Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Development of the GSTARIMA(1,1,1) model order for climate data forecasting

Ajeng Berliana Salsabila[a], Budi Nurani Ruchjana[b*] and Atje Setiawan Abdullah[c]

[a]Master of Mathematics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia
[b]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia
[c]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia

CHRONICLE

Article history:
Received: November 7, 2023
Received in revised format: November 30, 2023
Accepted: January 4, 2023
Available online: January 4, 2024

Keywords:
Data analytics lifecycle
GSTARIMA(3,1,1) model
Big Data
Climate

ABSTRACT

The space-time model combines spatial and temporal elements. One example is the Generalized Space-Time Autoregressive (GSTAR) Model, which improves the Space-Time Autoregressive (STAR) model. The GSTAR model assumes that each location has heterogeneity characteristics, and that the data is stationary. In this research, the moving average component is calculated by involving the relationship between variable values at a certain time and residual values at a previous time, and it is assumed that the data is not stationary, so the model used is the Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) Model. The model order for GSTARIMA is determined through the Space-Time Autocorrelation Function (STACF) and Space-Time Partial Autocorrelation Function (STPACF) to ensure accurate forecasting. Previous research only discussed the GSTARIMA(1,1,1) model, so in this research, the GSTARIMA(3,1,1) model will be addressed as a form of development of the GSTARIMA(1,1,1) model and applied to climate data. The climate data used in this research is sourced from NASA POWER and consists of rainfall variables with large data sizes, requiring the use of the data analytics lifecycle method to analyse Big Data. The lifecycle includes six phases: discovery, data preparation, model planning, model building, communicating results, and operationalization. Based on the data processing results with Python software, the GSTARIMA(3,1,1) model has a MAPE value of 9% for out-sample data and 11% for in-sample data. In contrast, the GSTARIMA(1,1,1) model has a MAPE value of 11% for out-sample data and 12% for in-sample data. So the GSTARIMA(3,1,1) model provides more accurate forecasting results. Therefore, selecting the correct model order is crucial for accurate forecasting.

1. Introduction

The time series is the realization of a stochastic process defined as a sequence of ordered observations over time indices (Wei, 2006). Referring to Box et al. (2015), there is a procedure for modelling time series based on the Box-Jenkins approach, which involves three stages: model identification, parameter estimation, and diagnostic checking. The identification stage aims to determine a suitable model for forecasting, including deciding the model order. The order in time series data serves to describe the number of lags or previous periods, and it also refers to the complexity level of the model used to represent the data. A higher order in a time series implies a more complex model. The appropriate order is crucial for producing a good and accurate model for forecasting or modelling time series. While a higher-order model may capture intricate patterns in time series data, estimating and interpreting can be more challenging. The principle of parsimony is applied to use a simpler model for a more straightforward interpretation. However, time series data may sometimes have intrinsic complexities requiring more complex models, such as in marine ecosystems, climate, global economy, and human social behavior. Therefore, the principle of parsimony is not an absolute rule but rather a guideline to be considered in time series modelling.

* Corresponding author.
E-mail address: budi.nurani@unpad.ac.id (B. N. Ruchjana)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.1.001

Time series can be combined with spatial and is commonly referred to as space-time series. An example of a space-time model is the Space-Time Autoregressive (STAR) model. The assumption in the STAR model is homogeneous location characteristics. According to Wei (2019), the STAR model is a special case form of the Vector Autoregressive (VAR) model. The STAR model was developed by Pfeifer & Deutsch (1980) applied to stationary data to analyze crime rates in Boston. Ruchjana (2002) extended the STAR model to the Generalized STAR (GSTAR) model for forecasting oil production in the Jatibarang field well locations. In the GSTAR model, locations are assumed to have heterogeneous characteristics, and the data is stationary. Borovkova et al. (2008) estimated the parameters of the GSTAR model using Ordinary Least Squares (OLS) and applied it to tea production. Nurhayati et al. (2012) compared GSTAR models of order 2 and order 1 to forecast the Gross Domestic Product (GDP) in Western European countries. Based on Mean Squared Forecast Error (MSFE) results, the GSTAR model of order 2 performed better. Prillantika et al. (2018) used GSTAR order 2 to compare parameter estimates with OLS and Kalman Filter for forecasting inflation rates, finding that parameter estimation using Kalman Filter was superior. Huda and Imro'ah (2023) modelled GSTAR of orders 1, 2, and 3, comparing five weight matrices, including uniform, queen contiguity, Minimum Spanning Tree (MST), inverse distance, and modified by railroad for forecasting Covid-19 in West Java Province. The results showed that the GSTAR model of order 3 using the MST weight matrix was the best forecasting model. Susanti et al. (2018) applied the GSTAR model to non-stationary data, resulting in the Integrated GSTAR model (GSTARI) for forecasting assets of Rural Credit Banks (BPR). Monika et al. (2022) utilizes a data mining approach to predict climate phenomena by combining the GSTARI model with exogenous variables and Autoregressive Conditional Heteroskedasticity (GSTARI-X-ARCH) to overcome heteroscedasticity conditions.

Di Giacinto (2006) developed the Generalized Space-Time Autoregressive Moving Average (GSTARMA) model of order 1 for autoregressive and moving average to analyze unemployment in Italy, using Maximum Likelihood Estimation (MLE) for parameter estimation. Akbar et al. (2020) applied the GSTARMA(1,0,1) model to forecast air pollution in Surabaya using two weight matrices: inverse distance and uniform. Parameter estimation methods included OLS and Seemingly Unrelated Regression (SUR). The research also discussed how the GSTARMA model can improve forecast errors compared to the GSTAR model. Andayani et al. (2017) used the GSTARMA model with added exogenous variables (GSTARMA-X). When applied to non-stationary data, the GSTARMA model requires differencing, resulting in the Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) model. Min and Hu (2010) developed the GSTARIMA(1,1,1) model and applied it to urban traffic network modelling and short-term traffic flow forecasting using the Least Square method for parameter estimation. Mubarak et al. (2022) applied the GSTARIMA(1,1,1) model to missing data to forecast gold prices. Sukarna et al. (2023) used the GSTARIMA (1,1,1) model to forecast COVID-19 in Sulawesi, comparing it with the GSTARI(1,1,0) and GSTIMA(0,1,1) models; both models showed similar accuracy based on Mean Absolute Percentage Error (MAPE).

One of the spatiotemporal phenomena is climate, as it can be observed based on location and time. According to the World Meteorological Organization (WMO) (2022), climate characterizes the average weather conditions for a specific area over a long period, requiring a minimum of 30 years for description. The European Union (2023) states that climate change affects all regions globally across various sectors, including agriculture, health, fisheries, tourism, and more. According to Ahlonsou et al. (2018), climate variables directly impacting daily life include rainfall, which has wet and dry seasons in tropical or subtropical regions. Wet months are typically characterized by higher rainfall and more frequent rain events, while dry months experience lower rainfall and less frequent rain. Wet months often occur in December, January, and February (DJF), and dry months in June, July, and August (JJA). The amount of climate data being observed is increasing rapidly, leading to the emergence of "Big Data" - huge and complex data sets that cannot be easily managed, processed, or analyzed using conventional techniques. Based on Dietrich et al. (2015) Big Data is characterized by 3Vs: volume, variety, and velocity. Volume refers to the size of the data, which can be in billions of rows and millions of columns. Variety describes the different forms in which data is produced, including structured and unstructured data. Velocity refers to the speed at which new data is created and grows. According to a survey by Transforming Data With Intelligence, organizations that implemented Big Data Analytics saw improvements in marketing focus, business insights, and client-based segmentation. A study by McKinsey Global Institute (2011) discovered that data holds comparable significance for organizations as both labor and capital. Those organizations adept at efficiently collecting, analyzing, visualizing, and leveraging insights from Big Data can set themselves apart from competitors, surpassing them in terms of operational efficiency. Organizations in any industry with Big Data can benefit from its careful analysis to gain insight and depth to solve real problems. The data analytics lifecycle method can be used to analyze Big Data. This method is designed specifically for Big Data and has six phases: discovery, data preparation, model planning, model building, communication results, and operationalized. By applying this method, Binbusayyis and Vaiyapuri (2019) were able to identify the main features of a cyber intrusion detection system. Based on the above exposition, this research develops a GSTARIMA(1,1,1) model order for climate data forecasting. The chosen model order aligns with model identification results using the Time Autocorrelation Function (STACF) and Space-Time Partial Autocorrelation Function (STPACF) since climate data falls into intrinsic complexities that require a more complex model. Additionally, parameter estimation for the GSTARIMA model is performed with model orders more significant than one using Maximum Likelihood Estimation (MLE). The climate data for this research is obtained from the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASA POWER), which falls under Big Data, and the data analytics lifecycle method is employed. Mean Absolute Percentage Error (MAPE) is used as the evaluation metric for the forecasting method.

2. *Materials*

This section discusses the theories that support research following the Box Jenkins procedure, starting from model identification, parameter estimation, and diagnostic checking—as well as the data analytics lifecycle method for analyzing Big Data.

2.1. Data Analytics Lifecycle

According to Dietrich et al. (2015), the data analytics lifecycle is specifically designed for data science projects and Big Data problems. The lifecycle consists of six phases, namely discovery, data preparation, model planning, model building, communicating results, and operationalization. These phases can be observed in Fig. 1. In most of the phases, the process can move forward to the next phase or return to the previous one.

