Comparison of Generalized Linear Model between Gamma and Tweedie Compound Response for Rainfall Prediction in Lampung Province

ABSTRACT

Rainfall plays a pivotal role in influencing agricultural production in Lampung province. The precision of rainfall predictions holds significant importance for enhancing agricultural yields in the region. One effective approach for modeling rainfall is Statistical Downscaling (SD), which employs statistical models to examine the correlation between large-scale (global) climatological data and small-scale (local) data. SD addresses the limitation of global scale data, such as the General Circulation Model (GCM), which lacks the resolution to directly forecast localized climate conditions like rainfall. Rainfall can be broadly categorized into continuous and discrete components. The continuous component delineates the intensity of rainfall, while the discrete component describes the occurrence of rain. Both components are integral to accurate rainfall predictions. The mixed Tweedie distribution, combining Gamma and Poisson distributions, is proficient in handling both continuous and discrete components of rainfall. GCMs commonly encounter multicollinearity issues in SD modeling, which can be mitigated through Principal Component Analysis. This study seeks to compare two regression models: the generalized linear model with a gamma response and the Tweedie compound response. Rainfall data from three distinct regions in Lampung province, representing high, medium, and lowlands, is utilized. The research findings indicate that, for high and lowlands, the Tweedie compound exhibits a smaller Root Mean Square Error of Prediction (RMSEP) compared to gamma. Conversely, in medium lands, gamma-GLM demonstrates a smaller RMSEP value than the Tweedie compound. Thus, the distribution of the Tweedie compound is better suited for use than Gamma-GLM, especially for high and lowland areas.

Keywords: Statistical Downscaling, Gamma Distribution, Tweedie Compound, Rainfall, General circulationModel..

1. INTRODUCTION

Agricultural activities are profoundly impacted by climatic factors, with temperature, relative air humidity, solar radiation, and rainfall playing significant roles. Rainfall, being a fundamental aspect of agricultural production, holds a crucial role, and comprehension of its potential and variability is imperative for the success of these endeavors [1]. The climate changes currently occurring are affecting rainfall patterns throughout the world. This occurs because higher average air temperatures produce higher evaporation rates, higher water vapor content, and result in an accelerated hydrological cycle [2]. Rainfall is one of the climate elements in tropical areas that has high variations so that it often requires quite complicated statistical modeling to make estimates [3]. Estimating rainfall through statistical modeling is very important to increase rice productivity in Indonesia. One statistical modeling

that can be used to model rainfall in a particular area is Statistical Downscaling (SD) modeling.

Statistical downscaling is a technique in climatology that uses statistical modeling to create functional relationships between large-scale (global) data and small-scale (local) data [4]. The SD model involves General Circulation Model (GCM) output data in the form of precipitation as an explanatory variable and has an important role in predicting rainfall and local scale data in the form of rainfall is used as a response variable. Climate data (GCM and rainfall) are generally non-stationary in space and time, dynamic and non-linear, non-Gaussian distribution and do not even have a standard distribution. [5]

The processes of rainfall can be divided into two categories: one pertains to the quantity of rain on days with precipitation, and the other involves the pattern of dry and wet days[6].The wet season is marked by a brief dry spell in the middle of summer, while the dry season often experiences prolonged periods without rainfall [7].Rainfall is essentially comprised of two elements, specifically the continuous and discrete components. The continuous facet delineates the intensity of rainfall, characterized by values exceeding 0. On the other hand, the discrete aspect elucidates the occurrence of either rain or no rain. Rainfall events are recorded when there is precipitation, whereas the absence of rain signifies an intensity value of 0, indicating no recorded rainfall [8, 9].

Typically, rainfall modeling focuses on a singular component. A versatile model for rainfall modeling is a regression model featuring a mixed Tweedie response. The mixed Tweedie distribution amalgamates Poisson and gamma distributions. The underlying regression model is built upon a generalized linear model (GLM), known for its adaptability to distributions beyond the normal distribution.

