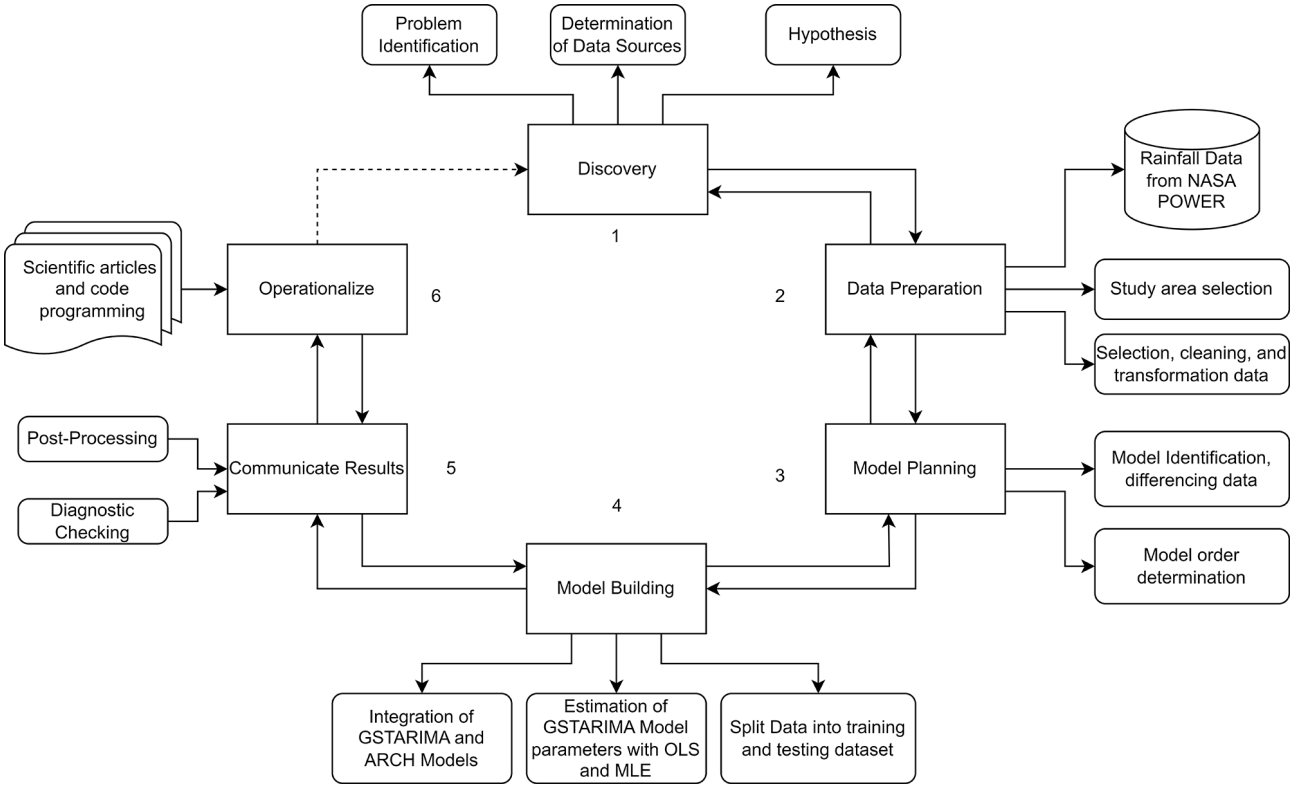


Fig. 1: Data Analytics Life Cycle Diagram for GSTARIMA-ARCH Modeling

developed in the future by explorers and practitioners in the fields of meteorology, stochastic, and spatiotemporal modeling mathematics.

The stages of GSTARIMA-ARCH modeling were described in detail as shown in Figure 2.

1. Data preparation results were used as output in GSTARIMA-ARCH modeling and were divided into training data and examining data.
2. Training data was used to train model and the first step performed based on Box-Jenkins method is model identification. Additionally, descriptive statistics and data correlation were calculated, while time series was plotted.
3. Data was checked for stationarity using ADF test for the mean and BoxCox Lambda for the variance. Non-stationary data was subjected to differencing and/or data transformation until the data became stationary.
4. Stationary data was used to determine the univariate model order with ARIMA.
5. The inverse distance weight matrix was calculated based on the actual distance of each location.
6. ST order identification with STACF and STPACF.
7. Estimation of GSTARI parameters with OLS and calculating model errors.
8. GSTARI errors were used to estimate GSTIMA parameters with MLE.
9. GSTARI and GSTIMA were combined into GSTARIMA.
10. Checking the homoscedasticity assumption of GSTARIMA error with ARCH-LM examination.
11. GSTARIMA errors that did not meet the assumption of homoskedasticity (heteroskedasticity) were re-modeled with ARCH. Parameter estimation of GSTARIMA-ARCH model using MLE and GLS.
12. GSTARIMA-ARCH errors were diagnosed to assess its suitability for modeling.

The GSTARIMA-ARCH was used for forecasting training as well as testing data, and the output was obtained as future rainfall forecasting results.

3.3 Research Area

Rainfall data was used as response variable, which was influenced by exogenous variables such as humidity and temperature. Moreover, the climate data was obtained from NASA POWER website and could be accessed at https://power.larc.nasa.gov/data-access-viewer/. The observation locations on Java Island consisted of 119 districts and cities, as shown in Figure 3. Different colors distinguished administrative boundaries between city districts. In addition, climate data was collected from December 1, 1982, to June 30, 2023, at daily intervals

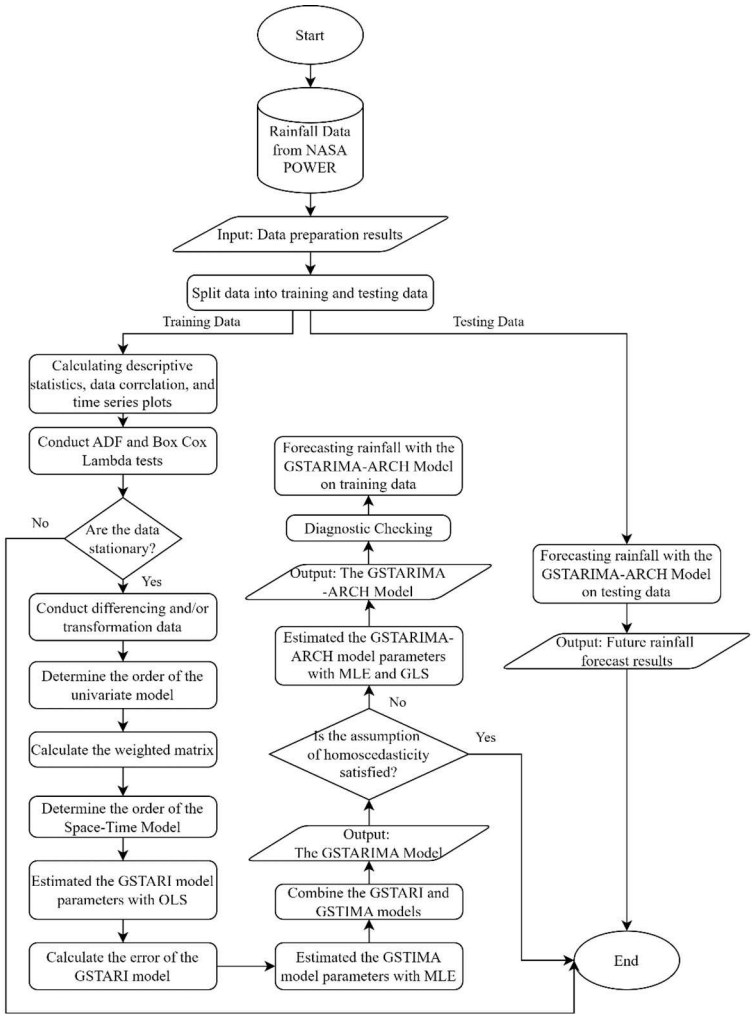


Fig. 2: Research Flow Chart for GSTARIMA-ARCH Model

which were stored in comma-separated values (.csv) format. The process of retrieving climate data on NASA POWER website required latitude and longitude coordinates.

