Available online at http://scik.org
J. Math. Comput. Sci. 2022, 12:35
https://doi.org/10.28919/jmcs/6965
ISSN: 1927-5307

THE SPATIO-TEMPORAL MODEL FOR THE TWEEDIE COMPOUND POISSON GAMMA RESPONSE IN STATISTICAL DOWNSCALING

MA'RUFAH HAYATI[1,2], AJI HAMIM WIGENA[1,*], ANIK DJURAIDAH[1], ANANG KURNIA[1]

[1]Department of Statistics, IPB University, Bogor 16680, Indonesia

[2]Department of Statistics, The University of Nahdlatul Ulama Lampung, Sukadana 34194, Indonesia

Abstract: This research aims to develop a Spatio-temporal generalized linear mixed model with the h-likelihood estimation method for Statistical Downscaling modeling with the Tweedie compound Poisson Gamma distribution which can produce estimates for fixed effects, random effects, and variance components simultaneously. The results showed that the proposed model has a good performance characterized by the lowest root mean square error prediction and able to reduce the variety of random effects caused by spatial and temporal dependencies.

Keywords: Tweedie compound Poisson-Gamma; hierarchical likelihood; spatio-temporal; generalized linear mixed model; statistical downscaling.

2010 AMS Subject Classification: 93A30.

1. INTRODUCTION

Statistical Downscaling (SD) is a technique in climatology that uses statistical modeling to develop functional relationships between large-scale (global) data and small-scale (local) data. SD

[*]Corresponding author

E-mail address: aji_hw@apps.ipb.ac.id

Received November 05, 2021

modeling involves General Circulation Model (GCM) output data in the form of precipitation used as an explanatory variables (global scale) and rainfall data (local scale) as the response. The GCM output is spatially and temporally related because the data is taken from several grids time to time which results in the multicollinearity problems because the GCM variables are correlated with each other. Multicollinearity can be overcome by dimensional reduction, variable selection, and shrinkage in parameter estimation such as principal component analysis methods, lasso, fused lasso, elasticnet and others.

Rainfall consists of two types of data, namely discrete and continuous. If there is rain or no rain then the type of data is discrete. Meanwhile, if it rains then the intensity of the rain is continuous [1]. Rainfall modeling in SD usually uses two different distributions, separate and does not involve a zero value because there is no rain event. Research on SD has been carried out by [2] which uses the normal distribution with the fused lasso dimension reduction method but is still limited to the intensity of rainfall without f non-rainy events.

[3] proposed a mixed distribution of the Tweedie family to model the two components of rain simultaneously. The Tweedie distribution is a family of distributions that is flexible to overcome non-negative data, highly right-skewed, symmetric data, and exact zero [4]. The Tweedie distribution is a special case of the exponential dispertion model (EDM) which has a variance function of the form $V(\mu) = \mu^p$ which is a function of variance that is proportional to several power or index parameters of the mean [5]. The Tweedie distribution involves a variety of discrete, continuous and mixed probability spreads depending on the index p parameter that is owned. The discrete probability distribution consists of the Poisson distribution with $p = 1$. The continuous distribution consists of the normal distribution with $p = 0$, gamma distribution with $p = 2$ and the inverse-gaussian distribution with $p = 3$. The mixed distribution between the Tweedie family consisting of Poisson and gamma is called Tweedie Compound Poisson Gamma (TCPG) with $1 < p < 2$. The selection of the appropriate distribution according to the occurrence and amount of rain is carried out simultaneously to overcome difficulties in modeling the two components of rain so as not to lose information in making predictions.

SD will be more difficult to model if the data are taken from several locations and times from each location. Such a data structure will cause spatial and temporal dependencies. Thus, modeling must pay attention to these two dependencies, namely spatio-temporal data modeling. Dependencies occur because the data between adjacent locations having a greater closeness of relationship and data from each location being interconnected will occur if the Spatio-temporal data is modeled with a generalized linear model.

Research on spatio-temporal data in SD has been carried out by [6] using the Normal distribution and [7] using three different distributions, namely the gamma distribution to model rainfall, the Bernouli distribution and the generalized Pareto distribution to select extreme rainfall. Both studies used the INLA parameter estimation method based on the Bayes method. However, these two studies have not modeled the two components of rainfall simultaneously.