GCM output in SD often deviates from the assumption of multicollinearity, necessitating attention to ensure meaningful predictions. To address this, the study employs the Principal Component Analysis (PCA) method as a dimension reduction technique. Previous research, such as the work by [10] focusing on normal distributed rainfall modeling using penalty fused lasso, and [5] employing a Tweedie distribution to forecast rainfall in West Java, serves as a backdrop. Against this backdrop, the present research seeks to compare rainfall predictions using Tweedie compound and gamma responses within a generalized linear model, incorporating PCA for handling multicollinearity. The study encompasses three diverse locations: high, medium, and lowland areas.

2. MATERIAL AND METHODS

This sub-chapter explains several materials related to research, including the distribution of the tweedie compound, general circulation models and statistical downscaling.

2.1 Tweedie Compound Distribution.

Tweedie is a specific component of the Exponential Dispersion Model (EDM). The density function of EDM is defined with two parameters, θ and ϕ, as follows [7]:

$$f_y(y|\theta,\phi) = a\,(y,\phi)\exp\left(\frac{1}{\phi}[y\theta - k(\theta)]\right) \qquad (1)$$

Here, θ is a canonical parameter in $\mathbb{R}$, $\phi > 0$ is a dispersion parameter in the range (0, $+\infty$) [11],$k(\theta)$is a cumulant function of the exponential dispersion model. The Tweedie model is a part of this, and its density function$a\,(y,\phi)$ is dependent on the parameter[13].

The Tweedie distribution encompasses various common distributions, such as the normal distribution for p=0, the Poisson distribution for ϕ=1p=1, the Gamma distribution for $p = 2$, the inverse Gaussian distribution for $p = 3$, and for $1 < p < 2$, it becomes a Tweedie compound, capable of modeling both discrete and continuous components simultaneously.In the context of climatology, Tweedie can be utilized to model rainfall,In the context of climatology, Tweedie can be utilized to model rainfall, assuming Y as the total monthly rainfall, N_t as the total number of rain events per month, and y_i as the precipitation from the i-th event. This is mathematically represented as:

$$P(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \forall n \in N_t$$

$$\boldsymbol{N} = \sum_{\boldsymbol{t}\geq \mathbf{1}} \mathbf{1}_{[\boldsymbol{t},\infty)}(\boldsymbol{t})$$

The amount of rainfall, Y, is represented as the total amount of rain from each rain event.If $N = 0$ then $Y = 0$, if $N > 0$ then$Y = \sum_i^{N_t} y_i$. The probability density function for Y for $N > 0$is given by [12]:

$$f(y) = \begin{cases} \frac{\gamma^\alpha}{\Gamma(\alpha)} y^{\alpha-1}e^{-\gamma y}, & y > 0 \\ 0 & y \leq 0 \end{cases}$$

The parameters of the Tweedie Compound {λ, α, γ} are related to the parameters of the Tweedie model {μ, ϕ, p} as follows:

$$\begin{cases} \mu = \lambda\alpha\gamma \\ p = \frac{\alpha+2}{\alpha+1} \\ \phi = \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2-p} \end{cases} \text{ is parameterized by } \begin{cases} \lambda = \frac{\mu^{2-p}}{\phi(2-p)} \\ \alpha = \frac{2-p}{p-1} \\ \gamma = \phi(p-1)\mu^{p-1} \end{cases} \quad (2)$$

According to [14], the probability that it will not rain is given by:

$$\pi = \Pr(Y = 0) = e^{-\lambda} = \exp\left(-\frac{\mu^{2-p}}{\phi(2-p)}\right) \quad (3)$$

This is equivalent to the equation:

$$P(Y, N = n|\lambda, \alpha, \gamma\} = d_0(y)e^{-\lambda} \|_{n=0} + \frac{y^{n\alpha-1}e^{-y/\beta}}{\beta^{n\alpha}\Gamma(n\alpha)}\frac{\lambda^n e^{-\lambda}}{n!} \|_{n>0} \quad (4)$$