3.4 GSTARIMA-ARCH model for Rainfall

GSTARIMA-ARCH modeling for rainfall data followed DALCM, and the first stage was a discovery, where explorations determined data sources and literature review. Data sources from NASA POWER that met big data criteria, were 3V (velocity, variety, volume). Furthermore, the data preparation stage was conducted by retrieving data on NASA POWER website. The selection process was performed on the selection of agroclimatology classes for rainfall variables with daily intervals from January 1, 1982, to June 30, 2023. Moreover, the results of rainfall data retrieval on NASA POWER website were stored in csv format. Daily rainfall data had a dimension of 15,159 series and 119 locations. Consequently, rainfall data passed through a pre-processing stage with the flow described in Figure 4.

Rainfall data was cleaned to check missing values and duplicate data, showing instances of missing values and several duplicate locations. Around 77 locations had the same observation data, therefore, the results of data cleaning were 42 locations. Rainfall data in 42 locations was aggregated into monthly data based on the amount of rain each month. Moreover, daily data totaling 15,159 series after being aggregated to monthly data became 502 series. In this research, monthly rainfall data was studied in months that had high rainfall intensity, including December, January, and February (DJF). Following this exploration, the results of DJF month selection obtained up to 123 series for 42 locations in Java Island.

Rainfall data in Java Island was modeled with univariate AutoRegressive Integrated Moving Average

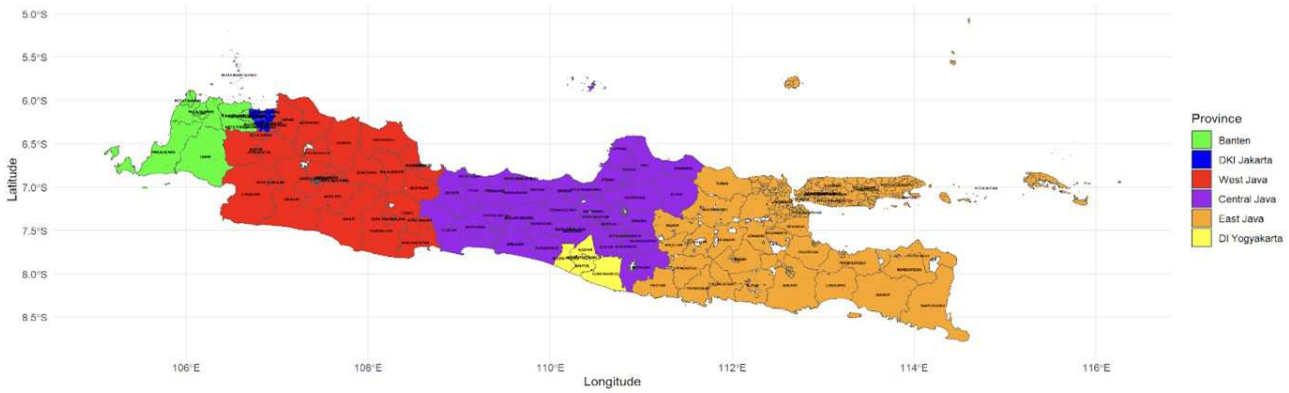


Fig. 3: Observation Location Map

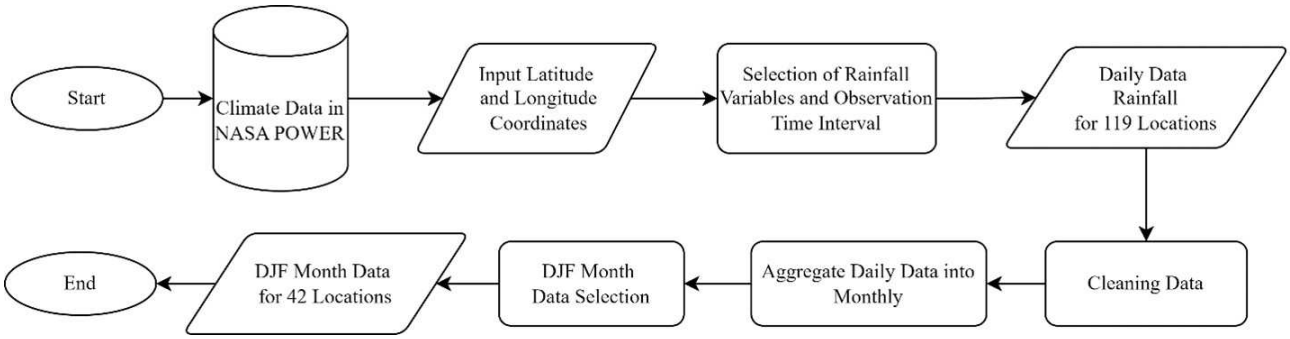


Fig. 4: Flow Chart for Pre-Processing Data

(ARIMA) time series. The best order obtained was ARIMA (3,1,1) model at 4 locations that showed the minimum Akaike Information Criterion (AIC) value. In addition, the 4 locations of rainfall data included Central Jakarta, Purwakarta, Serang City, and Lebak. GSTARIMA (3,1,1) was used to forecast rainfall at four locations on Java Island, with the correlation value of the observed data shown in Figure 5. Typically, rainfall data in the four observation locations had a very strong relationship, showing a tendency to follow the same up-and-down trend pattern.

Fig. 5: Correlation Values at 4 Locations in Java Island

Rainfall data at 4 locations were tested for stationarity and the results of the stationarity examination were presented in Table 1 and Table 2. The data at 4 observation locations were stationary on average with the acquisition of a value that was smaller than the significance level α, and ρ-value was smaller than the significance level $\alpha = 0.05$. Moreover, rainfall data still required to be stationary in variance with lambda values not close to one. Differencing and transformation processes were conducted on the data at 4 locations. However, rainfall data that was differentiated and transformed had a value smaller than the significance level. The ρ-value was lesser than α and lambda values were already close to one. Relating to this discussion, rainfall data was stationary in mean and variance.