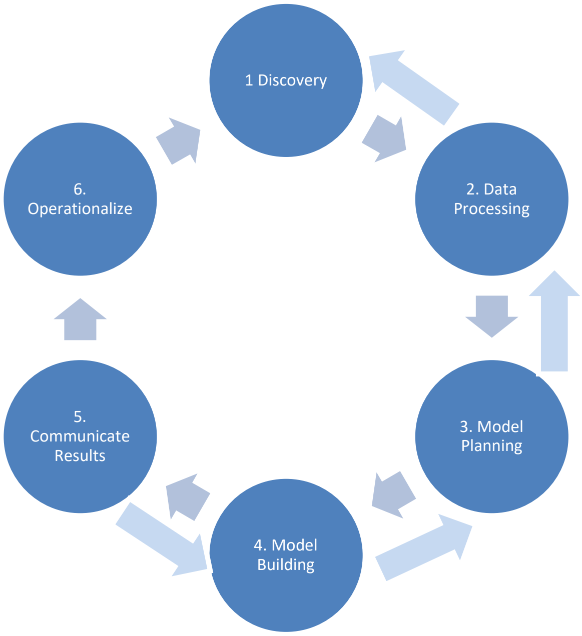


Fig. 1. Data analytics lifecycle phases (Dietrich et al. 2015)

2.2. Box-Jenkins Procedure

The Box-Jenkins method is a time series analyse technique that was introduced by George E. P. Box and Gwilym M. Jenkins. It is a three-stage procedure that involves model identification, parameter estimation, and diagnostic checking (Box et al. 2015). The model identification stage involves selecting a suitable model for the forecasting process and determining the model order. For univariate time series, the data must meet the stationarity requirements before proceeding with the Autocorrelation Function (ACF) and Partial ACF (PACF) plots. For multivariate time series, the ACF Matrix (MACF) and MPACF can be used, while for space-time series, Space-Time ACF (STACF) and STPACF are appropriate. The model parameter estimation stage is focused on estimating the model parameters, and the diagnostic checking stage aims to test the suitability and feasibility of the forecasting model. A feasible model should have significant parameters and the resulting errors should not have a particular pattern. Additionally, the process requires that white noise or errors are independent and have a normal distribution.

2.3. Space Time Autocorrelation Function and Space Time Partial Autocorrelation Function

Wei (2019) identified Space Time Autocorrelation Function (STACF) and Space-Time Partial Autocorrelation Function (STPACF) as essential tools in space-time analysis to identify correlation patterns. Eq. (1) is used to determine STACF.

$$\hat{\rho}_l(k) = \frac{\sum_{t=k+1}^{n}\left(\mathbf{W}^{(l)}\mathbf{Z}(t-k)\right)'\left(\mathbf{Z}(t)\right)}{\sqrt{\sum_{t=1}^{n}\left(\mathbf{W}^{(l)}\mathbf{Z}(t)\right)'\left(\mathbf{W}^{(l)}\mathbf{Z}(t)\right)\sum_{t=1}^{n}\left(\mathbf{Z}(t)\right)'\left(\mathbf{Z}(t)\right)}}. \tag{1}$$

To determine STPACF, the Yule-Walker equation is used which can be seen in Eq. (2), with spatial order λ being the last coefficient of $\phi_{lk} = (l = 0,1,2,\dots,\lambda$ dan $k = 1,2,3,\dots)$.

$$\begin{bmatrix}\begin{bmatrix}\gamma_{00}^{(1)}\\ \gamma_{10}^{(1)}\\ \vdots\\ \gamma_{\lambda0}^{(1)}\end{bmatrix}\\ \begin{bmatrix}\gamma_{00}^{(2)}\\ \gamma_{10}^{(2)}\\ \vdots\\ \gamma_{\lambda0}^{(2)}\end{bmatrix}\\ \vdots\\ \begin{bmatrix}\gamma_{00}^{(k)}\\ \gamma_{10}^{(k)}\\ \vdots\\ \gamma_{\lambda0}^{(k)}\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\begin{bmatrix}\gamma_{00}^{(0)} & \gamma_{01}^{(0)} & \cdots & \gamma_{0\lambda}^{(0)}\\ \gamma_{10}^{(0)} & \gamma_{11}^{(0)} & \cdots & \gamma_{1\lambda}^{(0)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(0)} & \gamma_{\lambda0}^{(0)} & \cdots & \gamma_{\lambda\lambda}^{(0)}\end{bmatrix} & \begin{bmatrix}\gamma_{00}^{(-1)} & \gamma_{01}^{(-1)} & \cdots & \gamma_{0\lambda}^{(-1)}\\ \gamma_{10}^{(-1)} & \gamma_{11}^{(-1)} & \cdots & \gamma_{1\lambda}^{(-1)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(-1)} & \gamma_{\lambda0}^{(-1)} & \cdots & \gamma_{\lambda\lambda}^{(-1)}\end{bmatrix} & \cdots & \begin{bmatrix}\gamma_{00}^{(1-k)} & \gamma_{01}^{(1-k)} & \cdots & \gamma_{0\lambda}^{(1-k)}\\ \gamma_{10}^{(1-k)} & \gamma_{11}^{(1-k)} & \cdots & \gamma_{1\lambda}^{(1-k)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(1-k)} & \gamma_{\lambda0}^{(1-k)} & \cdots & \gamma_{\lambda\lambda}^{(1-k)}\end{bmatrix}\\ \begin{bmatrix}\gamma_{00}^{(1)} & \gamma_{01}^{(1)} & \cdots & \gamma_{0\lambda}^{(1)}\\ \gamma_{10}^{(1)} & \gamma_{11}^{(1)} & \cdots & \gamma_{1\lambda}^{(1)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(1)} & \gamma_{\lambda0}^{(1)} & \cdots & \gamma_{\lambda\lambda}^{(1)}\end{bmatrix} & \begin{bmatrix}\gamma_{00}^{(0)} & \gamma_{01}^{(0)} & \cdots & \gamma_{0\lambda}^{(0)}\\ \gamma_{10}^{(0)} & \gamma_{11}^{(0)} & \cdots & \gamma_{1\lambda}^{(0)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(0)} & \gamma_{\lambda0}^{(0)} & \cdots & \gamma_{\lambda\lambda}^{(0)}\end{bmatrix} & \cdots & \begin{bmatrix}\gamma_{00}^{(2-k)} & \gamma_{01}^{(2-k)} & \cdots & \gamma_{0\lambda}^{(2-k)}\\ \gamma_{10}^{(2-k)} & \gamma_{11}^{(2-k)} & \cdots & \gamma_{1\lambda}^{(2-k)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(2-k)} & \gamma_{\lambda0}^{(2-k)} & \cdots & \gamma_{\lambda\lambda}^{(2-k)}\end{bmatrix}\\ \vdots & \vdots & \ddots & \vdots\\ \begin{bmatrix}\gamma_{00}^{(k-1)} & \gamma_{01}^{(k-1)} & \cdots & \gamma_{0\lambda}^{(k-1)}\\ \gamma_{10}^{(k-1)} & \gamma_{11}^{(k-1)} & \cdots & \gamma_{1\lambda}^{(k-1)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(k-1)} & \gamma_{\lambda0}^{(k-1)} & \cdots & \gamma_{\lambda\lambda}^{(k-1)}\end{bmatrix} & \begin{bmatrix}\gamma_{00}^{(k-2)} & \gamma_{01}^{(k-2)} & \cdots & \gamma_{0\lambda}^{(k-2)}\\ \gamma_{10}^{(k-2)} & \gamma_{11}^{(k-2)} & \cdots & \gamma_{1\lambda}^{(k-2)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(k-2)} & \gamma_{\lambda0}^{(k-2)} & \cdots & \gamma_{\lambda\lambda}^{(k-2)}\end{bmatrix} & \cdots & \begin{bmatrix}\gamma_{00}^{(0)} & \gamma_{01}^{(0)} & \cdots & \gamma_{0\lambda}^{(0)}\\ \gamma_{10}^{(0)} & \gamma_{11}^{(0)} & \cdots & \gamma_{1\lambda}^{(0)}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{\lambda0}^{(0)} & \gamma_{\lambda0}^{(0)} & \cdots & \gamma_{\lambda\lambda}^{(0)}\end{bmatrix}\end{bmatrix}\begin{bmatrix}\begin{bmatrix}\phi_{10}\\ \phi_{11}\\ \vdots\\ \phi_{1\lambda}\end{bmatrix}\\ \begin{bmatrix}\phi_{20}\\ \phi_{21}\\ \vdots\\ \phi_{2\lambda}\end{bmatrix}\\ \vdots\\ \begin{bmatrix}\phi_{k0}\\ \phi_{k1}\\ \vdots\\ \phi_{k\lambda}\end{bmatrix}\end{bmatrix}. \tag{2}$$

2.4. Generalized Space Time Autoregressive Integrated Moving Average Model

According to Wei (2019) the Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) model combines the GSTARI and GSTIMA models. The GSTARIMA model assumes heterogeneous location characteristics, non-stationary data, and white noise or zero average errors with constant uncorrelated variance, independent and normally distributed. The GSTARIMA(p, d, q) model can be expressed in Eq. (3).

$$\mathbf{Y}(t) = \sum_{k=1}^{p}\sum_{l=0}^{\lambda_k} \mathbf{\Phi}_{k,l}\mathbf{W}^{(l)}\mathbf{Y}(t-k) - \sum_{k=1}^{q}\sum_{l=0}^{\gamma_k} \boldsymbol{\theta}_{k,l}\mathbf{W}^{(l)}\mathbf{e}(t-k) + \mathbf{e}(t). \tag{3}$$

where $\mathbf{Y}(t) = \mathbf{Z}(t) - \mathbf{Z}(t-1), \dots, \mathbf{Y}(t-k) = \mathbf{Z}(t-k) - \mathbf{Z}(t-k-1)$ and $\mathbf{e}(t)\overset{\boldsymbol{iid}}{\sim}N(0, \sigma^2)$,

$\mathbf{Z}(t)$: vector of observation variables $(N \times 1)$ at time t,

$\mathbf{Z}(t-k)$: vector of observation variables $(N \times 1)$ at time $(t-k)$,

$\mathbf{\Phi}_{k,l}$: autoregressive and space time parameters at time order k and spatial order l measuring $(N \times N)$ in the form of a diagonal matrix $\left(\mathbf{\Phi}_{kl}^{(1)}, \mathbf{\Phi}_{kl}^{(2)}, \mathbf{\Phi}_{kl}^{(3)}, \dots, \mathbf{\Phi}_{kl}^{(N)}\right)$,

λ_k : spatial order in k-th autoregressive,

$\mathbf{W}^{(l)}$: weight matrix $(N \times N)$ at spatial order l $(l = 1,2,3, \dots)$ contains $w_{ii} = 0$ and $\sum_{i \neq j} w_{ij} = 1$,

$\boldsymbol{\theta}_{k,l}$: moving average and space time parameters at time order k and spatial order l measuring $(N \times N)$ in the form of a diagonal matrix $\left(\boldsymbol{\theta}_{kl}^{(1)}, \boldsymbol{\theta}_{kl}^{(2)}, \boldsymbol{\theta}_{kl}^{(3)}, \dots, \boldsymbol{\theta}_{kl}^{(N)}\right)$,

γ_k : spatial order in k-th *moving average*,

$\mathbf{e}(t)$: error vector $(N \times N)$ at time t,

$\mathbf{e}(t-k)$: error vector $(N \times N)$ at time $(t-k)$.