Where$d_0(y)$is the Dirac delta function at zero. The joint distribution $P(Y, N = n|\lambda, \alpha, \gamma\}$based on [5] attains a closed-form expression by substituting equation (2) into equation (4). This results in the joint density function represented by by$\{\mu, \phi, p\}$ as:

$$P(Y, N = n|\mu, \phi, p\} = \left[\exp\left(-\frac{\mu^{2-p}}{\phi(2-p)}\right)\right]^{\|_{n=0}}$$

$$* \left[\exp\left(n\left(-\frac{\log(\phi)}{p-1} + \frac{2-p}{p-1}\log\left(\frac{y}{p-1}\right) - \log(2-p)\right) - log\Gamma(n+1)\right.\right.$$
$$\left.\left. -\frac{1}{\phi}\left(\frac{\mu^{1-p}y}{p-1} + \frac{\mu^{2-p}}{2-p}\right) - log\Gamma\left(\frac{2-p}{p-1}n\right) - \log(y)\right)\right]^{\|_{n>0}}$$

2.2. General Circulation Models (GCM) and Statistical Downscaling (SD)

Statistical Downscaling (SD) technique is utilized to address the limitations of General Circulation Models (GCM), which have lower resolution, in predicting climate conditions at a local scale with higher resolution. SD models can be represented by the following formula:

$$y_{n\times 1} = f(X_{n\times k})$$

where:

$y_{t\times 1}$ = represent rainfall
$X_{n\times k}$ = the precipitation data from GCM output.
n = denotes number of observations
k = number of explanatory variables

The General Circulation Model plays a crucial role in studying climate diversity and change. Several reasons contribute to the GCM output data's inability to provide direct information at the local scale: (1). The depiction of spatial solutions regarding the earth's surface structure, particularly topography, is unclear. (2). Atmospheric hydrodynamics exhibit nonlinearity, and there are nonlinear interactions among small-scale grids(3). The abundance of parameters may not be suitable for small-scale processes [15].An illustration of the Statistical Downscaling process is presented in Figure 1.