Table 1: Stationarity Test Results Before Differencing

Location	Variables	Before Differencing		Con.
		p-value	Lambda	
Serang City	$Z^1(t)$	0.01	1.055366	TS
Lebak	$Z^2(t)$	0.01	0.323666	TS
Central Jakarta	$Z^3(t)$	0.01	-0.197534	TS
Purwakarta	$Z^4(t)$	0.01	-0.135984	TS

Table 2: Stationarity Test Results After Diferencing

Location	Variables	After Differencing		Con.
		p-value	Lambda	
Serang City	$Z^1(t)$	0.01	1.125785	S
Lebak	$Z^2(t)$	0.01	1.063062	S
Central Jakarta	$Z^3(t)$	0.01	1.056984	S
Purwakarta	$Z^4(t)$	0.01	0.970487	S

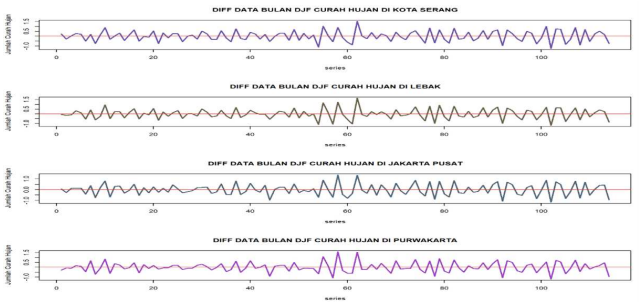


Fig. 6: Plot of Differencing and Transformation of Rainfall Data at 4 Locations

Figure 6 showed that time series plot for rainfall data after the differencing and transformation process was around the mean value represented by the red line. Moreover, the series plot of differencing data showed the rainfall data had the same up-and-down trend pattern. The inverse weight matrix of the distance between the 4 observation locations was calculated by converting the latitude and longitude coordinates. The calculation of weight matrix was calculated using Equation (1) and standardized with Equation (2). In addition, the results of the inverse weight matrix were shown in Equation (53).

$$\mathbf{W} = \begin{pmatrix} 0 & 0.4632 & 0.3593 & 0.1775 \\ 0.4769 & 0 & 0.3207 & 0.2023 \\ 0.3596 & 0.3117 & 0 & 0.3287 \\ 0.2527 & 0.2798 & 0.4676 & 0 \end{pmatrix} \tag{53}$$

The order of GSTARIMA (3,1,1) was identified using STACF as well as STPACF calculations, and plots were shown in Figure 7 and Figure 8. STACF plot in spatial lag 0 was truncated at time lags 1,2,3,5,6 and 9, and at spatial lag 1, the plot was shortened at 1, 2,3,5,6, and 9. Furthermore, STPACF plot in spatial lag 0 was truncated at time lags 1,2,3,5,6 and 8, and at spatial lag 1, it was shortened at time lags 2,5 and 8. When viewed from STACF and STPACF plots, models formed were GSTARIMA (1,1,1), (2,1,1), and (3,1,1). However, referring to the determination of the order using ARIMA, this observation location had ARIMA order (3,1,1). The specific model selected for this process modeling was GSTARIMA (3,1,1).

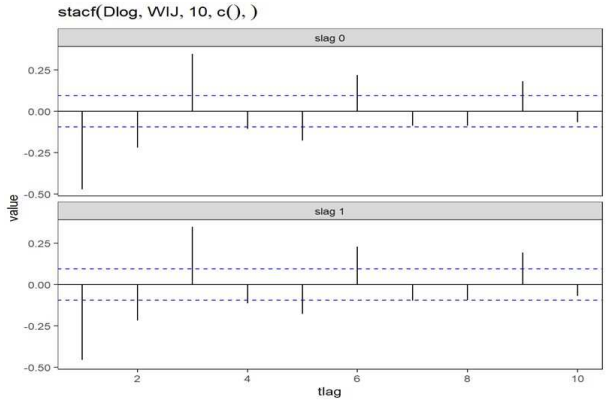


Fig. 7: STACF Plots for 4 Locations in Java Island

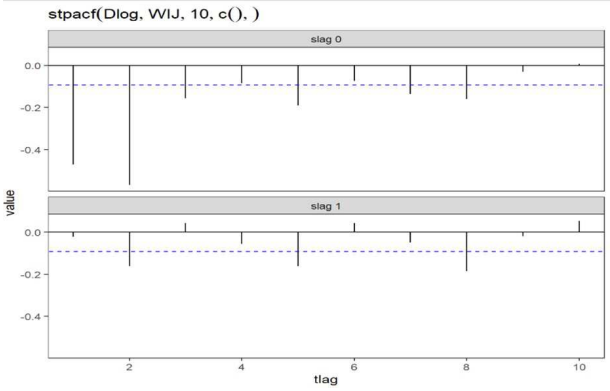


Fig. 8: STPACF Plots for 4 Locations in Java Island

GSTARIMA (3,1,1) was the same as GSTARI (3,1,1) and GSTIMA (1,1). Moreover, the parameter estimation procedure of GSTARI (3,1,1) was conducted by OLS method referring to Equation (9) up to (11). The results of the parameter estimation of GSTARI (3,1,1) model were shown in Table 3 and Table 4.

Table 3 and 4 showed the value of p-value of each parameter was still dominant, greater than the significant level, $\alpha = 0.05$. The p-value greater than α marked with (*) showed that the parameters were not partially significant. Moreover, parameters that had p-value lesser than α were partially significant. The estimated parameter

Table 3: Parameters Estimation of GSTARI (3,1,1) at spatial lag 0

ϕ	Estimated Value	Standard Error	t-Count	p-value
$\hat{\phi}_{10}^{(1)}$	-0.6056	0.2840	-2.1324	0.0336
$\hat{\phi}_{10}^{(2)}$	-1.7533	0.3885	-4.5136	0.0000
$\hat{\phi}_{10}^{(3)}$	-0.1960	0.3401	-0.5762	0.5648*
$\hat{\phi}_{10}^{(4)}$	-0.7493	0.2956	-2.5350	0.0116
$\hat{\phi}_{20}^{(1)}$	-0.0671	0.3211	-0.2088	0.8347*
$\hat{\phi}_{20}^{(2)}$	-1.6688	0.4547	-3.6704	0.0003
$\hat{\phi}_{20}^{(3)}$	-0.1261	0.3813	-0.3308	0.7410*
$\hat{\phi}_{20}^{(4)}$	-0.7050	0.3369	-2.0925	0.0370
$\hat{\phi}_{30}^{(1)}$	-0.0059	0.2825	-0.0210	0.9833*
$\hat{\phi}_{30}^{(2)}$	-0.7059	0.3926	-1.7979	0.0729*
$\hat{\phi}_{30}^{(3)}$	0.1369	0.3384	0.4045	0.6860*
$\hat{\phi}_{30}^{(4)}$	-0.3586	0.2914	-1.2306	0.2192*

Table 4: Parameters Estimation of GSTARI (3,1,1) at spatial lag 1

ϕ	Estimated Value	Standard Error	t-Count	p-value
$\hat{\phi}_{11}^{(1)}$	-0.2123	0.2771	-0.7660	0.4441*
$\hat{\phi}_{11}^{(2)}$	0.8841	0.3834	2.3057	0.0216
$\hat{\phi}_{11}^{(3)}$	-0.5899	0.3038	-1.9417	0.0528*
$\hat{\phi}_{11}^{(4)}$	-0.1638	0.4060	-0.4035	0.6868*
$\hat{\phi}_{21}^{(1)}$	-0.6076	0.3142	-1.9336	0.0538*
$\hat{\phi}_{21}^{(2)}$	0.9885	0.4410	2.2414	0.0255
$\hat{\phi}_{21}^{(3)}$	-0.4965	0.3451	-1.4388	0.1509*
$\hat{\phi}_{21}^{(4)}$	-0.0149	0.4634	-0.0321	0.9744*
$\hat{\phi}_{31}^{(1)}$	-0.1835	0.2769	-0.6628	0.5078*
$\hat{\phi}_{31}^{(2)}$	0.5807	0.3810	1.5239	0.1283*
$\hat{\phi}_{31}^{(3)}$	-0.2225	0.3064	-0.7261	0.4682*
$\hat{\phi}_{31}^{(4)}$	0.2648	0.4029	0.6574	0.5113*

results of the GSTARI (3,1,1) were formed into a matrix equation referred to Equation (4) as follows.