The SD study using the Tweedie distribution has been carried out by [8] comparing three different models to see which model is the best. The results showed that the mixed distribution of the Tweedie family, also known as the Tweedie compound Poisson gamma (TCPG) distribution with lasso regularization, had good modeling abilities, indicated by the smallest RMSEP value and the large correlation close to one. The model used is still limited to one location and does not consider the temporal dependencies.

The modeling of Spatio-temporal data can be done using a generalized linear mixed model (GLMM) which can handle correlated data due to spatial and temporal dependencies. Estimation of GLMM parameters involves complex integration when obtaining the marginal model [9]. Several methods have been proposed to solve parameter estimation with integral approximation such as the Laplace method, Gauss Hermite Quadrature (GHQ) method. These methods cannot be used for data analysis with more than two random effects because the computational speed will decrease rapidly if more random effects are used [10].

Bayes method is widely used to estimate GLMM parameters with Spatio-temporal random effects [11]. However, the Bayes method requires knowledge of the prior distribution of each estimated parameter and is computationally difficult, especially to obtain the convergence of the predicted parameters. Thus, GLMM requires a parameter estimation method that can handle some of these problems. [12] proposed a method of estimating hierarchical likelihood (h-likelihood) to avoid complex integration, slow or non-converging convergence, and do not require prior distributions for each of the estimated parameters [13][14].

SD research using the GLMM method has been carried out by [15] and [16] but are still limited to the Gaussian distribution. Research on TCPG distribution with GLMM model has been done by Zhang [17] in the field of insurance, but spatial and temporal dependencies have not been included in the model. Thus, this study aims to build a GLMM model that can overcome spatio-temporal dependencies with the Tweedie Compound Poisson-Gamma (TCPG) response for SD modeling called the STGLMM model. The estimation of regression parameters was carried out using the h-likelihood method and multicollinearity in GCM output data was solved using the principal component analysis (PCA) method. Modeling with the TCPG distribution is not only able to predict the intensity of rain like the distribution commonly used but is also able to predict the probability of not raining, and the number of rain events every month.

2. The Development of Spatio-Temporal GLMM Model (STGLMM) with Tweedie Compound Poisson Gamma Response

This study will describe the results of the development of the GLMM model with a spatially and temporally dependent. Tweedie compound Poisson gamma (TCPG) distribution is used as response with involving explanatory variables for fixed effects. The method of estimating parameters used is h-likelihood for estimating fixed effects, random effects, and variance components in the model. The steps of estimating fixed effect parameters, random effects, and variance components are as follows:

1. Specification of the STGLMM model with TCPG distribution.
2. Estimation of fixed and random effect intercept parameters use equation (5)
3. Estimation of the spatial and temporal random variance components random range effects use equation (6)

Solutions for stapes two and three are obtained using the Newton Rapson iteration method.

2.1. Spatio Temporal GLMM Model Spesification with TCPG Distribution

TCPG distribution can be used for modeling in the field of meteorology, especially rainfall. The of characteristics rainfall data are continuous positive and exact zero. In TCPG models of rainfall, Y is the total monthly rainfall, N is the total number of rain events per month and Y_i is the precipitation from the i-th event which has a Poisson distribution $N \sim \text{Pois}(\lambda)$ 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. Suppose $(y_i)_{i\geq 1}$is assumed to have gamma distribution, namely:

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

$y_i \sim Gamma(\alpha, \gamma)$ is a probability density function with mean $\alpha\gamma$ and variance $\alpha\gamma^2$. If $N = 0$ then $Y = 0$, if $N > 0$ then $Y = \sum_i^{N_t} y_i$ [3][18]. For example, the spatio-temporal data $Y(s,t)$ is the observation data assumed to be distributed in TCPG with the $Y(s,t) \sim Tw_p(\mu(s,t), \phi)$ with $1 < p < 2$. (s,t) is the spatio-temporal data notation observed at the location $s = 1,2, \ldots., m$ and month $t = 1,2, \ldots., T$.