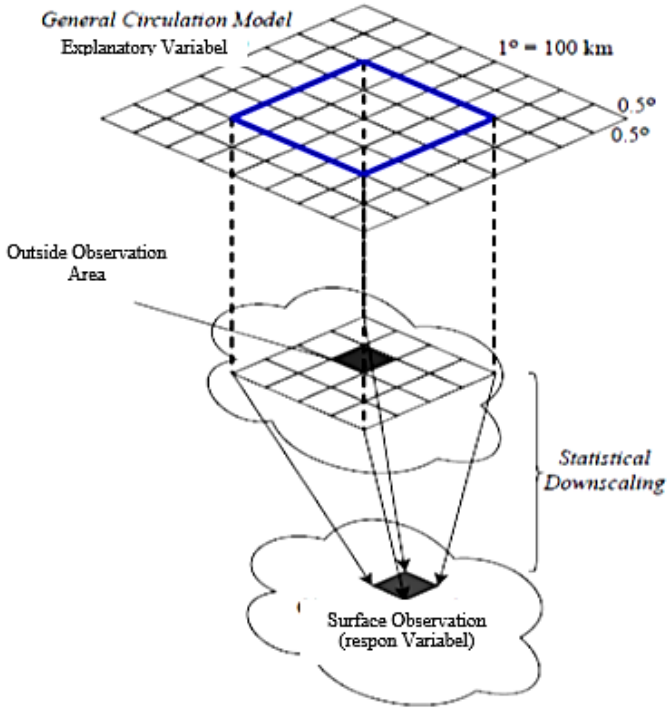

Fig 1. Illustration of Statistical Downscaling

2.3. Data Description Research Stages

Circulation Model (GCM) output, specifically monthly precipitation, serves as the predictor variable. The temporal scope of the analysis spans from January 2013 to December 2019, encompassing a total of 84 months. The geographical focus is on rainfall data situated between -3^{0}North Latitude and -6^0 South Latitude, as well as 103^{0}East Longitude to 106^0 East Latitude in Lampung province. The unit of measurement for the data is mm/day, and the information was sourced from the Meteorology and Geophysics Agency (BMKG) of Lampung Province.The GCM data, utilized as an explanatory variable, was sourced from The National Centers for Environmental Prediction (NCEP) in the form of a Climate Forecast System Reanalysis (CFSR) model, featuring a resolution of 2.5^0 x 2.5^0. This data can be accessed through the website https://rda.ncar.edu/[3, 5]. The analysis involves three rain stations that represent distinct terrains. This approach aims to assess the model's capability to effectively model and predict each terrain, considering their unique characteristics.

Table 1. Description of Data

Variable	$Y_{n\ x\ 1}$	$X_{n\ xk}$
Description	The three rain stations used include: 1. Highlands: Balik Bukit 2. Medium plains: Gisting Atas 3. Lowland: Biha	GCM (General Circulation Model)

n represents the number of observations, and k is the number of variables. Rainfall modeling is conducted using a generalized linear model with Tweedie Compound and Gamma responses in statistical downscaling modeling techniques. R software was employed for data analysis, utilizing the statmod and Tweedie packages to determine index parameters, dispersion, and Tweedie compound regression parameters. The regression models used in this study for Tweedie and gamma are as follows:

$$\log(\mu) = \beta_0 + \beta^T x \quad (5)$$

The regression parameters are denoted as $\beta_0$andβ^T, where x represents the predictor variable, and log (μstands for the link function for the distribution. The data analysis involves the following steps:

1. Examine the data characteristics through density plots, histograms, and boxplots to comprehend the features of rainfall.
2. Identify the index parameters (p) at each station to specify particular distributions within the Tweedie family.
3. Estimate both the phi parameter(ϕ) and the index parameterusing the tweedie package, achieved through the tweedie.profile() function.
4. Apply a generalized linear model for both Tweedie Compound and gamma responses during the modeling process.
5. Generate predictions for rainfall.
6. Assess the model's performance by scrutinizing the Root Mean Squared Error of Prediction (RMSEP) value.

3. RESULTS AND DISCUSSION

The examination of rainfall data will be addressed in this subsection. Following the collection of data, it is partitioned into two segments: training and testing data. Subsequently, a data exploration process is conducted to observe the attributes of rainfall at each rain station, as illustrated in Figure 1.

