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

Comment [F1]: Compound. Check your spellings.

ABSTRACT

Comment [F2]: Make a statement of conclusion after the results as the last sentence.

Rainfall is one of the factors that influences agricultural production in Lampung province. Accurate rainfall predictions are very necessary to increase agricultural production in Lampung Province. One method that can be used to model rainfall is Statistical Downscaling (SD), which uses statistical models to analyze the relationship between large-scale (global) data and small-scale (local) data in climatology. SD overcomes the inability of global scale data such as the General Circulation Model (GCM) as a low resolution predictor to directly predict high resolution local scale climate conditions such as rainfall as a response. In general, rainfall consists of two components, namely continuous and discrete. The continuous component explains the intensity of rainfall while the discrete component explains the incidence of rain. Both components have an important role in rainfall prediction efforts. One distribution that is able to handle both components of rain is the mixed Tweedie distribution, namely the Gamma and Poisson distribution, hereinafter referred to as the Tweedie compound. GCMs generally have multicollinearity problems in SD modeling which can be handled using Principle component analysis. This research aims to compare two regression models, namely the generalized linear model with gamma and tweedie compound response. The data used is rainfall data from three regions in Lampung province representing the high, medium and lowlands. Based on the research results, it was found that for the high and lowlands the tweedie compound had a smaller RMSEP value than gamma, while the medium land gamma -GLM had a smaller RMSEP value than the Tweedie compound.

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

1. INTRODUCTION

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 [1]. 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 [2]. 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. 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, spread non-Gaussian and do not even have a standard distribution. [3]

Rainfall actually consists of two components, namely continuous and discrete. The continuous component explains the intensity of rainfall whose value is more than 0, while the discrete component explains the occurrence of rain and no rain. The occurrence of rain indicates that there is recorded rainfall and no rain event means the intensity has a value of 0 because there is no recorded rainfall[4, 5]. Rainfall modeling generally only models one component. One model that is flexible to use for modeling rainfall is a regression model with a mixed tweedie response. The mixed Tweedie distribution is a mixture of the Poisson and gamma distributions. The regression model used is based on a generalized linear model (GLM). This model is flexible to use for distributions other than the normal distribution.

GCM output in SD often violates the multicollinearity assumption. This needs to be addressed in order to obtain meaningful predictions. This research addresses the multicollinearity problem using a dimension reduction method in the form of the principal component analysis (PCA) method. Research that has been carried out previously includes research by [6] modeling normal distributed rainfall using a penalty fused lasso and [3] using a Tweedie distribution to predict rainfall in West Java. Based on this background, this research aims to compare rainfall predictions using Tweedie compound and gamma responses using a generalized linear model with multicollinearity handling using PCA. Three different locations, namely high, medium and lowland locations, were used in this research.

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 special part of the Exponential Dispersion Model (EDM). The density function of EDM is defined as a function of two parameters, namely:

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

θ is a canonical parameter in $\mathbb{R}$, $\phi > 0$ is a dispersion parameter in (0, $+\infty$) [7],$k(\theta)$is a cumulant function of the exponential dispersion model, $a\,(y,\phi)$ is a size-independent normality quantity of parameter θ [5]. The normalized quantity $a\,(y,\phi)$ can be obtained as follows:

$$(y,\phi) = \begin{cases} \dfrac{\mu_i^{1-p}}{1-p} if y = 0 \\ \dfrac{1}{y}\displaystyle\sum_{n=1}^{\infty} a_n(y,\phi,p) \quad if y > 0 \end{cases}$$

$\sum_{n=1}^{\infty} a_n(y,\phi,p)$ is Wright's generalized Bessel function. $a\,(y,\phi)$ in the Tweedie model it is also a function of$p$. The p parameter value is used as a determinant of the Tweedie distribution. Some common distributions that are included in the tweedie family are known

for their amnalytic forms, including,$p = 0$ is normal distribution, $p = 1, \phi = 1$is Poisson distribution $Tw_1(\mu, 1) \sim Poisson\,(\mu)$,$p = 2$is Gamma distribution $Tw_2(\mu, \phi) \sim Gamma\,(\mu, \phi)$, $p = 3$ is the inverse Gaussian distribution $Tw_3(\mu, \phi) = Invers - Gaussian\,(\mu, \phi)$, $1 < \rho < 2$is a Tweedie compound which can model discrete and continuous components simultaneously, so This distribution can model the occurrence and amount of rainfall simultaneously, $\rho \geq 2$can model positive data, right-skewed data [8]. Tweedie can be used for modeling in the field of climatology by assuming Y is the total monthly rainfall, $N_t$is the total number of rain events per month and $y_i$is the precipitation from the ith event [9] mathematically written as:

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

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

The amount of rainfall is represented as the total amount of rain from each rain event, say $(y_i)_{i \geq}1$, assumed to have an independent and identical Gamma distribution over the time of the rain event:

$$Y = \begin{cases} \sum_{i=1}^{N} y_i & N = 1,2,3,\ldots \\ 0 & N = 0, \end{cases}$$

So $y_i \sim Gamma(\alpha, \gamma)$is a probability density function with mean $\alpha\gamma$ and variance $\alpha\gamma^2$. If$N = 0$ then $Y = 0$, if $N > 0$ then$Y = \sum_i^{N_t} y_i$[12]. The probability density function for Y for$N > 0$ is:

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

Rainfall contains a value of 0 and positive continuous values. So, Y is the total amount of monthly rain, which is represented by the Poisson sum of the random variable Gamma. N is the number of rain events and y_i is the intensity of rainfall for the ith rain event or the amount of rain per day. When Y is the total monthly amount of rain. So Y has a Poisson-Gamma distribution with the following parameters:

λ : average number of rain events per month
γ : Shape of the precipitation event
$\alpha\gamma$:Average amount of precipitation per event.

The relationship between the parameters $\{\lambda, \alpha, \gamma\}$of the Tweedie Compound and the parameters $\{\mu, \phi, p\}$of the Tweedie model is 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 [10] the probability that it will not rain is:

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

Equivalent to the equation:

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

$d_0(y)$Delta dirac function at zero. The joint distribution $P(Y, N = n|\lambda, \alpha, \gamma\}$based on [5] has a close form expression by substituting equation (2) into equation (4). So we get the joint density function represented by $\{\mu, \phi, p\}$as

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

$$* \left[\exp\{n\left(-\frac{\log(\phi)}{p-1} + \frac{2-p}{p-1}\log\left(\frac{y}{p-1}\right) - \log(2-p)\right) - log\Gamma(n+1)\right.$$

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

2.2. Statistical Downscaling and General Circulation Models

Statistical Downscaling (SD) technique is used to overcome the inability of GCM (low resolution) to predict local scale climate conditions (high resolution). SD models have the following formula:

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

where :

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

The General Circulation Model is an important tool in the study of diversity and climate change .Some reasons that GCM output data cannot produce information for local scale directly are: (1) The description of spatial solutions about the structure of the earth's surface, especially topography, is unclear; (2) Atmospheric hydrodynamics are nonlinear and there are nonlinear interactions between small scale grids; (3) Too many parameters that might not be right for small-scale processes [11]. Illustration of Statistical Downscaling is shown in Figure 1.

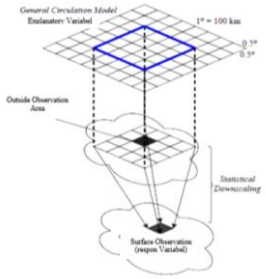

Fig 1. Illustration of Statistical Downscaling

2.3. Data description Research stages

This research uses rainfall data as a response variable and the General Circulation Model (GCM) output in the form of monthly precipitation as a predictor variable with a time period from January 2013 to December 2019 of 84 months. Rainfalldataislocated between-3^0SouthLatitudeand-6^0SouthLatitudeand103^0 southLatitude to 106^0 East Latitude in Lampung province. The data unit used is mm/day. This data was obtained from the Meteorology and Geophysics Agency (BMKG) of Lampung Province. GCM data as an explanatory variable was obtained from The National Centers for Environmental Prediction (NCEP) in the form of a Climate Forcast System Reanalysis (CSFR) model with a resolution of 2.5^0 x 2.5^0 which can be downloaded on the website https://rda.ncar.edu / (Saha, et al. 2010). The locations used are 3 rain stations representing each plain. This is considered to see the ability of the model used to model and predict each terrain which has different characteristics.

Table 1. Data Description

Variable	Description
$\boldsymbol{Y_{nx1}}$	The three rain stations used include: 1. Highlands: Balik Bukit 2. Medium plains: Gisting Atas 3. Lowland: Biha
$\boldsymbol{X_{nxp}}$	GCM (General Circulation Model)

n is the number of observations, and p is the number of variables. Rainfall modeling is modeled using generalized linear Tweedie model with Compound and Gamma response in statistical downscaling modeling techniques. R software was used to assist data analysis with the statmod and tweedie packages for determining index parameters, dispersion and tweedie compound regression parameters. The regression model used in this research for Tweedie and gamma is as follows:

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

β_0 , $\beta^T$is the regression parameter, x is the predictor variable, log (μ) is the link function for distribution.The data analysis steps include the following:

1. Explore data through density plots, histograms and boxplots to see the characteristics of rainfall.

2. Determine the index parameters (p) at each station so that you can specify certain distributions from the tweedie family.
3. The phi parameter (ϕ) is estimated together with the index parameter. Using the tweedie package. With the tweedie.profile( ) function
4. Modeling uses a generalized linear model for the Tweedie Compound and gamma responses
5. Rainfall prediction
6. Evaluate the model by looking at the RMSEP value

3. RESULTS AND DISCUSSION

Analysis of data regarding rainfall will be discussed in this sub-chapter. After the data is collected, it is divided into two parts, namely training and testing data, next is data exploration to see the characteristics of rainfall at each rain station, which can be seen in figure 1.