$$\begin{bmatrix}\hat{Y}^{(1)}(t)\\ \hat{Y}^{(2)}(t)\\ \hat{Y}^{(3)}(t)\\ \hat{Y}^{(4)}(t)\end{bmatrix} = \begin{bmatrix}-0.7745 & 0.4632 & 0.3593 & 0.1775\\ 0.4769 & -1.0808 & 0.3207 & 0.2023\\ 0.3596 & 0.3117 & -0.4860 & 0.3287\\ 0.2527 & 0.2798 & 0.4676 & -0.7094\end{bmatrix}\begin{bmatrix}y^{(1)}(t-1)\\ y^{(2)}(t-1)\\ y^{(3)}(t-1)\\ y^{(4)}(t-1)\end{bmatrix} + \begin{bmatrix}-0.4109 & -0.1620 & -0.1257 & -0.0621\\ 0.0630 & -0.8712 & 0.0424 & 0.0267\\ -0.0995 & -0.0862 & -0.4446 & -0.0909\\ -0.0545 & -0.0630 & -0.1008 & -0.5949\end{bmatrix}\begin{bmatrix}y^{(1)}(t-2)\\ y^{(2)}(t-2)\\ y^{(3)}(t-2)\\ y^{(4)}(t-2)\end{bmatrix} + \begin{bmatrix}-0.0825 & -0.0425 & -0.0329 & -0.0162\\ -0.0083 & -0.1283 & -0.0056 & -0.0035\\ 0.0073 & 0.0063 & -0.1500 & 0.0067\\ 0.0205 & 0.0227 & 0.0380 & -0.2419\end{bmatrix}\begin{bmatrix}y^{(1)}(t-3)\\ y^{(2)}(t-3)\\ y^{(3)}(t-3)\\ y^{(4)}(t-3)\end{bmatrix} \quad (54)$$

Equation (54) was written into GSTARI (3,1,1) equation for the location of Serang City as follows.

$$\begin{aligned}\hat{Y}^{(1)}(t) = &-0.7745y^{(1)}(t-1)+0.4632y^{(2)}(t-1)\\ &+0.3593y^{(3)}(t-1)+0.1775y^{(4)}(t-1)\\ &-0.4109y^{(1)}(t-2)-0.1620y^{(2)}(t-2)\\ &-0.1257y^{(3)}(t-2)-0.0621y^{(4)}(t-2)\\ &-0.0825y^{(1)}(t-3)-0.0425y^{(2)}(t-3)\\ &-0.0329y^{(3)}(t-3)-0.0162y^{(4)}(t-3)\end{aligned} \quad (55)$$

GSTARIMA (3,1,1) for the other locations was constructed similarly to Equation (42). The model was used to perform forecasting on in-sample and out-sample. Additionally, forecasting results produced errors that were then used for input to GSTIMA (1,1,1) and were estimated using the MLE method. The estimated parameter results of the model were shown in Table 5.

Table 5: Parameters Estimation of GSTIMA(1,1,1)

Θ	Estimated Value	Standard Error	t-Count	p-value
$\hat{\Theta}_{10}^{(1)}$	-0.0029	0.0004	-7.5753	0
$\hat{\Theta}_{10}^{(2)}$	-0.0025	0.0004	-6.2398	0
$\hat{\Theta}_{10}^{(3)}$	-0.0028	0.0003	-8.4214	0
$\hat{\Theta}_{10}^{(4)}$	-0.0041	0.0004	-10.8018	0
$\hat{\Theta}_{11}^{(1)}$	-0.0006	0.0003	-1.6361	0.1032*
$\hat{\Theta}_{11}^{(2)}$	-0.0011	0.0004	-3.0603	0.0025
$\hat{\Theta}_{11}^{(3)}$	-0.0003	0.0003	-1.0024	0.3172*
$\hat{\Theta}_{11}^{(4)}$	0.0009	0.0005	1.8630	0.0638*

The results of the calculation of the estimated value of GSTIMA (1,1,1) parameters showed that the p-value of model parameters was dominantly and partially significant. Moreover, the parameters that had a p-value less than $\alpha = 0.05$. The estimated value of GSTIMA (1,1,1) parameters was presented in the form of a matrix in Equation (56):

$$\begin{bmatrix}\hat{Y}^{(1)}(t)\\ \hat{Y}^{(2)}(t)\\ \hat{Y}^{(3)}(t)\\ \hat{Y}^{(4)}(t)\end{bmatrix} = \begin{bmatrix}-0.0029 & -0.0002 & -0.0002 & -0.0001\\ -0.0005 & -0.0025 & -0.0003 & -0.0002\\ -0.0001 & -0.0001 & -0.0028 & -0.0001\\ 0.0002 & 0.0002 & 0.0004 & -0.0041\end{bmatrix}\begin{bmatrix}e^{(1)}(t-1)\\ e^{(2)}(t-1)\\ e^{(3)}(t-1)\\ e^{(4)}(t-1)\end{bmatrix} \quad (56)$$

GSTIMA (1,1,1) for the location of Serang City referred to the calculation of Equation (56) which was written into Equation (57):

$$\begin{aligned}\hat{Y}^{(1)}(t) = &-0.00295e^{(1)}(t-1)-0.0002e^{(2)}(t-1)\\ &-0.0002e^{(3)}(t-1)-0.0001e^{(4)}(t-1)\end{aligned} \quad (57)$$

GSTIMA (1,1,1) model equation for other locations was the same as Equation (57). GSTARI (3,1,1) and GSTIMA (1,1,1) were combined into GSTARIMA (3,1,1). Relating to this, the GSTARIMA (3,1,1) equation was shown in Equation (58).

$$\begin{bmatrix}\hat{Y}^{(1)}(t)\\ \hat{Y}^{(2)}(t)\\ \hat{Y}^{(3)}(t)\\ \hat{Y}^{(4)}(t)\end{bmatrix} = \begin{bmatrix}-0.7745 & 0.4632 & 0.3593 & 0.1775\\ 0.4769 & -1.0808 & 0.3207 & 0.2023\\ 0.3596 & 0.3117 & -0.4860 & 0.3287\\ 0.2527 & 0.2798 & 0.4676 & -0.7094\end{bmatrix}\begin{bmatrix}y^{(1)}(t-1)\\ y^{(2)}(t-1)\\ y^{(3)}(t-1)\\ y^{(4)}(t-1)\end{bmatrix} + \begin{bmatrix}-0.4109 & -0.1620 & -0.1257 & -0.0621\\ 0.0630 & -0.8712 & 0.0424 & 0.0267\\ -0.0995 & -0.0862 & -0.4446 & -0.0909\\ -0.0545 & -0.0603 & -0.1008 & -0.5949\end{bmatrix}\begin{bmatrix}y^{(1)}(t-2)\\ y^{(2)}(t-2)\\ y^{(3)}(t-2)\\ y^{(4)}(t-2)\end{bmatrix} + \begin{bmatrix}-0.0825 & -0.0425 & -0.0329 & -0.0162\\ -0.0083 & -0.1283 & -0.0056 & -0.0035\\ 0.0073 & 0.0063 & -0.1500 & 0.0067\\ 0.0205 & 0.0227 & 0.0380 & -0.2419\end{bmatrix}\begin{bmatrix}y^{(1)}(t-3)\\ y^{(2)}(t-3)\\ y^{(3)}(t-3)\\ y^{(4)}(t-3)\end{bmatrix} + \begin{bmatrix}-0.0029 & -0.0002 & -0.0002 & -0.0001\\ -0.0005 & -0.0025 & -0.0003 & -0.0002\\ -0.0001 & -0.0001 & -0.0028 & -0.0001\\ 0.0002 & 0.0002 & 0.0004 & -0.0041\end{bmatrix}\begin{bmatrix}e^{(1)}(t-1)\\ e^{(2)}(t-1)\\ e^{(3)}(t-1)\\ e^{(4)}(t-1)\end{bmatrix} \quad (58)$$