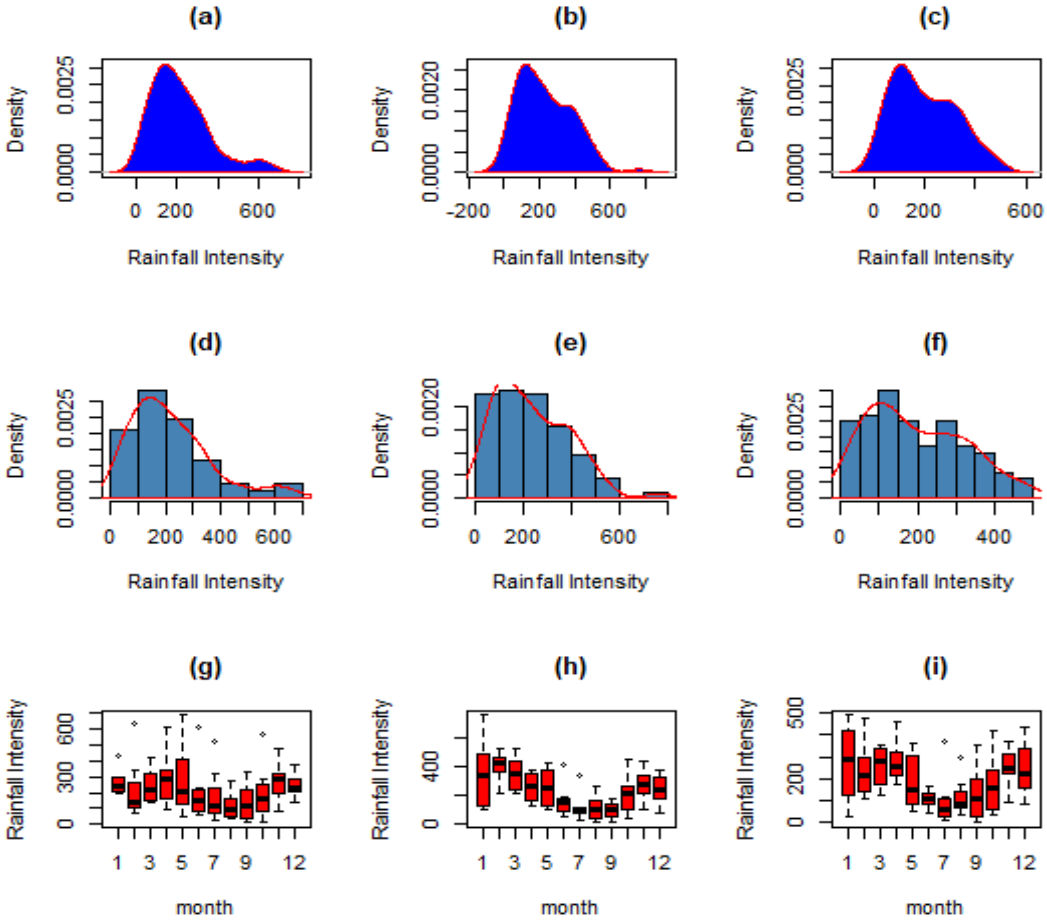

Fig. 2. (a)(b)(c) is a density plot, (d),(e)(f) is a histogram and (g)(h)(i) is a box plot for Balik Bukit, Gisting Atas and Biha stations

Figure 2 (a) (b)(c) is a rainfall density plot from each rain station. The data distribution pattern depicted appears to be skew to the rightand the exact zero is not symmetrical so that the rainfall data. Figure 2 (b) (c) (e) is a histogram for the entire observation area. The picture shows that each station has rainfall data with the characteristics of the data skew to the right and exact zero so that the data at each station is suspected of being included in the Tweedie compound distribution. Figure 2. (g)(h)(i) is a box-plot of rainfall for each rain station which shows that each station has a monsoon rain pattern where the rainfall forms a U shape which has the lowest rainfall intensity from June to September.

Doubts about the distribution of the Tweedie compound need validation through numerical estimation, specifically by estimating the value of the index parameter (p). Two methods are employed to ascertain the value of the index parameter (p): visualization and numerical estimation. Visualization allows for the observation of the index parameter (p) values for each station, as depicted in Figure 3.