The GSTARIMA(1,1,1) model is expressed in Eq. (4).

$$\mathbf{Y}(t) = \mathbf{\Phi}_{10}\mathbf{Y}(t-1) + \mathbf{\Phi}_{11}\mathbf{W}^{(1)}\mathbf{Y}(t-1) - \boldsymbol{\theta}_{10}\mathbf{e}(t-1) - \boldsymbol{\theta}_{11}\mathbf{W}^{(1)}\mathbf{e}(t-1) + \mathbf{e}(t). \tag{4}$$

The GSTARIMA(2,1,1) model is expressed in Eq. (5).

$$\begin{aligned}\mathbf{Y}(t) = {} & \mathbf{\Phi}_{10}\mathbf{Y}(t-1) + \mathbf{\Phi}_{20}\mathbf{Y}(t-2) + \mathbf{\Phi}_{11}\mathbf{W}^{(1)}\mathbf{Y}(t-1) + \mathbf{\Phi}_{21}\mathbf{W}^{(1)}\mathbf{Y}(t-2) - \\ & \boldsymbol{\theta}_{10}\mathbf{e}(t-1) - \boldsymbol{\theta}_{11}\mathbf{W}^{(1)}\mathbf{e}(t-1) + \mathbf{e}(t).\end{aligned} \tag{5}$$

The GSTARIMA(3,1,1) model is expressed in Eq. (6).

$$\begin{aligned}\mathbf{Y}(t) = {} & \mathbf{\Phi}_{10}\mathbf{Y}(t-1) + \mathbf{\Phi}_{20}\mathbf{Y}(t-2) + \mathbf{\Phi}_{30}\mathbf{Y}(t-3) + \mathbf{\Phi}_{11}\mathbf{W}^{(1)}\mathbf{Y}(t-1) + \\ & \mathbf{\Phi}_{21}\mathbf{W}^{(1)}\mathbf{Y}(t-2) + \mathbf{\Phi}_{31}\mathbf{W}^{(1)}\mathbf{Y}(t-3) - \mathbf{\theta}_{10}\mathbf{e}(t-1) - \mathbf{\theta}_{11}\mathbf{W}^{(1)}\mathbf{e}(t-1) + \mathbf{e}(t).\end{aligned} \tag{6}$$

2.5. Distance Inverse Weight Matrix

Based on Wei (2019) a weight matrix is a square matrix that contains the weights of the corresponding locations as its elements. In the GSTARIMA model, the weight matrix is calculated based on the distance between locations to determine its elements. On the other hand, the inverse distance weight matrix is a weight matrix that is based on the actual distances between the locations. The inverse distance weight can be calculated using Eq. (7).

$$w_{ij} = \frac{1}{d_{ij}}, \tag{7}$$

where w_{ij} is an element of the distance inverse weight matrix at locations i and j, d_{ij} is the distance from location i to location j. Standardization is performed in the form w_{ij} to obtain the value $\sum_{i\neq j} w_{ij}^{(l)} = 1$. If it is assumed that there are three locations then, the inverse distance weight matrix is in the Eq. (8).

$$\mathbf{W} = [w_{ij}] = \begin{bmatrix} 0 & \frac{w_{12}}{w_{12}+w_{13}+w_{14}} & \frac{w_{13}}{w_{12}+w_{13}+w_{14}} \\ \frac{w_{21}}{w_{21}+w_{23}+w_{24}} & 0 & \frac{w_{23}}{w_{21}+w_{23}+w_{24}} \\ \frac{w_{31}}{w_{31}+w_{32}+w_{34}} & \frac{w_{32}}{w_{31}+w_{32}+w_{34}} & 0 \end{bmatrix}. \tag{8}$$

2.6. Maximum Likelihood Estimation

Based on Terzi (1995), Maximum Likelihood Estimation (MLE) is a method for estimating regression parameters by maximizing the likelihood function. The probability density function used is expressed in Eq. (9).

$$f(\varepsilon|\boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right). \tag{9}$$

The likelihood function is as follows:

$$L(\boldsymbol{\beta}|\varepsilon) = f(\varepsilon|\boldsymbol{\beta}).$$

The natural logarithm of the likelihood function is used, so that it becomes an Eq. (10).

$$\begin{aligned}\ln L(\boldsymbol{\beta}|\varepsilon) &= \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right)\right), \\ &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}.\end{aligned} \tag{10}$$

The Eq. (10) is derived from $\boldsymbol{\beta}$ to maximize the likelihood function, then we get eEq. (11).

$$\begin{aligned}\frac{\partial \ln L(\boldsymbol{\beta}|\varepsilon)}{\partial\boldsymbol{\beta}} &= -\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(-\mathbf{X})}{2\sigma^2} = 0, \\ &\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}}{2\sigma^2} = 0.\end{aligned} \tag{11}$$

The Eq. (11) multiplied by $2\sigma^2\mathbf{X}'$, becomes,

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\mathbf{X}' = 0.$$

Next multiply both sides by $(\mathbf{X}\mathbf{X}')^{-1}$, so it is obtained as follows:

$$\begin{aligned}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} &= 0, \\ (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' &= 0.\end{aligned}$$

Transpose both sides, to obtain the Eq. (12).

$$\begin{aligned}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta} &= 0, \\ \mathbf{X}\boldsymbol{\beta} &= \mathbf{Y}.\end{aligned} \tag{12}$$

Next multiply $\mathbf{X}'$ with both sides of Eq. (12) to obtain Eq. (13).

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

From Eq. (13), the parameter estimates for the MLE method are obtained in Eq. (14)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{14}$$

The GSTARIMA(3,1,1) model for three locations ($N = 3$), can be expressed as a matrix like Eq. (15).

$$\begin{bmatrix} Y_{(1)}(t) \\ Y_{(2)}(t) \\ Y_{(3)}(t) \end{bmatrix} = \begin{bmatrix} \phi_{10}^{(1)} & 0 & 0 \\ 0 & \phi_{10}^{(2)} & 0 \\ 0 & 0 & \phi_{10}^{(3)} \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-1) \\ Y_{(2)}(t-1) \\ Y_{(3)}(t-1) \end{bmatrix} + \begin{bmatrix} \phi_{20}^{(1)} & 0 & 0 \\ 0 & \phi_{20}^{(2)} & 0 \\ 0 & 0 & \phi_{20}^{(3)} \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-2) \\ Y_{(2)}(t-2) \\ Y_{(3)}(t-2) \end{bmatrix} + \begin{bmatrix} \phi_{30}^{(1)} & 0 & 0 \\ 0 & \phi_{30}^{(2)} & 0 \\ 0 & 0 & \phi_{30}^{(3)} \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-3) \\ Y_{(2)}(t-3) \\ Y_{(3)}(t-3) \end{bmatrix} +$$
$$\begin{bmatrix} \phi_{11}^{(1)} & 0 & 0 \\ 0 & \phi_{11}^{(2)} & 0 \\ 0 & 0 & \phi_{11}^{(3)} \end{bmatrix} \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-1) \\ Y_{(2)}(t-1) \\ Y_{(3)}(t-1) \end{bmatrix} + \begin{bmatrix} \phi_{11}^{(1)} & 0 & 0 \\ 0 & \phi_{11}^{(2)} & 0 \\ 0 & 0 & \phi_{11}^{(3)} \end{bmatrix} \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-1) \\ Y_{(2)}(t-1) \\ Y_{(3)}(t-1) \end{bmatrix} +$$
$$\begin{bmatrix} \phi_{11}^{(1)} & 0 & 0 \\ 0 & \phi_{11}^{(2)} & 0 \\ 0 & 0 & \phi_{11}^{(3)} \end{bmatrix} \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} Y_{(1)}(t-1) \\ Y_{(2)}(t-1) \\ Y_{(3)}(t-1) \end{bmatrix} - \begin{bmatrix} \theta_{10}^{(1)} & 0 & 0 \\ 0 & \theta_{10}^{(2)} & 0 \\ 0 & 0 & \theta_{10}^{(3)} \end{bmatrix} \begin{bmatrix} e_{(1)}(t-1) \\ e_{(2)}(t-1) \\ e_{(3)}(t-1) \end{bmatrix} - \tag{15}$$
$$\begin{bmatrix} \theta_{11}^{(1)} & 0 & 0 \\ 0 & \theta_{11}^{(2)} & 0 \\ 0 & 0 & \theta_{11}^{(3)} \end{bmatrix} \begin{bmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{bmatrix} \begin{bmatrix} e_{(1)}(t-1) \\ e_{(2)}(t-1) \\ e_{(3)}(t-1) \end{bmatrix} + \begin{bmatrix} e_{(1)}(t) \\ e_{(2)}(t) \\ e_{(3)}(t) \end{bmatrix}.$$

For estimate parameters using the MLE method, Eq. (15) was transformed following the simple linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Express $\mathbf{U}_i = \sum_{j=1}^{N} w_{ij}\mathbf{Y}_j$ and $\mathbf{V}_i = \sum_{j=1}^{N} w_{ij}\mathbf{e}_j$, so it becomes Eq. (16).