GSTARIMA (3,1,1) model for the location of Serang City was shown in Equation (59),

$$\begin{aligned}\hat{Y}^{(1)}(t) = &-0.7745y^{(1)}(t-1)+0.4632y^{(2)}(t-1)\\ &+0.3593y^{(3)}(t-1)+0.1775y^{(4)}(t-1)\\ &-0.4109y^{(1)}(t-2)-0.1620y^{(2)}(t-2)\\ &-0.1257y^{(3)}(t-2)-0.0621y^{(4)}(t-2)\\ &-0.0825y^{(1)}(t-3)-0.0425y^{(2)}(t-3)\\ &-0.0329y^{(3)}(t-3)-0.0162y^{(4)}(t-3)\\ &-0.00295e^{(1)}(t-1)-0.0002e^{(2)}(t-1)\\ &-0.0002e^{(3)}(t-1)-0.0001e^{(4)}(t-1)\end{aligned} \tag{59}$$

The examination results of ARCH error effect showed the value of p-value = 0.0003137, which was smaller than the value of $\alpha = 0.05$. Rejecting H_0 showed that GSTARIMA (3,1,1) contained ARCH effects on model errors. Therefore, to overcome the non-constant variance of the errors, the parameters were drawn by the ARCH method.

According to error testers in GSTARIMA (3,1,1) containing ARCH effect, GSTARIMA (3,1,1)-ARCH(1) was used to forecast rainfall on Java Island. In the context of this research, the estimation of the model was conducted using MLE and GLS methods. The first estimation was performed at each location as follows:

$$\hat{\sigma}_1^2(t) = 19460 + 0.0894\hat{e}_1^2(t-1) \tag{60}$$

$$\hat{\sigma}_2^2(t) = 17750 + 0.02732\hat{e}_2^2(t-1) \tag{61}$$

$$\hat{\sigma}_3^2(t) = 24490 + 0.00000001\hat{e}_3^2(t-1) \tag{62}$$

$$\hat{\sigma}_4^2(t) = 17750 + 0.09026\hat{e}_4^2(t-1) \tag{63}$$

Once the conditional variance estimation equation for each location was obtained, the conditional variance for each time was calculated, with $t = 2,3,4,\ldots,114$. Moreover, the conditional variance was calculated using the following equation.

$$\hat{\sigma}_i^2(t) = \frac{\alpha_{0i}}{1-\alpha_{1i}}, \quad i = 1,2,3,4 \tag{64}$$

The results of the conditional variance calculation were inputted in a diagonal matrix for each location as follows:

$$H_1 = \text{diag}(\hat{\sigma}_1^2(2), \hat{\sigma}_1^2(3), \ldots, \hat{\sigma}_1^2(114)) \tag{65}$$

$$H_2 = \text{diag}(\hat{\sigma}_2^2(2), \hat{\sigma}_2^2(3), \ldots, \hat{\sigma}_2^2(114)) \tag{66}$$

$$H_3 = \text{diag}(\hat{\sigma}_3^2(2), \hat{\sigma}_3^2(3), \ldots, \hat{\sigma}_3^2(114)) \tag{67}$$

$$H_4 = \text{diag}(\hat{\sigma}_4^2(2), \hat{\sigma}_4^2(3), \ldots, \hat{\sigma}_4^2(114)) \tag{68}$$

Parameter estimation of GSTARIMA (3,1,1)-ARCH(1) with the GLS method was performed using the diagonal matrix of conditional variance for each location. Subsequently, the estimation of the model was calculated using Equations (33) to (38) and the results were shown in Table 6 and Table 7.

Table 6: Parameters Estimation of GSTARIMA (3,1,1) - ARCH (1) at spatial lag 0

Param	Estimated Value	Standard Error	t-Count	p-value
$\hat{\phi}_{10}^{(1)}$	-0.7108	0.2715	-2.8532	0.0045
$\hat{\phi}_{10}^{(2)}$	-1.0850	0.4362	-2.4779	0.0136
$\hat{\phi}_{10}^{(3)}$	-0.4860	0.3358	-1.4475	0.1485*
$\hat{\phi}_{10}^{(4)}$	-0.7599	0.3034	-2.3381	0.0198
$\hat{\phi}_{20}^{(1)}$	-0.3612	0.3003	-1.3683	0.1720*
$\hat{\phi}_{20}^{(2)}$	-0.8639	0.4958	-1.7574	0.0796*
$\hat{\phi}_{20}^{(3)}$	-0.4446	0.3779	-1.1766	0.2400*
$\hat{\phi}_{20}^{(4)}$	-0.6295	0.3485	-1.7074	0.0885*
$\hat{\phi}_{30}^{(1)}$	-0.0986	0.2720	-0.3036	0.7616*
$\hat{\phi}_{30}^{(2)}$	-0.1296	0.4346	-0.2954	0.7679*
$\hat{\phi}_{30}^{(3)}$	-0.1501	0.3357	-0.4471	0.6550*
$\hat{\phi}_{30}^{(4)}$	-0.1895	0.3046	-0.7944	0.4274*
$\hat{\Theta}_{10}^{(1)}$	-0.0032	0.0004	-8.6250	0.0000
$\hat{\Theta}_{10}^{(2)}$	-0.0025	0.0004	-6.3893	0.0000
$\hat{\Theta}_{10}^{(3)}$	-0.0032	0.0003	-9.5471	0.0000
$\hat{\Theta}_{10}^{(4)}$	-0.0033	0.0004	-8.6409	0.0000