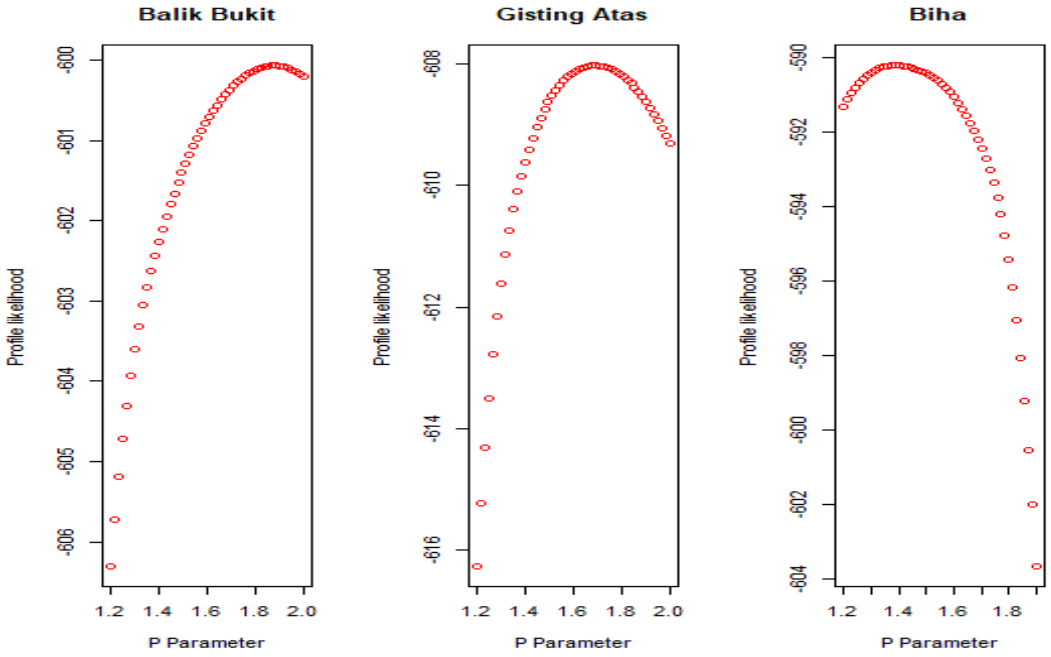

Fig. 3. Profile plot of likelihood parameter p index for rainfall data for Balik Bukit, Gisting Atas and Biha stations.

Figure 3 is a profile likelihood plot of the estimated value of index parameter(p) for rainfall data at each station by selecting the smallestprofile likelihood in each region. In all plots, it can be seen that the index parameters with profile likelihood values for the three locations are around 1.2 to 1.8, meaning that the data for each location follows the Tweedie compound distribution pattern.

The GCM output data encounters various challenges, such as predictor variables with numerous dimensions, spatial correlation among grids, and multicollinearity among variables. To tackle the correlation between predictor variables, this study employs the principal component analysis (PCA) method as a pre-processing technique. Principal component analysis (PCA) techniques are statistical methods employed in the SD model to address the high correlation observed between GCM data grids. PCA specifically emphasizes the variability present in the predictor variables [16].This approach aims to derive latent variables that are mutually orthogonal and form linear combinations of other variables.The number of latent variables used for further analysis is generally determined by at least two of the following three things, namely: screeplot graph, proportion of cumulative diversity and the magnitude of the variance value indicated byeigen value. The number of main components using a screeplot graph is determined by changes in variance that are not significant (shown by a stationary bar/plot graph), while the proportion of cumulative diversity is generally taken to be > 75%. PCA selection is based on eigen values that are >1. This

research selects the main component based on the eigen value which is greater than 1. The main component value based on the eigen value can be seen in Table 2.

Table 2 Eigen values for GCM outcome predictor variables

Principal Component	Eigen value	Principal Component	Eigen value	Principal Component	Eigen value	Principal Component	Eigen value
PC.1	5,25	PC.11	0,48	PC.21	0,15	PC.31	0,00
PC.2	2,37	PC.12	0,36	PC.22	0,14	PC.32	0,00
PC.3	1,60	PC.13	0,26	PC.23	0,14	PC.33	0,00
PC.4	1,27	PC.14	0,26	PC.24	0,13	PC.34	0,00
PC.5	1,05	PC.15	0,23	PC.25	0,12	PC.35	0,00
PC.6	0,80	PC.16	0,21	PC.26	0,01	PC.36	0,00
PC.7	0,66	PC.17	0,18	PC.27	0,00	PC.37	0,00
PC.8	0,05	PC.18	0,18	PC.28	0,00	PC.38	0,00
PC.9	0,75	PC.19	0,17	PC.29	0,00	PC.39	0,00
PC.10	0,52	PC.20	0,15	PC.30	0,00	PC.30	0,00

Table 2 shows that there are 5 main components that have eigen value of more than 1. Next, generalized linear model of the Gamma and Tweedie compound distribution models uses 5 latent variables, namely (PC.1 – PC.5 scores) as predictor variables. Regression analysis was carried out and rainfall predictions for actual 2019 data using the GLM distribution Gamma and Tweedie compound method can be seen in Figure 4.