$$\begin{bmatrix} Y_{(1)}(t) \\ Y_{(2)}(t) \\ Y_{(3)}(t) \end{bmatrix} = \begin{bmatrix} Y_{(1)}(t-1) & 0 & 0 & Y_{(1)}(t-2) & 0 & 0 & Y_{(1)}(t-3) & 0 & 0 & U_{(1)}(t-1) & 0 & 0 & U_{(1)}(t-2) & 0 & 0 & U_{(1)}(t-3) & 0 & 0 & -e_{(1)}(t-1) & 0 & 0 & -V_{(1)}(t-1) & 0 & 0 \\ 0 & Y_{(2)}(t-1) & 0 & 0 & Y_{(2)}(t-2) & 0 & 0 & Y_{(2)}(t-3) & 0 & 0 & U_{(2)}(t-1) & 0 & 0 & U_{(2)}(t-2) & 0 & 0 & U_{(2)}(t-3) & 0 & 0 & -e_{(2)}(t-1) & 0 & 0 & -V_{(2)}(t-1) & 0 \\ 0 & 0 & Y_{(3)}(t-1) & 0 & 0 & Y_{(3)}(t-2) & 0 & 0 & Y_{(3)}(t-3) & 0 & 0 & U_{(3)}(t-1) & 0 & 0 & U_{(3)}(t-2) & 0 & 0 & U_{(3)}(t-3) & 0 & 0 & -e_{(3)}(t-1) & 0 & 0 & -V_{(3)}(t-1) \end{bmatrix} \begin{bmatrix} \phi_{10}^{(1)} \\ \phi_{10}^{(2)} \\ \phi_{10}^{(3)} \\ \phi_{20}^{(1)} \\ \phi_{20}^{(2)} \\ \phi_{20}^{(3)} \\ \phi_{30}^{(1)} \\ \phi_{30}^{(2)} \\ \phi_{30}^{(3)} \\ \phi_{11}^{(1)} \\ \phi_{11}^{(2)} \\ \phi_{11}^{(3)} \\ \phi_{21}^{(1)} \\ \phi_{21}^{(2)} \\ \phi_{21}^{(3)} \\ \phi_{31}^{(1)} \\ \phi_{31}^{(2)} \\ \phi_{31}^{(3)} \\ \theta_{10}^{(1)} \\ \theta_{10}^{(2)} \\ \theta_{10}^{(3)} \\ \theta_{11}^{(1)} \\ \theta_{11}^{(2)} \\ \theta_{11}^{(3)} \end{bmatrix} + \begin{bmatrix} e_{(1)}(t) \\ e_{(2)}(t) \\ e_{(3)}(t) \end{bmatrix}. \tag{16}$$

The Eq. (16), already has the same structure as a simple linear model, where

$$\mathbf{Y} = \begin{bmatrix} Y_{(1)}(t) \\ Y_{(2)}(t) \\ Y_{(3)}(t) \end{bmatrix},$$

$$\mathbf{X} = \left[\begin{matrix} Y_{(1)}(t-1) & 0 & 0 & Y_{(1)}(t-2) & 0 & 0 & Y_{(1)}(t-3) & 0 & 0 & U_{(1)}(t-1) & 0 \\ 0 & Y_{(2)}(t-1) & 0 & 0 & Y_{(2)}(t-2) & 0 & 0 & Y_{(2)}(t-3) & 0 & 0 & U_{(2)}(t-1) \\ 0 & 0 & Y_{(3)}(t-1) & 0 & 0 & Y_{(3)}(t-2) & 0 & 0 & Y_{(3)}(t-3) & 0 & 0 \end{matrix}\right.$$

$$\begin{matrix} 0 & U_{(1)}(t-2) & 0 & 0 & U_{(1)}(t-3) & 0 & 0 & -e_{(1)}(t-1) & 0 & 0 \\ 0 & 0 & U_{(2)}(t-2) & 0 & 0 & U_{(2)}(t-3) & 0 & 0 & -e_{(2)}(t-1) & 0 \\ U_{(3)}(t-1) & 0 & 0 & U_{(3)}(t-2) & 0 & 0 & U_{(3)}(t-3) & 0 & 0 & -e_{(3)}(t-1) \end{matrix}$$

$$\left.\begin{matrix} -V_{(1)}(t-1) & 0 & 0 \\ 0 & -V_{(2)}(t-1) & 0 \\ 0 & 0 & -V_{(3)}(t-1) \end{matrix}\right],$$

$$\boldsymbol{\beta} = \begin{bmatrix} \phi_{10}^{(1)} \\ \phi_{10}^{(2)} \\ \phi_{10}^{(3)} \\ \phi_{20}^{(1)} \\ \phi_{20}^{(2)} \\ \phi_{20}^{(3)} \\ \phi_{30}^{(1)} \\ \phi_{30}^{(2)} \\ \phi_{30}^{(3)} \\ \phi_{11}^{(1)} \\ \phi_{11}^{(2)} \\ \phi_{11}^{(3)} \\ \phi_{21}^{(1)} \\ \phi_{21}^{(2)} \\ \phi_{21}^{(3)} \\ \phi_{31}^{(1)} \\ \phi_{31}^{(2)} \\ \phi_{31}^{(3)} \\ \theta_{10}^{(1)} \\ \theta_{10}^{(2)} \\ \theta_{10}^{(3)} \\ \theta_{11}^{(1)} \\ \theta_{11}^{(2)} \\ \theta_{11}^{(3)} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} e_{(1)}(t) \\ e_{(2)}(t) \\ e_{(3)}(t) \end{bmatrix}.$$

Eq. (9) to Eq. (14) are used to obtain parameter estimates $\widehat{\boldsymbol{\beta}}' = \big(\hat{\phi}_{10}^{(1)}, \hat{\phi}_{10}^{(2)}, \hat{\phi}_{10}^{(3)}, \hat{\phi}_{20}^{(1)}, \hat{\phi}_{20}^{(2)}, \hat{\phi}_{20}^{(3)}, \hat{\phi}_{30}^{(1)}, \hat{\phi}_{30}^{(2)}, \hat{\phi}_{30}^{(3)}, \hat{\phi}_{11}^{(1)}, \hat{\phi}_{11}^{(2)}, \hat{\phi}_{11}^{(3)}, \hat{\phi}_{21}^{(1)}, \hat{\phi}_{21}^{(2)}, \hat{\phi}_{21}^{(3)}, \hat{\phi}_{31}^{(1)}, \hat{\phi}_{31}^{(2)}, \hat{\phi}_{31}^{(3)}, \hat{\theta}_{10}^{(1)}, \hat{\theta}_{10}^{(2)}, \hat{\theta}_{10}^{(3)}, \hat{\theta}_{11}^{(1)}, \hat{\theta}_{11}^{(2)}, \hat{\theta}_{11}^{(3)}\big)$.

2.7. Diagnostic Checking

The diagnostic checking process is conducted to identify errors and determine whether the assumptions have been met or not. The error assumptions that need to be fulfilled are the multivariate nature of white noise and its normal multivariate distribution. The Portmanteau test is executed to verify whether the multivariate properties of white noise are satisfied. Similarly, to avoid the assumption of a normal multivariate distribution, Chi-Square QQ plots are used. The Portmanteau test was first developed by Box and Pierce (1970) then Ljung and Box (1978) set it using standardized autocorrelation values. The Portmanteau test is conducted using Eq. (17). Referring to Tsai and Yang (2005), Chi-Square QQ plots are a graphical technique that aims to check the validity of assumptions in data by calculating the expected values based on the distribution. A normal distribution can approximate data if the plot resembles a straight line.

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\rho_k^2, \tag{17}$$

where n are many samples, and ρ_k^2 autocorrelation at the k-th lag.

2.8. Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) is an evaluation of a forecasting model that considers the influence of actual values. MAPE can be calculated using the Eq. (18) (Lawrence et al., 2009).

$$MAPE = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} \left[\left| \frac{\hat{e}_i(t)}{Z_i(t)} \right| \right] \times 100\%, \tag{18}$$

where,

$Z_i(t)$: actual value of the t-th period at the i-th location,

$\hat{e}_i(t)$: residual of the t-th period at the i-th location,

T : the number of time series observations,

N : many observation locations.

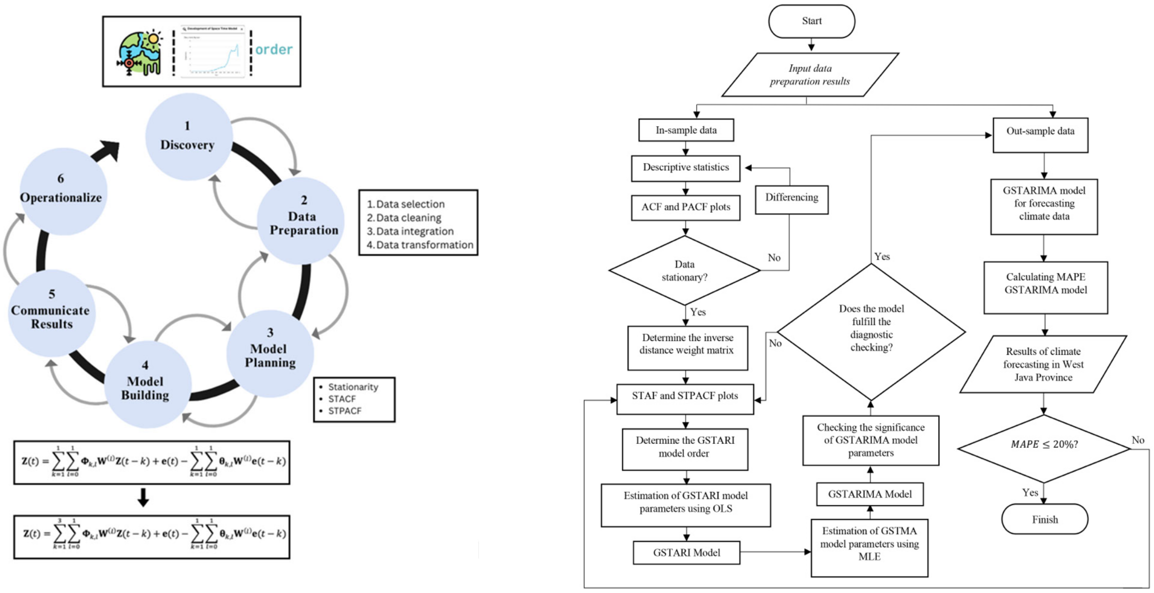
Referring to Lewis (1982) in Lawrence et al. (2009), there is a scale for assessing model accuracy based on the MAPE value which can be seen in Table 1. The smaller the MAPE value, the more accurate the model forecasting results.