Table 7: Parameters Estimation of GSTARIMA (3,1,1) - ARCH (1) at spatial lag 1

Param	Estimated Value	Standard Error	t-Count	p-value
$\hat{\phi}_{11}^{(1)}$	-0.1600	0.2533	-0.4135	0.6795*
$\hat{\phi}_{11}^{(2)}$	0.1958	0.3924	0.4878	0.6259*
$\hat{\phi}_{11}^{(3)}$	-0.3841	0.3450	-1.1132	0.2662*
$\hat{\phi}_{11}^{(4)}$	-0.1838	0.4147	-0.5621	0.5744*
$\hat{\phi}_{21}^{(1)}$	-0.3916	0.2828	-1.2372	0.2167*
$\hat{\phi}_{21}^{(2)}$	0.1276	0.4391	0.3012	0.7634*
$\hat{\phi}_{21}^{(3)}$	-0.2768	0.3883	-0.7130	0.4762*
$\hat{\phi}_{21}^{(4)}$	-0.1780	0.4762	-0.4531	0.6507*
$\hat{\phi}_{31}^{(1)}$	-0.0803	0.2580	-0.3559	0.7221*
$\hat{\phi}_{31}^{(2)}$	-0.0151	0.3829	-0.0458	0.9635*
$\hat{\phi}_{31}^{(3)}$	0.0204	0.3463	0.0588	0.9531*
$\hat{\phi}_{31}^{(4)}$	0.0276	0.4145	0.1962	0.8446*
$\hat{\Theta}_{11}^{(1)}$	-0.0002	0.0003	-0.6797	0.4974*
$\hat{\Theta}_{11}^{(2)}$	-0.0010	0.0004	-2.9527	0.0035
$\hat{\Theta}_{11}^{(3)}$	0.0001	0.0003	0.2230	0.8237*
$\hat{\Theta}_{11}^{(4)}$	-0.0001	0.0005	-0.2892	0.7727*

The estimated parameters of GSTARIMA (3,1,1)-ARCH (1) were presented in matrix form in Equation (69),

$$\begin{aligned}\begin{bmatrix}\hat{Y}^{(1)}(t)\\ \hat{Y}^{(2)}(t)\\ \hat{Y}^{(3)}(t)\\ \hat{Y}^{(4)}(t)\end{bmatrix} &= \begin{bmatrix}-0.7108 & -0.0741 & -0.0575 & -0.0284\\ 0.0934 & -1.0850 & 0.0628 & 0.0396\\ -0.1381 & -0.1197 & -0.4860 & -0.1263\\ -0.0465 & -0.0514 & -0.0860 & -0.7599\end{bmatrix}\begin{bmatrix}y^{(1)}(t-1)\\ y^{(2)}(t-1)\\ y^{(3)}(t-1)\\ y^{(4)}(t-1)\end{bmatrix}\\ &+\begin{bmatrix}-0.3612 & -0.1814 & -0.1407 & -0.0695\\ 0.0609 & -0.8639 & 0.0409 & 0.0258\\ -0.0996 & -0.0863 & -0.4446 & -0.0910\\ -0.0450 & -0.0498 & -0.0832 & -0.6295\end{bmatrix}\begin{bmatrix}y^{(1)}(t-2)\\ y^{(2)}(t-2)\\ y^{(3)}(t-2)\\ y^{(4)}(t-2)\end{bmatrix}\\ &+\begin{bmatrix}-0.0986 & -0.0372 & -0.0289 & -0.0143\\ -0.0072 & -0.1296 & -0.0048 & -0.0030\\ 0.0073 & 0.0064 & -0.1501 & 0.0067\\ 0.0070 & 0.0077 & 0.0129 & -0.1895\end{bmatrix}\begin{bmatrix}y^{(1)}(t-3)\\ y^{(2)}(t-3)\\ y^{(3)}(t-3)\\ y^{(4)}(t-3)\end{bmatrix}\\ &-\begin{bmatrix}-0.0033 & -0.0001 & -0.0001 & 0.0000\\ -0.0005 & -0.0026 & -0.0003 & -0.0002\\ 0.0000 & 0.0000 & -0.0032 & 0.0000\\ -0.0000 & -0.0000 & -0.0001 & -0.0033\end{bmatrix}\begin{bmatrix}e^{(1)}(t-1)\\ e^{(2)}(t-1)\\ e^{(3)}(t-1)\\ e^{(4)}(t-1)\end{bmatrix}\end{aligned} \tag{69}$$

GSTARIMA (3,1,1) - ARCH (1) for the location of Serang City was presented in Equation (70),

$$\begin{aligned}\hat{Y}^{(1)}(t) = &-0.71082y^{(1)}(t-1)-0.07409y^{(2)}(t-1)\\ &-0.05747y^{(3)}(t-1)-0.02839y^{(4)}(t-1)\\ &-0.36116y^{(1)}(t-2)-0.18139y^{(2)}(t-2)\\ &-0.14070y^{(3)}(t-2)-0.06951y^{(4)}(t-2)\\ &-0.09862y^{(1)}(t-3)-0.03720y^{(2)}(t-3)\\ &-0.02890y^{(3)}(t-3)-0.01430y^{(4)}(t-3)\\ &+0.00325e^{(1)}(t-1)+0.00011e^{(2)}(t-1)\\ &+0.00008e^{(3)}(t-1)-0.00004e^{(4)}(t-1)\end{aligned} \tag{70}$$

GSTARIMA (3,1,1) - ARCH (1) for other locations was the same as Equation (70). In addition, model was used to forecast rainfall on in-sample and out-sample data.

4 Conclusions

In conclusion, GSTARIMA (3,1,1)-ARCH (1) in this research was developed referring to the stages of Box-Jenkins procedure. During the identification stage, the data stationarity check was conducted, which was a requirement for GSTARIMA modeling, determining the order of time series and ST. Furthermore, GSTARIMA (3,1,1)-ARCH passed through several stages of parameter estimation procedures. First, OLS method estimated GSTARI (3,1,1), followed by model errors which were re-modeled using MA elements through GSTIMA (1,1,1), which was estimated using MLE. Furthermore, GSTARI (3,1,1) and GSTIMA (1,1,1) were combined into GSTARIMA (3,1,1), which passed through a diagnosed stage to show the assumptions of errors. Consequently, errors with no constant variance were modeled through ARCH procedure, while GSTARIMA (3,1,1)-ARCH (1) was estimated using MLE and GLS methods. This modeling procedure was more comprehensive and structured to produce a good model.

GSTARIMA (3,1,1)-ARCH (1) was used to model rainfall in Java Island following the stages of DALCM. Data analytics life cycle performed an important role in analyzing this research's large amount of rainfall data. The stage commenced with discovery which included, identification of the research problem, determination of the data source used, and selection of the observation location. Furthermore, data was collected from NASA POWER website at the data preparation stage. NASA POWER website contained climate data around the world, including a large number of variables and a long period. Moreover, location selection was conducted by inputting the desired location's latitude as well as longitude coordinate values, and the selection of the period from 1982 to 2023 with daily intervals. The collected rainfall data passed through data cleaning to remove noise and fill in missing values. Additionally, the data was aggregated into monthly data and filtered for wet months or DJF. Data preparation results were inputted into model planning, including model identification according to Box-Jenkins procedure. Given this scenario, building was performed by modeling GSTARIMA (3,1,1) - ARCH (1) by dividing the data into training and testing data. Rainfall forecasting results were interpreted at the communication results stage, and recommendations from research were used at the operationalized stage.

The research contributed to the development of GSTARIMA (3,1,1) - ARCH (1), which was used for rainfall forecasting and also used in general on ST data. Suggestions for further exploration included examining exogenous variables that affected rainfall modeling. Furthermore, the research could combine GSTARIMA (3,1,1) - ARCH (1) with exogenous variables that influenced the variables. This research hopes to be an inspiration for other explorations in the field of statistics and mathematics to develop ST model.