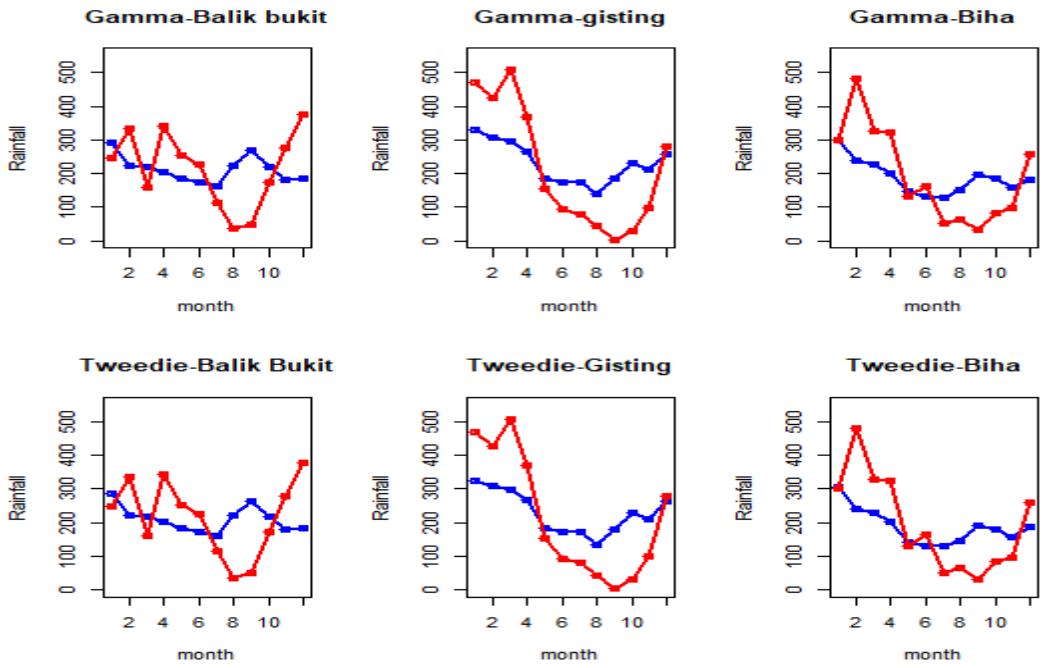

Fig.4. Plot between prediction and actual data for 2019 using the GLM method with Gamma and Tweedie distribution for Balik Bukit, Gisting Atas and Biha stations.

Referring to Figure 4, it is evident that both the GLM approach with gamma distribution and the Tweedie compound exhibit similar patterns in predicting actual data. The assessment of the models is available through the Root Mean Square Error Prediction (RMSEP) value, as indicated in Table 3.

Table 3 RMSEP for GLM with Gamma and Tweedie distributions from the three Rain Stations

Technique	Balik Bukit	Gisting Atas	Biha
Gamma-GLM-PCA	121,78	88,47	110,53
Tweedi-GLM-PCA	120,96	127,96	108,62

Based on the research results, it was found that for the high and lowlands, Tweedie compound has a smaller RMSEP value than Gamma. Meanwhile, the gamma -GLM method has a medium-range RMSEP value that is smaller than the Tweedie compound.