Table 1
MAPE value scale

MAPE Scale	Level of accuracy
$\boldsymbol{MAPE} \leq \mathbf{10\%}$	Highly Accurate Forecasting
$\mathbf{10\%} < \boldsymbol{MAPE} \leq \mathbf{20\%}$	Accurate Forecasting
$\mathbf{20\%} < \boldsymbol{MAPE} \leq \mathbf{50\%}$	Reasonable Forecasting
$\mathbf{50\%} < \boldsymbol{MAPE}$	Inaccurate Forecasting

3. Methodology

The research method used is the data analytics lifecycle, which has six phases: discovery, data preparation, model planning, model building, communication results, and operationalization. This research aims to develop the GSTARIMA(1,1,1) model using the Maximum Likelihood Estimation (MLE) method for forecasting climate data in West Java Province.


order
1 Discovery
2 Data Preparation
1. Data selection
2. Data cleaning
3. Data integration
4. Data transformation
3 Model Planning
• Stationarity
• STACF
• STPACF
4 Model Building
5 Communicate Results
6 Operationalize
$\mathbf{Z}(t) = \sum_{k=1}^{1}\sum_{l=0}^{1} \mathbf{\Phi}_{k,l}\mathbf{W}^{(l)}\mathbf{Z}(t-k) + \mathbf{e}(t) - \sum_{k=1}^{1}\sum_{l=0}^{1} \mathbf{\theta}_{k,l}\mathbf{W}^{(l)}\mathbf{e}(t-k)$
$\mathbf{Z}(t) = \sum_{k=1}^{3}\sum_{l=0}^{1} \mathbf{\Phi}_{k,l}\mathbf{W}^{(l)}\mathbf{Z}(t-k) + \mathbf{e}(t) - \sum_{k=1}^{1}\sum_{l=0}^{1} \mathbf{\theta}_{k,l}\mathbf{W}^{(l)}\mathbf{e}(t-k)$
Start
Input data preparation results
In-sample data
Descriptive statistics
ACF and PACF plots
Differencing
Data stationary?
No
Yes
Determine the inverse distance weight matrix
STAF and STPACF plots
Determine the GSTARI model order
Estimation of GSTARI model parameters using OLS
GSTARI Model
Estimation of GSTMA model parameters using MLE
GSTARIMA Model
Checking the significance of GSTARIMA model parameters
Does the model fulfill the diagnostic checking?
No
Yes
Out-sample data
GSTARIMA model for forecasting climate data
Calculating MAPE GSTARIMA model
Results of climate forecasting in West Java Province
$MAPE \leq 20\%$?
No
Yes
Finish

(a) (b)

Fig. 2. Research flow chart

The data used comes from NASA POWER, which can be accessed at https://power.larc.nasa.gov/data-access-viewer/. The data collected on NASA POWER is 51,818,587 TB, which includes big data. In NASA POWER, the data consists of three communities: Agroclimatology, Renewable Energy, and Sustainable Buildings. Climate data is available in the Agroclimatology community. The climate variable used in this research is rainfall.

In general, this research flow uses the Data Analytics Lifecycle which can be seen in Fig. 2 which is also used to help analyze Big Data. Starting from determining the research gap, data sources, hypotheses, data selection, choosing the suitable model, model development, and data processing, to interpreting research results, which can be an appeal to relevant agencies regarding rainfall.

4. *Result and Discussion*

This section contains the results and a discussion of this research, which follows the phases of the data analytics lifecycle.

4.1. Discovery

During the discovery phase, the first step is to identify the problem by studying and investigating it and developing an understanding process. In addition, the data sources needed and available for the research to be conducted are examined, and initial hypotheses, which will later be tested with the data, are formulated.

1) Framing Problem

The ongoing issue of climate change has become a global concern due to its devastating consequences, which pose a significant risk to the survival of all living creatures and the future of upcoming generations. According to the (European Union, 2023) climate change has affected various sectors worldwide, including agriculture, health, fisheries, and tourism. Precipitation is a critical climate variable that influences several aspects of our lives. The scarcity of rainfall can cause droughts, which have a detrimental impact on the ecosystem. Conversely, excessive rainfall can trigger floods and landslides, leading to severe consequences.

Climate data is a crucial component of space-time data, which can be modelled using space-time models. To identify research gaps, a literature review was conducted on space-time models utilizing various sources such as scientific journals, e-books, and previous research. Moreover, bibliometric analysis was employed to investigate the research gap.

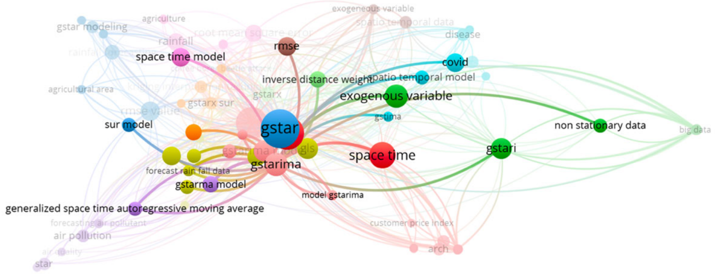


Fig. 3. Bibliometric analysis with keywords "Generalized Space Time Autoregressive" OR "GSTAR"

According to the bibliometric analysis results in Fig. 3, the GSTARIMA model still needs to be developed further. This can be observed from the size of the circle in the GSTARIMA model, which is smaller than that of the GSTAR model. Therefore, this research focused on developing the GSTARIMA model order. Additionally, the GSTARIMA model can capture fluctuations in time series data by considering the error between observed and predicted values. The research aims to forecast climate data using the GSTARIMA model with the appropriate order. Rainfall variables obtained from NASA POWER and included in Big Data. The data analytics lifecycle method was employed to handle Big Data.

2) Identify Data Sources

The research utilizes climate data from West Java Province, specifically focusing on rainfall variables. The data is obtained from NASA POWER, a system and dataset developed by the United States Space Agency to provide information about global energy resources. NASA POWER offers information on various weather parameters such as solar radiation, air temperature, humidity, rainfall, and more, sourced from different sources, including satellites and global climate models. The data volume used in this research is 512,101,087 TB, which is widely distributed.

3) Hypothesis

The development of the GSTARIMA(1,1,1) model order will provide more accurate climate data forecasting using MAPE value criteria.

4.2. Data Preparation

The data preparation phase includes data exploration, pre-processing, or conditioning the data before modeling and analysis.

1) Data Selection

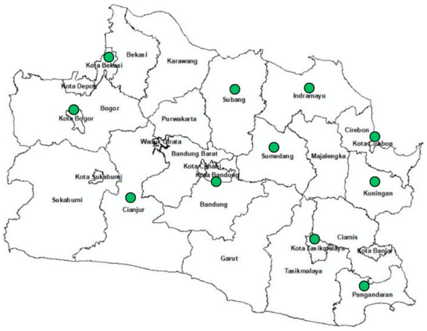


Fig. 4. Map of West Java Province

Daily rainfall data for 27 regencies/cities in West Java Province, collected from the NASA POWER database, amounts to 340,173 or 12,599 data for each regency/city from January 1989 until February 2023. Data from NASA POWER is satellite data with a radius of $57km^2$, so some several regencies/cities have the same data. Locations were selected to represent the same data, and 11 regencies/cities were selected, Bandung City, Bekasi City, Bogor City, Tasikmalaya City, Cianjur Regency, Cirebon City, Sumedang Regency, Indramayu Regency, Subang Regency, Kuningan Regency, and Regency Pangandaran which can be seen in Fig. 4 marked with a green dot on the map.

2) Data Cleaning

Data from 11 regencies/cities that underwent the data selection stage were cleaned to handle missing values using Python software. No missing values were found in 11 regencies/cities data in West Java Province.

3) Data Transformation

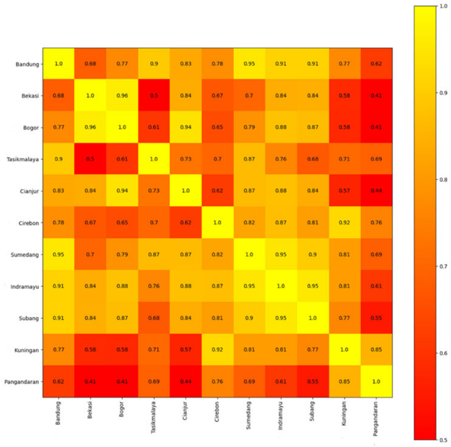
Daily data that has undergone the data cleaning stage is aggregated into monthly data. Obtained 414 monthly data from 12,599 daily data for each regency/city. In this research, the data taken was for December, January, and February (DJF), so from 414 data, 104 monthly data were obtained.

4) Data Integration

In the integration stage, data is combined from various sources. In this research, the data combined is rainfall data in the month of DJF with latitude and longitude coordinate data.

4.3. Model Planning

After the data selection phase, the resulting data is split into two groups: 80% in-sample data, which consists of 83 data (from January 1989-February 2016), and 20% out-sample data, which consists of 21 data (from December 2016-February 2023).