Acknowledgment

The authors are grateful to the Rector, Directorate of Research and Community Service (DRPM), Center for Modeling and Computation Studies, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. This research was funded by the Padjadjaran Excellence Fast Track Scholarship (BUPP) grant number 1425/UN6.3.1/PT.00/2024 and The Academic Leadership Grant (ALG) grant number 1817/UN6.3.1/PT.00/2024. Additionally, the authors are also grateful to Prof. Dr. Eddy Hermawan, M.Sc., Prof. Dr. Sukono, M.M., M.Si., and Prof. Dr. Diah Chaerani, M.Si. for valuable discussions. This research also supported by RISE_SMA Project funded by European Union year 2019-2024.

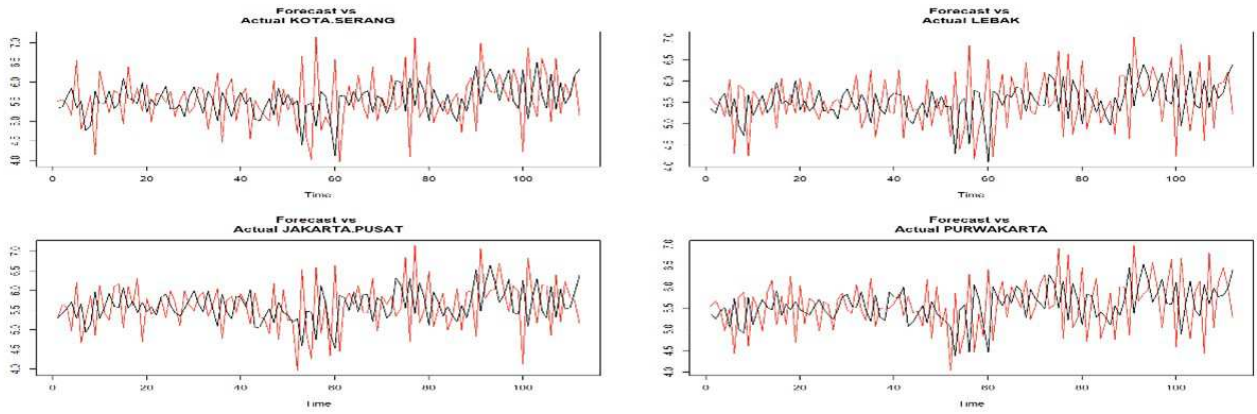


Fig. 9: Plot of Actual and Forecast of GSTARIMA (3,1,1) for 4 Locations on In-Sample Data

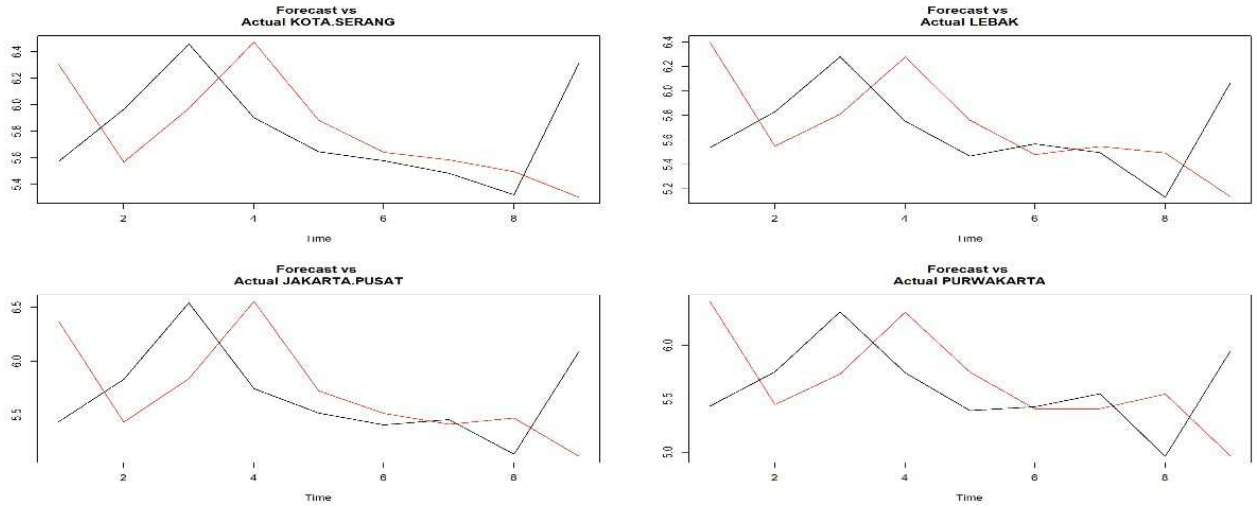


Fig. 10: Plot of Actual and Forecast of GSTARIMA (3,1,1) for 4 Locations on Out-Sample Data

References

[1] Wei, W. W. S. Time Series Analysis: Univariate and Multivariate Methods (Issue 1, pp. 1–614) (2006).

[2] Wei, W. W. S. Multivariate time series analysis and applications. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119502951 (2019).

[3] Lamigueiro, O. P. Displaying time series, spatial, and space-time data with R. Chapman and Hall/CRC (2018).

[4] Box, G. E. P., & Jenkins, G. M. Time series analysis forecasting and control. Holden-Day (1976).

[5] Pfeifer, P.E., Deutsch, S.J., A Three-Stage Iterative Procedure for Space-Time Modeling, Technometrics **22**, 35-47 (1980).

[6] Pfeifer, P.E., Deutsch, S.J., A STARIMA Model-Building Procedure with Application to Description and Regional Forecasting, Transactions of the Institute of British Geographers **5**, 330-349 (1980). doi:10.2307/621846

[7] Borovkova, S.A., Lopuhaä, H.P., Nurani, B., Generalized STAR Model with Experimental Weights, In Proceedings of the Proceedings of the 17th International Workshop on Statistical Modelling, 139–147 (2002).

[8] J. Hu, S. Wang, and J. Mao, "Short time PM2.5 prediction model for Beijing-Tianjin-Hebei region based on Generalized Space Time Autoregressive (GSTAR)," in IOP Conference Series: Earth and Environmental Science, Institute of Physics Publishing, Dec. (2019). doi: 10.1088/1755-1315/358/2/022075.

[9] Y. Yundari, N. M. Huda, U. S. Pasaribu, U. Mukhaiyar, and K. N. Sari, "Stationary Process in GSTAR(1;1) through Kernel Function Approach," in AIP Conference Proceedings, American Institute of Physics Inc., Sep. (2020). doi: 10.1063/5.0016808.

[10] M. Alawiyah, D. A. Kusuma, and B. N. Ruchjana, "Application of generalized space time autoregressive integrated (GSTARI) model in the phenomenon of covid-19," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jan. (2021). doi: 10.1088/1742-6596/1722/1/012035.

[11] U. S. Pasaribu, U. Mukhaiyar, N. M. Huda, K. N. Sari, and S. W. Indratno, "Modelling COVID-19 growth cases of provinces in java Island by modified spatial weight matrix GSTAR through railroad passenger's mobility," Heliyon, vol. 7, no. 2, Feb. (2021), doi: 10.1016/j.heliyon.2021.e06025.

[12] N. M. Huda and N. Imro'ah, "Determination of the best weight matrix for the Generalized Space Time Autoregressive (GSTAR) model in the Covid-19 case on Java Island, Indonesia," Spat Stat, vol. 54, Apr. (2023) doi: 10.1016/j.spasta.2023.100734.

[13] X. Min, J. Hu, and Z. Zhang, "Urban traffic network modeling and short-term traffic flow forecasting based on GSTARIMA model," IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 1535–1540, (2010), doi: 10.1109/ITSC.2010.5625123.