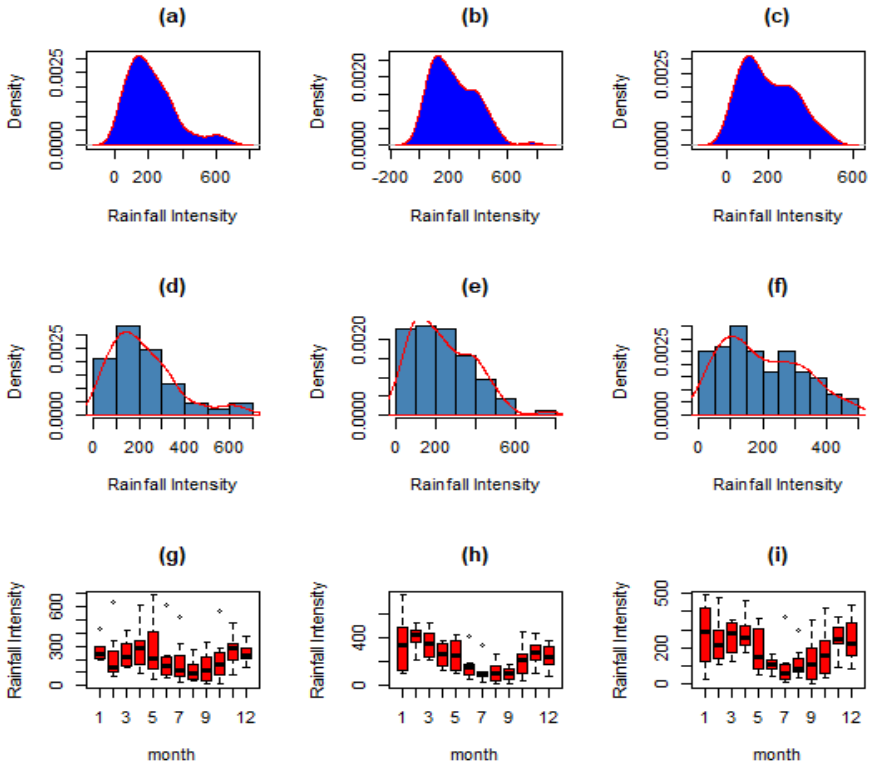

Fig. 2. (a)(b)(c) is a density plot, (d),(e)(f) is a histogram and (g)(h)(i) is a box plot for Balik Bukit, GistingAtas and Biha stations

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 with September. Suspicions regarding the distribution of Tweedie compound need to be proven using numerical estimates by estimating the value of the index parameter (p). There are two ways to determine the value of index parameter (p), namely using visualization and estimating Thevalue of index parameter. Visually, the value ofindexparameter (p) for each station can be seen in Figure 3.

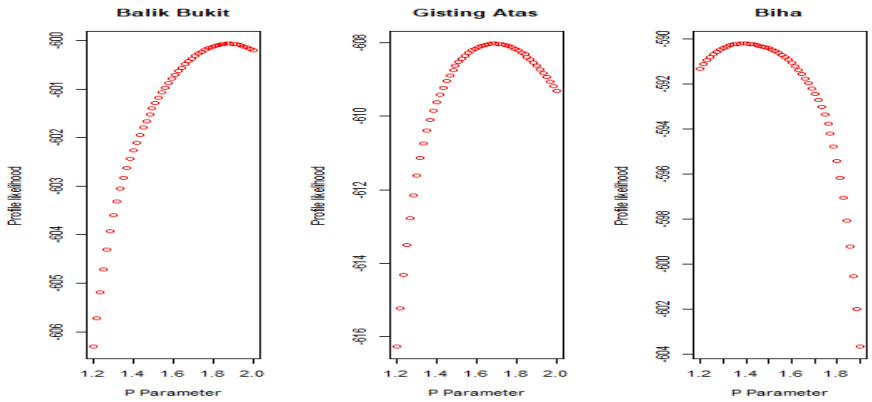

Fig. 3. Profile plot of likelihood parameter p index for rainfall data for Balik Bukit, GistingAtas and Biha stations

Figure 2 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.