	Bandung	Bekasi	Bogor	Tasikmalaya	Cianjur	Cirebon	Sumedang	Indramayu	Subang	Kuningan	Pangandaran
Bandung	1.0	0.68	0.77	0.9	0.83	0.78	0.95	0.91	0.91	0.77	0.62
Bekasi	0.68	1.0	0.96	0.5	0.84	0.67	0.7	0.84	0.84	0.58	0.41
Bogor	0.77	0.96	1.0	0.61	0.94	0.65	0.79	0.88	0.87	0.58	0.41
Tasikmalaya	0.9	0.5	0.61	1.0	0.73	0.7	0.87	0.76	0.68	0.71	0.69
Cianjur	0.83	0.84	0.94	0.73	1.0	0.62	0.87	0.88	0.84	0.57	0.44
Cirebon	0.78	0.67	0.65	0.7	0.62	1.0	0.82	0.87	0.81	0.92	0.76
Sumedang	0.95	0.7	0.79	0.87	0.87	0.82	1.0	0.95	0.9	0.81	0.69
Indramayu	0.91	0.84	0.88	0.76	0.88	0.87	0.95	1.0	0.95	0.81	0.61
Subang	0.91	0.84	0.87	0.68	0.84	0.81	0.9	0.95	1.0	0.77	0.55
Kuningan	0.77	0.58	0.58	0.71	0.57	0.92	0.81	0.81	0.77	1.0	0.85
Pangandaran	0.62	0.41	0.41	0.69	0.44	0.76	0.69	0.61	0.55	0.85	1.0

Fig. 5. The correlation value between locations

The in-sample data is used for training purposes, while the out-sample data is used for testing the accuracy of rainfall forecasting using the GSTARIMA model. The correlation value was calculated to determine the attachment between locations; in this case Python software was used; the results can be seen in Fig. 5. Based on Fig. 5, the strongest attachment occurs in the Bekasi City area with Bogor City, with a value of 0.96. The process of data centering involved subtracting the data recorded at time t $(Z(t))$ from the average $(\bar{Z})$. Afterward, a stationarity test was performed using the Augmented Dickey-Fuller (ADF) test with the aid of Python software. The *adfuller* function was utilized for this purpose. The results of the ADF test, which can be observed in Table 2, were obtained through this process.

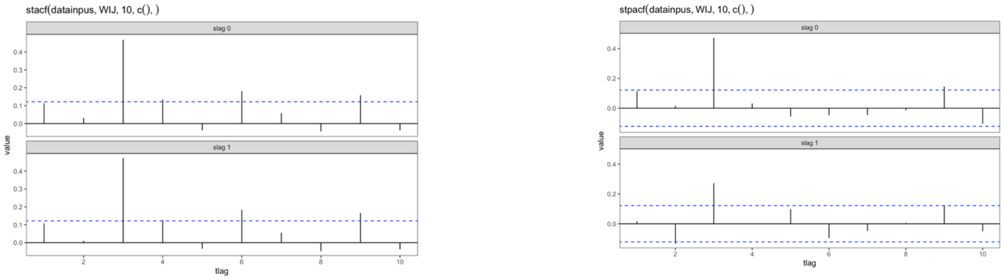
Table 2
Stationarity test of climate data in West Java Province

Location	Variable	Before *Differencing*		After *Differencing*	
		$p-value$	Condition	$p-value$	Condition
Bandung City	$Z_1(t)$	0.1808	Not Stationary	0.01	Stationary
Bekasi City	$Z_2(t)$	0.06637	Not Stationary	0.01	Stationary
Bogor City	$Z_3(t)$	0.0904	Not Stationary	0.01	Stationary
Tasikmalaya City	$Z_4(t)$	0.3268	Not Stationary	0.01	Stationary
Cianjur Regency	$Z_5(t)$	0.1469	Not Stationary	0.01	Stationary
Cirebon City	$Z_6(t)$	0.1483	Not Stationary	0.01	Stationary
Sumedang Regency	$Z_7(t)$	0.1902	Not Stationary	0.01	Stationary
Indramayu Regency	$Z_8(t)$	0.2103	Not Stationary	0.01	Stationary
Subang Regency	$Z_9(t)$	0.1995	Not Stationary	0.01	Stationary
Kuningan Regency	$Z_{10}(t)$	0.09398	Not Stationary	0.01	Stationary
Pangandaran Regency	$Z_{11}(t)$	0.02382	Stationary	0.01	Stationary

Based on Table 2, ten of the eleven locations have $p-value > \alpha$ where $\alpha = 0.05$ so accept H_0 or the data is not stationary. However, Pangandaran Regency's has a $p-value < \alpha$ so reject H_0 or the data is stationary. To achieve data stationarity, a differencing process was conducted by subtracting the t-th time series data from the previous one. Python software was used to assist with this process by utilizing the diff function. The differencing process was applied only once during the research, and the results are shown in Table 2. After the differencing process, the $p-value$ for all locations is smaller than α or reject H_0, which states that the data is stationary. Next, identification model with calculate the Akaike's Information Criterion (AIC) value to determine the best model order at each location. The AIC values obtained for each location can be seen in Table 3. Wei (2019) states that the GSTARIMA model is a special case of the Vector Autoregressive Integrated Moving Average (VARIMA) model, and the VARIMA model is a combination of ARIMA models of the same order and correlation attachment between locations. In this research, locations that have the same order were selected, that the observation locations that have the same order include Bekasi City, Bogor City and Indramayu Regency with the best model order ARIMA(3,1,1); Cianjur Regency and Cirebon City with the best ARIMA model order (4,1,1); and Tasikmalaya City, Sumedang Regency, Kuningan Regency, and Pangandaran Regency with the best model order ARIMA(0,0,1).

Table 3
AIC value to determine the best model order

Location	AIC	Best Model Order
Bandung City	539.085	ARIMA(4,1,2)
Bekasi City	590.838	ARIMA(3,1,1)
Bogor City	561.406	ARIMA(3,1,1)
Tasikmalaya City	537.788	ARIMA(0,1,1)
Cianjur Regency	549.074	ARIMA(4,1,1)
Cirebon City	549.074	ARIMA(4,1,1)
Sumedang Regency	521.643	ARIMA(0,1,1)
Indramayu Regency	536.683	ARIMA(3,1,1)
Subang Regency	561.307	ARIMA(3,1,0)
Kuningan Regency	519.662	ARIMA(0,1,1)
Pangandaran Regency	522.478	ARIMA(0,1,1)



(a) (b)

Fig. 6. STAF and STPACF rainfall data in West Java Province

In this research, Bekasi City, Bogor City, and Indramayu Regency were selected with third-order autoregressive, first-order integrated, and moving average because the model used in this research is the GSTARIMA model and also based on Fig. 5, the attachment between these three locations has a higher correlation, 0.96, 0.88, and 0.84 compared to the correlation between Cianjur Regency and Cirebon City which is only 0.62. STACF and STPACF were computed for all three locations, and the resulting plots can be found in Fig. 6. STPACF was truncated at the third lag, indicating that the best model for this data is a third-order autoregressive model. Therefore, in this research, the $\mathrm{GSTARIMA}(3,1,1)$ model was chosen to develop the $\mathrm{GSTARIMA}(1,1,1)$ model.

4.4. Model Building

Based on model identification results, the GSTARIMA (3,1,1) model was used as a development of the $\mathrm{GSTARIMA}(1,1,1)$ model. The $\mathrm{GSTARIMA}(3,1,1)$ model forecasts rainfall in West Java Province. Locations that meet the criteria are Bekasi City, Bogor City, and Indramayu Regency.

The inverse distance weight matrix is used to determine the weight of a location based on the relationship between locations. Eqs. (7) are used to determine the inverse distance weight matrix. The results of the inverse distance weight matrix calculation can be seen in Eqs. (19), which will then be used to estimate the model parameters.

$$\mathbf{W} = \begin{bmatrix} 0 & 0.737 & 0.263 \\ 0.788 & 0 & 0.212 \\ 0.569 & 0.431 & 0 \end{bmatrix}. \tag{19}$$

The parameters of the $\mathrm{GSTARIMA}(3,1,1)$ model were estimated using MLE as in subchapter 2.6 with the assumption that the error is white noise or zero average with constant uncorrelated, independent variance and a normal distribution $N(0, \sigma_e^2\mathbf{I})$. A Python script was built to assist the data processing process. The results of estimating the parameters of the $\mathrm{GSTARIMA}(3,1,1)$ model using MLE can be seen in Table 4.

Table 4
$\mathrm{GTARIMA}(3,1,1)$ model parameters for forecasting climate data in West Java province

Parameter	Estimator	Parameter	Estimator	Parameter	Estimator	Parameter	Estimator
$\boldsymbol{\phi_{10}^{(1)}}$	-1.076	$\phi_{30}^{(1)}$	0.286	$\phi_{21}^{(1)}$	-0.091	$\theta_{10}^{(1)}$	-0.701
$\boldsymbol{\phi_{10}^{(2)}}$	-0.894	$\phi_{30}^{(2)}$	-0.288	$\phi_{21}^{(2)}$	0.222	$\theta_{10}^{(2)}$	-0.952
$\boldsymbol{\phi_{10}^{(3)}}$	-0.584	$\phi_{30}^{(3)}$	-0.313	$\phi_{21}^{(3)}$	-0.339	$\theta_{10}^{(3)}$	-0.566
$\boldsymbol{\phi_{20}^{(1)}}$	-0.627	$\phi_{11}^{(1)}$	0.339	$\phi_{31}^{(1)}$	-0.434	$\theta_{11}^{(1)}$	-0.131
$\boldsymbol{\phi_{20}^{(2)}}$	-0.959	$\phi_{11}^{(2)}$	0.076	$\phi_{31}^{(2)}$	0.166	$\theta_{11}^{(2)}$	0.133
$\boldsymbol{\phi_{20}^{(3)}}$	-0.253	$\phi_{11}^{(2)}$	-0.165	$\phi_{31}^{(3)}$	0.106	$\theta_{11}^{(3)}$	-0.195

The parameter values that have been obtained in Table 4, when entered into the $\mathrm{GSTARIMA}(3,1,1)$ model at the location of Bekasi City can be seen in Eq. (20), Bogor City in Eq. (21), and Indramayu Regency in Eq. (22).