[14] Di Giacinto, V. A generalized space-time ARMA model with an application to regional unemployment analysis in Italy. International Regional Science Review, 29 **2**, 159–198 (2006). https://doi.org/10.1177/0160017605279457

[15] Akbar, M. S., Setiawan, Suhartono, Ruchjana, B. N., Prastyo, D. D., Muhaimin, A., & Setyowati, E. A Generalized Space-Time Autoregressive Moving Average (GSTARMA) Model for Forecasting Air Pollutant in Surabaya. Journal of Physics: Conference Series, 1490 **1** (2020). https://doi.org/10.1088/1742-6596/1490/1/012022

[16] Aulia, N., & Saputro, D. R. S. Generalized Space Time Autoregressive Integrated Moving Average with Exogenous (GSTARIMA-X) Models. IOP Conference Series: Earth and Environmental Science, 1808 **1** (2021). https://doi.org/10.1088/1742-6596/1808/1/012052

[17] Andayani, N., Sumertajaya, I. M., Ruchjana, B. N., & Aidi, M. N. Comparison of GSTARIMA and GSTARIMA-X Model by using Transfer Function Model Approach to Rice Price Data. IOP Conference Series: Earth and Environmental Science, 187 **1** (2018). https://doi.org/10.1088/1755-1315/187/1/012052

[18] Salsabila, A. B., Ruchjana, B. N., & Abdullah, A. S. Development of the GSTARIMA(1,1,1) model order for climate data forecasting. International Journal of Data and Network Science, 8 **2**, 773–788 (2024). https://doi.org/10.5267/j.ijdns.2024.1.001

[19] Nainggolan, N., & Titaley, J. Development of generalized space time autoregressive (GSTAR) model. AIP Conference Proceedings, 1827 (2017). https://doi.org/10.1063/1.4979450

[20] Engle, R. F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50 **4**, 987 (1982). https://doi.org/10.2307/1912773

[21] Monika, P., Ruchjana, B. N., & Abdullah, A. S. The implementation of the ARIMA-ARCH model using data mining for forecasting rainfall in Bandung city. International Journal of Data and Network Science, 6 **4**, 1309–1318 (2022). https://doi.org/10.5267/j.ijdns.2022.6.004

[22] Bonar, H., Ruchjana, B. N., & Darmawan, G. Development of generalized space time autoregressive integrated with ARCH error (GSTARI - ARCH) model based on consumer price index phenomenon at several cities in North Sumatera province. AIP Conference Proceedings, 1827 (2017). https://doi.org/10.1063/1.4979425

[23] Monika, P., Ruchjana, B.N., Abdullah, A.S., GSTARI-X-ARCH Model with Data Mining Approach for Forecasting Climate in West Java, Computation **10**, 204 (2022). doi:10.3390/computation10120204

[24] Monika, P., Ruchjana, B. N., Abdullah, A. S., & Budiarto, R. Integration of GSTARIMA Model with Heteroskedastic Error and Kriging for Climate Forecasting: A Systematic Review. Applied Mathematics and Information Sciences, 18(3), 551–567 (2024). https://doi.org/10.18576/amis/180307

[25] Kudyba, S., & Davenport, T. H. Big Data, Mining, and Analytics. CRC Press (2014).

[26] Monika, P., Ruchjana, B. N., & Abdullah, A. S. The implementation of GSTARI-X model for forecasting climate change with data mining approach. AIP Conference Proceedings, 3082(1) (2024). https://doi.org/10.1063/5.0201267

[27] Dietrich, D., Heller, B., & Yang, B. Data Science & Big Data Analytics. John Wiley & Sons, Inc (2015).

[28] Ruchjana, B. N. Pemodelan Kurva Produksi Minyak Bumi Menggunakan Model Generalisasi Star. Forum Statistika Dan Komputasi, September(September), 1–6 (2002).

[29] Cramer, J. S. Econometric Applications of Maximum Likelihood Methods. In Econometric Applications of Maximum Likelihood Methods (1986). https://doi.org/10.1017/cbo9780511572050

[30] Ljung, G. M. Diagnostic testing of univariate time series models. Biometrika, 73 **3**, 725–730 (1986). https://doi.org/10.21538/0134-4889-2019-25-2-177-184

[31] P. Sjölander, "A stationary unbiased finite sample ARCH-LM test procedure," Appl Econ, vol. 43, no. 8, pp. 1019–1033, Mar. 2011, doi: 10.1080/00036840802600046.

[32] Epaphra, M. Modeling Exchange Rate Volatility: Application of the GARCH and EGARCH Models. Journal of Mathematical Finance, 07 **01**, 121–143 (2017). https://doi.org/10.4236/jmf.2017.71007

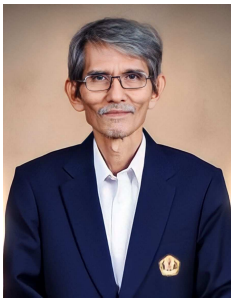
[33] Ishwarappa, & Anuradha, J. A brief introduction on big data 5Vs characteristics and hadoop technology. Procedia Computer Science, 48(C), 319–324 (2015). https://doi.org/10.1016/j.procs.2015.04.188

[34] Larose, D. T. Discovering Knowledge in Data: An Introdcution to Data Mining. John Wiley & Sons, Inc (2005).

Putri Monika, received B.Sc. degree in Mathematics,Faculty Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia in 2021, Magister in Mathematics,Faculty Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia in 2022. Currently, she is a student in doctor program of Mathematics, Faculty Mathematics and Natural Sciences, Universitas

Padjadjaran, Indonesia. Her research interest include Spatio Temporal Modelling, Time Series Analysis, Stochastics Processes, and Big Data Analytics.

Budi Nurani Ruchjana, received B.Sc. degree in Mathematics from Universitas Padjadjaran, Indonesia in 1987, Magister in Applied Statistics from Institut Pertanian Bogor, Indonesia in 1992, and Doctor of Mathematics and Natural Sciences from Institut Teknologi Bandung, Indonesia, in 2002. Currently, she is a full professor at Department of Mathematics, Universitas Padjadjaran, Indonesia. Her research interest include Spatio Temporal Modeling, Stochastics Processes, Time Series Analysis, Spatial Analysis, Geostatistics and Ethnomathematics.



Atje Setiawan Abdullah, received B.Sc. degree in Mathematics from Universitas Padjadjaran, Indonesia in 1985, Magister in Management and Industrial Technology from Institut Teknologi Bandung, Indonesia in 1989, Magister and Doctor of Computer Science from Universitas Gadjah Mada, Indonesia in 2004 and 2009, respectively. Currently, he is a full professor at Department of Computer Science, Universitas Padjadjaran, Indonesia. His research interest include Spatial Data Mining, Management and Information Systems, Decision Support System, and Ethno-informatics.

Rahmat Budiarto, received B.Sc. degree in Mathematics from Bandung Institute of Technology, Indonesia in 1986, M.Eng. and Dr.Eng. in Computer Science from Nagoya Institute of Technology, Japan in 1995 and 1998, respectively. Currently, he is a full professor at Dept. of Computer Science, Albaha University, Saudi Arabia. His research interests include intelligent systems, brain modeling, IPv6, network security, Wireless sensor networks, and MANETs.