GCM output data has several problems, including predictor variables that have many dimensions, spatial correlation between grids and multicollinearity between variables. This research addresses the correlation between predictor variables using the principal component analysis (PCA) method as a pre-processing technique to obtain latent variables that are mutually orthogonal and are a linear combination 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 3.

Table 2 Eigen values for GCM outcome predictor variables

Component	Eigen value	Component	Eigen value	Component	Eigen value	Component	Eigen value
KU.1	5,25	KU.11	0,48	KU.21	0,15	KU.31	0,00
KU.2	2,37	KU.12	0,36	KU.22	0,14	KU.32	0,00
KU.3	1,60	KU.13	0,26	KU.23	0,14	KU.33	0,00
KU.4	1,27	KU.14	0,26	KU.24	0,13	KU.34	0,00
KU.5	1,05	KU.15	0,23	KU.25	0,12	KU.35	0,00
KU.6	0,80	KU.16	0,21	KU.26	0,01	KU.36	0,00
KU.7	0,66	KU.17	0,18	KU.27	0,00	KU.37	0,00
KU.8	0,05	KU.18	0,18	KU.28	0,00	KU.38	0,00
KU.9	0,75	KU.19	0,17	KU.29	0,00	KU.39	0,00
KU.10	0,52	KU.20	0,15	KU.30	0,00	KU.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 (KU1 – KU5 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.

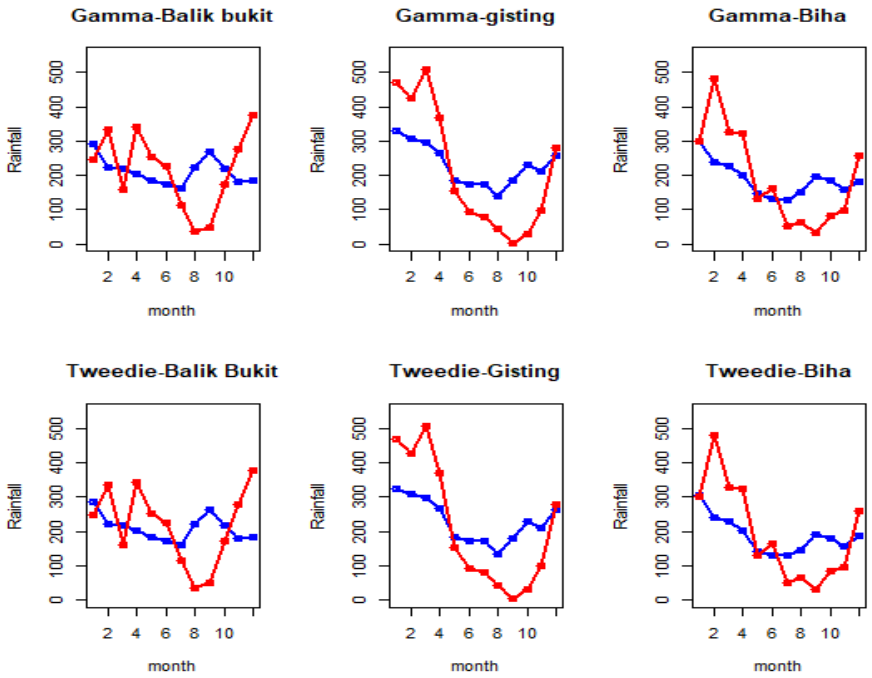

Fig.4. Plot between prediction and actual data for 2019 using the GLM method with Gamma and Tweedie distribution for Balik Bukit, GistingAtas and Biha stations

Based on Figure 4, it can be seen that both the GLM method with gamma distribution and Tweedie compound have prediction results with patterns that are not much different in predicting actual data. Model evaluation can be seen through the Root Mean Square Error Prediction (RMSEP) value shown in Table 3.

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

Metode	Balik Bukit	GistingAtas	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 results of the data analysis above can be concluded that. The rainfall data analysis method using the generalized linear Tweedie distribution model is better used to predict rainfall at stations located in the high and lowlands. Meanwhile, the temperate plains are not good enough if analyzed using the Tweedie compound distribution GLM method and it is better if using the Gamma distribution GLM method. This research used a GCM with a resolution of 2.5^0 x 2.5^0. It would be good if further research was carried out using a lower resolution and narrowed it down to the area around Lampung province only. Because this research uses a GCM which is quite extensive and covers all regions of Indonesia.

Comment [F3]: Do grammar check.

REFERENCES

Comment [F4]: Update the cited articles. Cite at least five additional current articles of not more than three years from the date of last publication.

[1] 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.

[2] 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. 538; 10(3): 537-546.

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

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

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

[6] 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.

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

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

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

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

[11] 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.