4. CONCLUSION

The outcomes of the aforementioned data analysis lead to the conclusion that the rainfall data analysis method employing the generalized linear Tweedie distribution model proves more effective in predicting rainfall in stations situated in high and lowlands. On the other hand, for temperate plains, the Tweedie compound distribution GLM method is not as effective, and it is preferable to utilize the Gamma distribution GLM method. It's noteworthy that this research utilized a GCM with a resolution of 2.5^0 x 2.5^0. Future studies might benefit from exploring lower resolutions, specifically focusing on the Lampung province area. This approach would be advantageous as the current research employs a GCM covering extensive regions throughout Indonesia.

REFERENCES

[1] Carlos C S J, Casaroli D, Junior J A, Evangelista A W P, Battisti R. Statistical downscaling in the TRMM satellite rainfall estimates for the Goiás state and the Federal District, Brazil. Pesquisa Agropecuária Tropical . 2023; 53: 75552.

[2] Jeon J J, Sung J H, Cung E S, Abrupt change point detection of annual maximum precipitation using fused Lasso, Journal of Hydrology. 2016; 538: 831–841.

[3] Hayati M, Permatasari R. Rainfall Prediction in Statistical Downscaling Using Tweedie Compound Response and Lasso Penalty. International Journal of Scientific Research in Science, Engineering and Technology. 2023; 538; 10(3): 537-546.

[4] Kardiana A, Wigena A H, Djuraidah A, Soleh A M. Asian Journal of Mathematics and Computer Research. 2022; 29(2): 43-55.

[5] Hayati M, Wigena A H, Djuraidah A, Kurnia A. A New ApproachToStatistical Downscaling Using Tweedie Compound Poisson Gamma Response And Lasso Regularization. Commun. Math. Biol. Neurosci, 2021:60.

[6] Suhaila J. Tweedie models for Malaysia rainfall simulations with seasonal variabilities. Journal of Water and Climate Change. 2023; 14 (10): 3648–3670.

[7] Altam R M, Harari O, Moisseeva N, Steyn D. Statistical Modelling of the Annual Rainfall Pattern in Guanacaste, Costa Rica. Water 2023; 15: 700.

[8] Hasan M M, Dunn PK. A simple Poisson–gamma model for modelling rainfall occurrence and amount simultaneously, Agric. Forest Meteorol. 2010; 150: 1319–1330.

[9] Yang Y, Luo R, Khorsidi R, Liu Y, Inferringtweedie compound poisson mixed models with adversarial variational methods, Presented at NIPS Workshop on Advances in Approximate Bayesian Inference, 2017.

[10] Novkaniza F, Hayati M, Sartono B, Notodiputro K A, Fused lasso for modeling monthly raifall in Indramayu sub distric West Java Indonesia, IOP Conf. Ser.: Earth Environ. Sci. 2018; 187: 012046.

[11] Qian W, Yang Y, Zou H, Tweedie's compound poisson model with grouped elastic net, J. Comput. Graph. Stat. 2016; 25. 606–625.

[12] Ma R. Jϕrgensen B. Nested Generalized Linier Mixed Models: an ortodhox Best Linear Unbiased Predictor Approach, J.R. Statist.Soc.B. 2006; 69(4): 625-641.

[13] Dunn P K, Occurrence and quantity of precipitation can be modeled simultaneously. Int. J. Climatol. 2004; 24: 1231–1239.

[14] Dunn PK, Smyth GK. 2005. Series Evaluation of Tweedie Exponential Dispersion Model Densities. Stat. Comput. 2005; 15:267–280. [CrossRef]

[15] Zorita E, von Storch H, The Analog method as a simple statistical downscaling technique: comparison with more complicated methods. J. Climate. 1999; 12: 2474-2489.

[16] Sahriman S, Yulianti A S. Statistical Downscaling Model With Principal Component Regression And Latent Root Regression To Forecast Rainfall In PangkepRegency. BAREKENG: JurnalMatematika dan Terapan. 2023; 17(1): 0401-0410.