$$\begin{aligned}\hat{Y}_1(t) = {} & -1.076Y_1(t-1) + 0.251Y_2(t-1) + 0.089Y_3(t-1) - 0.627Y_1(t-2) - \\ & 0.066Y_2(t-2) - 0.023Y_3(t-2) + 0.286Y_1(t-3) - 0.321Y_2(t-3) - \\ & 0.114Y_3(t-3) + 0.701e_1(t-1) - 0.096e_2(t-1) - 0.034e_3(t-1),\end{aligned} \tag{20}$$

$$\begin{aligned}\hat{Y}_2(t) = {} & -1.076Y_1(t-1) + 0.251Y_2(t-1) + 0.089Y_3(t-1) - 0.627Y_1(t-2) - \\ & 0.066Y_2(t-2) - 0.023Y_3(t-2) + 0.286Y_1(t-3) - 0.321Y_2(t-3) - \\ & 0.114Y_3(t-3) + 0.701e_1(t-1) - 0.096e_2(t-1) - 0.034e_3(t-1),\end{aligned} \tag{21}$$

$$\begin{aligned}\hat{Y}_3(t) = {} & -1.076Y_1(t-1) + 0.251Y_2(t-1) + 0.089Y_3(t-1) - 0.627Y_1(t-2) - \\ & 0.066Y_2(t-2) - 0.023Y_3(t-2) + 0.286Y_1(t-3) - 0.321Y_2(t-3) - \\ & 0.114Y_3(t-3) + 0.701e_1(t-1) - 0.096e_2(t-1) - 0.034e_3(t-1),\end{aligned} \tag{22}$$

where $Y(t) = Z(t) - Z(t-1), \ldots, Y(t-k) = Z(t-k) - Z(t-k-1)$, for example, Eq. (20) if return to $Z(t)$ can be expressed in the Eq. (23).

$$\begin{aligned}\hat{Z}_1(t) - Z_1(t-1) = {} & -1.076[Z_1(t-1) - Z_1(t-2)] + 0.251[Z_2(t-1) - Z_2(t-2)] + \\ & 0.089[Z_3(t-1) - Z_3(t-2)] - 0.627[Z_1(t-2) - Z_1(t-3)] - \\ & 0.066[Z_2(t-2) - Z_2(t-3)] - 0.023[Z_3(t-2) - Z_3(t-3)] + \\ & 0.286[Z_1(t-3) - Z_1(t-4)] - 0.321[Z_2(t-3) - Z_2(t-4)] - \\ & 0.114[Z_3(t-3) - Z_3(t-4)] + 0.701\mathrm{e}_1(t-1) - 0.096\mathrm{e}_2(t-1) - \\ & 0.034\mathrm{e}_3(t-1) - Z_1(t-1),\end{aligned} \tag{23}$$

so $\hat{Z}_1(t)$ is expressed in the Eq. (24).

$$\begin{aligned}\hat{Z}_1(t) = {} & -0.076Z_1(t-1) + 0.449Z_1(t-2) + 0.913Z_1(t-3) - 0.286Z_1(t-4) + \\ & 0.251Z_2(t-1) - 0.317Z_2(t-2) - 0.255Z_2(t-3) + 0.321Z_2(t-4) + \\ & 0.089Z_3(t-1) - 0.112Z_3(t-2) - 0.091Z_3(t-3) + 0.114Z_3(t-4) + \\ & 0.701\mathrm{e}_1(t-1) - 0.096\mathrm{e}_2(t-1) - 0.034\mathrm{e}_3(t-1).\end{aligned} \tag{24}$$

Do the same thing for the Eq. (21) and Eq. (22).

Diagnostic checks are performed to ensure that the error is white noise and follows a normal multivariate distribution. The results of the Portmanteau test indicate that the p-value $> \alpha$, where $\alpha = 0.05$, which means the error satisfies the characteristics of white noise. To confirm the normal multivariate distribution, a Chis-Square QQ plot is used, which is shown in Fig.7. The error in the GSTARIMA(3,1,1) model is almost close to the normal line, indicating that it has a normal multivariate distribution. Therefore, the GSTARIMA(3,1,1) model is suitable for forecasting climate data as it meets the necessary assumptions.

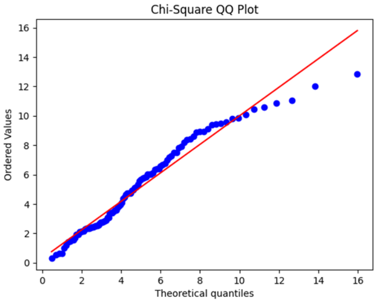
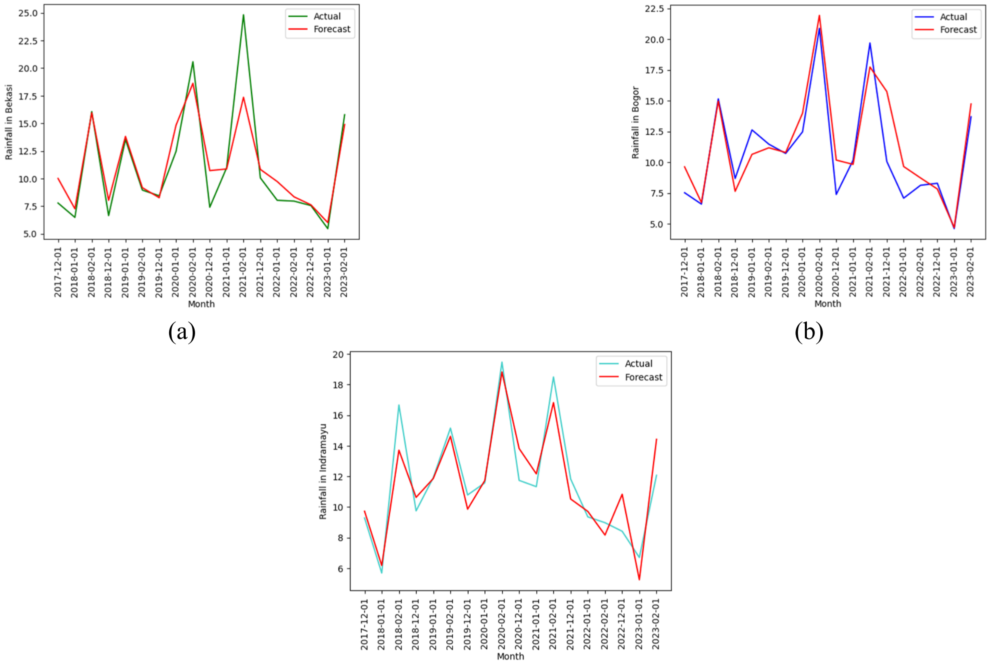


Fig. 7. Chi-Square QQ plot of GSTARIMA$(\mathbf{3},\mathbf{1},\mathbf{1})$ model

The GSTARIMA(3,1,1) model was used to forecast climate data on the out-sample data. The results have been plotted in Fig. 8, which shows that the forecasting pattern matches the actual data. This indicates that using the GSTARIMA(3,1,1) model for forecasting is effective. However, forecasting using the GSTARIMA(3,1,1) model on climate data is only for short-term forecasting, which can be seen in Fig. 8; only the first month to the third month is close to the actual data.



(a)

(b)

(c)

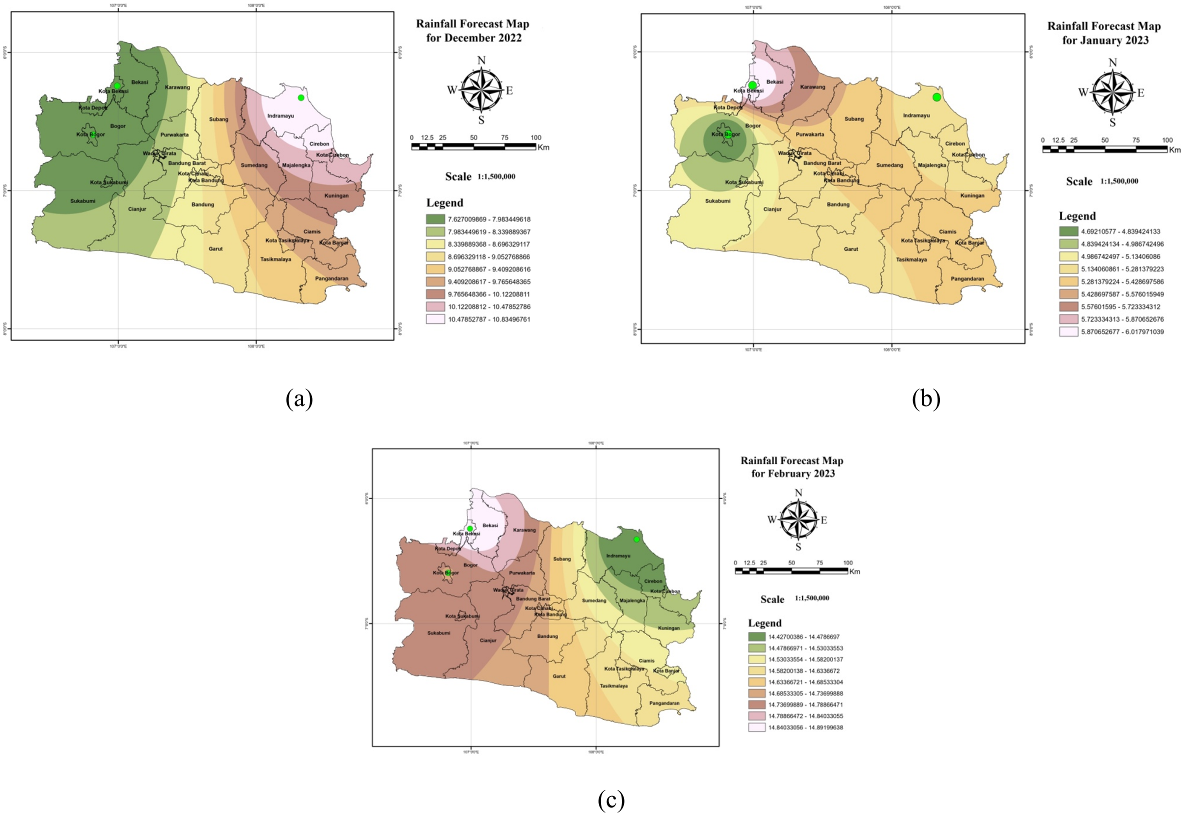
Fig. 8. Plot of actual data and forecasting results of the GSTARIMA$(\mathbf{3},\mathbf{1},\mathbf{1})$ model in Bekasi City (a), Bogor City (b), and Indramayu Regency (c)

Forecasting was conducted using the GSTARIMA(1,1,1) model, following the same process as the GSTARIMA(3,1,1) model. In this case, the principle of parsimony was applied by selecting order one in autoregression. The MAPE values were then calculated for both models to evaluate their forecasting performance. The GSTARIMA(3,1,1) model achieved a MAPE value of 11% for in-sample data and 9% for out-sample data. Meanwhile, the GSTARIMA(1,1,1) model gained a MAPE value of 12% for in-sample data and 11% for out-sample data. As the MAPE value of the GSTARIMA(3,1,1) model is smaller than that of the GSTARIMA(1,1,1) model, the former is better at forecasting climate data, especially rainfall in West Java Province.

4.5. Communicate Result

Based on the model building results, the GSTARIMA(3,1,1) model was used to forecast climate data, and the MAPE results for the out-sample data were 9%, which is considered very accurate. This answers the hypothesis in discovery phase, which states that the development of the GSTARIMA(1,1,1) model order provides more accurate forecasting results.

A visualization was conducted on rainfall forecasting data in West Java Province, specifically in Bekasi City, Bogor City, and Indramayu Regency. The results of the visualization of the forecasting data for December 2022 until February 2023 can be seen in Fig. 9.



(a)

(b)

(c)

Fig. 9. Rainfall forecast map for December 2022 (a), January 2023 (b), February 2023 (c)

Based on Fig. 9, the location with the highest rainfall in December 2022 was Indramayu Regency, with a recorded amount of $10.83496761mm$. On the other hand, Bekasi City and Bogor City had the lowest rainfall in December 2022, ranging from $7.627009869mm$ to $7.983449618mm$. In January 2023, Bekasi City had the highest rainfall, and Bogor City had the lowest rainfall, while in February, Bekasi City had the highest rainfall and Indramayu Regency had the lowest. The rainfall values across the locations were quite similar, as indicated in the legend. These findings suggest that there was an almost even distribution of rainfall intensity in West Java Province, so there was no extreme rainfall in the three regions in December, January, and February (DJF).

4.6. Operationalized

The development of the GSTARIMA(1,1,1) model order results in more accurate forecasting. Selecting the correct order is crucial for the model's effectiveness. This advancement is expected to have a significant impact on the field of modeling. Additionally, using the GSTARIMA(3,1,1) model for forecasting holds significant potential in offering advantages as suggestions to relevant organizations such as the Meteorology, Climatology and Geophysics Agency (BMKG). It also functions as an early warning system for forecasting climate, with a primary focus on rainfall in the province of West Java.

5. Conclusion

The research on climate data forecasting uses the data analytics lifecycle approach to develop the GSTARIMA(1,1,1) model. This approach makes research more structured by beginning with the discovery phase, which helps identify problems and research gaps for model development. The data preparation phase is also helpful for analyzing big data, ranging from 12,599 data to 104 data for each region/city.

In the model planning phase, the best model order for forecasting is determined using STACF and STPACF—estimation using the MLE method for the GSTARIMA(3,1,1) model in the model building phase. The result communication phase helps to analyze forecasting results, and the operationalized stage is a form of implementing solutions or actions based on insights and analysis results.

The research shows that the GSTARIMA (3,1,1) model provides better forecasting results than the GSTARIMA(1,1,1) model, as seen from the MAPE value. In the GSTARIMA(3,1,1) model, the MAPE value for out-sample data is 9%, and for in-sample data is 11%. In contrast, in the GSTARIMA(1,1,1) model, the MAPE value for out-sample data is 11%, and for in-sample data, it is 12%. This confirms that the hypothesis given at the discovery stage is accepted. Based on this research, it can be concluded that selecting the correct order is crucial for the model's feasibility and accuracy in spatiotemporal forecasting. Therefore, it can be concluded that choosing the correct order will produce a more accurate model.

Acknowledgments

The authors thank the Rector, Directorate of Research and Community Service (DRPM), and the Studies Center of Modeling and Computation Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. This research was funded by the Padjadjaran Excellence Fastrack Scholarship (BUPP), Academic Leadership Grant (ALG) grant number 1549/UN6.3.1/PT.00/2023 and supported by international Consortium RISE_SMA project 2019-2024. The authors also grateful to the reviewers of journal for valuable discussion, as well as to Prof. Dr. Sukono, MM., M.Si., Prof. Dr. Diah Chaerani, M.Si., Dr. Firdaniza, M.Si.

References

Ahlonsou, E., Ding, Y., Schimel, D., & Baede, A. P. M. (2018). *The Climate System: an Overview*.

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

Andayani, N., Sumertajaya, I. M., Ruchjana, B. N., & Aidi, M. N. (2017). Development of space time model with exogenous variable by using transfer function model approach on the rice price data. *Applied Mathematical Sciences*, *11*, 1779–1792. https://doi.org/10.12988/ams.2017.74150

Binbusayyis, A., & Vaiyapuri, T. (2019). Identifying and Benchmarking Key Features for Cyber Intrusion Detection: An Ensemble Approach. *IEEE Access*, *7*, 106495–106513. https://doi.org/10.1109/ACCESS.2019.2929487

Borovkova, S., Lopuhaä, H. P., & Ruchjana, B. N. (2008). Consistency and asymptotic normality of least squares estimators in generalized STAR models. *Statistica Neerlandica*, *62*(4), 482–508. https://doi.org/10.1111/j.1467-9574.2008.00391.x

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *TIME SERIES ANALYSIS* (Fifth edition). Wiley.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, *65*(332), 1509–1526. https://doi.org/10.1080/01621459.1970.10481180

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

Dietrich, D., Heller, B., & Yang, B. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Inc.

European Union. (2023). *Consequences of climate change*. https://climate.ec.europa.eu/climate-change/consequences-climate-change_en (accessed May 2023)

Huda, N. M., & Imro'ah, N. (2023). Determination of the best weight matrix for the Generalized Space Time Autoregressive (GSTAR) model in the Covid-19 case on Java Island, Indonesia. *Spatial Statistics*, *54*. https://doi.org/10.1016/j.spasta.2023.100734

Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2009). *Fundamentals of forecasting using Excel*. Industrial Press.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *68*(2), 297–303. http://biomet.oxfordjournals.org/

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. www.mckinsey.com/mgi.

Min, X., & Hu, J. (2010). Urban Traffic Network Modeling and Short-term Traffic Flow Forecasting Based on GSTARIMA Model. *IEEE Conference on Annual Conference on Intelligent Transportation Systems*. https://doi.org/https://doi.org/10.1109/ITSC.2010.5625123

Monika, P., Ruchjana, B. N., & Abdullah, A. S. (2022). GSTARI-X-ARCH Model with Data Mining Approach for Forecasting Climate in West Java. *Computation*, *10*(12). https://doi.org/10.3390/computation10120204

Mubarak, F., Aslanargun, A., & Sıklar, İ. (2022). GSTARIMA Model with Missing Value for Forecasting Gold Price. *Indonesian Journal of Statistics and Its Applications*, *6*(1), 90–100. https://doi.org/10.29244/ijsa.v6i1p90-100

Nurhayati, N., Pasaribu, U. S., & Neswan, O. (2012). Application of generalized space-time autoregressive model on GDP data in West European countries. *Journal of Probability and Statistics*. https://doi.org/10.1155/2012/867056

Pfeifer, P. E., & Deutsch, S. J. (1980). A Three-Stage Iterative Procedure for Space-Time Modeling Space-time modeling STARIMA STAR STMA Time series modeling Three-stage model building procedure. *Technometrics*, *0*(1).

Prillantika, J. R., Apriliani, E., & Wahyuningsih, N. (2018). Comparison between GSTAR and GSTAR-Kalman Filter models on inflation rate forecasting in East Java. *Journal of Physics: Conference Series*, *974*(1). https://doi.org/10.1088/1742-6596/974/1/012039

Sukarna, S., Syahrul, N. F., Sanusi, W., Aswi, A., Abdy, M., & Irwan, I. (2023). Estimating and forecasting covid-19 cases in sulawesi island using generalized space-time autoregressive integrated moving average model. *Media Statistika*, *15*(2), 186–197. https://doi.org/10.14710/medstat.15.2.186-197

Susanti, S., Handajani, S. S., & Indriati, D. (2018). GSTARI model of BPR assets in West Java, Central Java, and East Java. *Journal of Physics: Conference Series*, *1025*(1). https://doi.org/10.1088/1742-6596/1025/1/012119

Terzi, S. (1995). Maximum Likelihood Estimation of A Generalize STAR(p; lp) Model. *Journal of ltalian Statistical Society* (Vol. 3).

Tsai, D.-M., & Yang, C.-H. (2005). A quantile-quantile plot based pattern matching for defect detection A quantile-quantile plot based pattern matching for defect inspection. *Pattern Recognition Letters*.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods* (Second Edition). Pearson.

Wei, W. W. S. (2019). *Multivariate Time Series Analysis and Applications* (First Edition). John Wiley & Sons Ltd. http://www.wiley.com/go/wsps

World Meteorological Organization (WMO). (2022). *State of the Global Climate in 2022*. https://public.wmo.int/en/our-mandate/climate/wmo-statement-state-of-global-climate (accessed May 2023)

https://power.larc.nasa.gov/data-access-viewer/ (accessed June 2023)