[19] state that GLMM is a model with a hierarchical structure. Level one is given for discrete or continuous response variable that follow family of exponential distribution. Level two assigned to unobservable latent is referred to as random effects. The GLMM model for Spatio-temporal data can be written as a hierarchical model, namely:

Level 1 $Y(s,t)|\, b(s), c(t) \overset{iid}{\sim} f\big(y(s,t)\big|\, b(s), c(t)\big) = Tw_p(\mu(s,t)\,, \phi),\ 1 < p < 2.$

Level 2 $b(s), \sim N(0, \sigma_b^2 \boldsymbol{\gamma}(\boldsymbol{d}))$ and $c(t) \sim N\left(0, \frac{\sigma_{e_t}^2}{1-\rho^2}\boldsymbol{\rho}_{\boldsymbol{t},\boldsymbol{t'}}\right)$ for $|\rho| < 1$

$b(s)$ is a spatial random effect, $c(t)$ is a temporal random effect following the first order of the autoregressive process which has the form $c(t) = \rho c(t+1) + \varepsilon_t$, $|\rho| < 1$. $\boldsymbol{\gamma}(\boldsymbol{d})$ is an exponential spatial correlation matrix with $d = \sqrt{(r_i - r_j)^2 + (s_i - s_j)^2}$ as euclidean distance between locations and a is range parameter, $\frac{\mathrm{\boldsymbol{\rho}_{t,t'}}}{1-\mathrm{\rho}^2}$ is temporal correlation matrix [20]. The TCPG distribution belongs to the exponential family. Thus, data modeling can use a link function that connects the observed data expectations with the regression equation, namely:

$$\eta = log(\mu|b,c) = \boldsymbol{X}\beta + \boldsymbol{Z_1}b + \boldsymbol{Z_2}c \tag{1}$$

This research adopts the research results [17] and [21]. Based on [17], the regression model in equation (1) is transformed in the form of a relative correlation factor matrix $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$ which are Cholesky decomposition matrices such that $\boldsymbol{\Sigma_s} = \sigma_b^2\boldsymbol{\gamma}(\boldsymbol{d}) = \sigma_b^2\boldsymbol{LL'}$ and $\boldsymbol{\Sigma_t} = \sigma_t^2 \, \frac{\boldsymbol{\rho_{t,t'}}}{1-\rho^2} = \sigma_t^2 \; \boldsymbol{\Lambda\Lambda'}$. This regression model change aims to adjust to the model developed by [21]. The change in the form of the regression equation in equation (1) becomes the regression equation with the relative correlation matrix $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$ shown in the equation (2)

$$Log(\boldsymbol{\mu}|\ \boldsymbol{b}, \boldsymbol{c}) = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{L}\ \boldsymbol{b} + \boldsymbol{Z}_2\ \boldsymbol{\Lambda}\ \boldsymbol{c} \tag{2}$$

Equation (2) can be written in the form:

$$\boldsymbol{\eta} = Log(\boldsymbol{\mu}|\ \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{X\beta} + \boldsymbol{Z_1^*}\ \boldsymbol{u} + \boldsymbol{Z_2^*}\ \boldsymbol{v} \tag{3}$$

In which $\boldsymbol{Z_1^*} = \boldsymbol{Z_1L}$, $\boldsymbol{Z_2^*} = \boldsymbol{Z_1\Lambda}$, $\boldsymbol{u} \sim N(0, \sigma_b^2\boldsymbol{I})$ and $\boldsymbol{v} \sim N(0, \sigma_t^2\ \boldsymbol{I})$.

2.2. The Estimation of Fixed and Random Effect

The estimation of regression parameters uses the hierarchical likelihood estimation method which has a form as in equation (4):

$$\boldsymbol{h} = log\big(f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v})\big) = log\left(f_{\boldsymbol{\beta},\phi}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{c})\right) + log\left(f_{\sigma_b^2}(\boldsymbol{u})\right) + log\left(f_{\sigma_t^2}(\boldsymbol{v})\right) \tag{4}$$

In which

$f_{\boldsymbol{\beta},\phi}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{v})$: The joint probability density function for conditional observational data $\boldsymbol{u}$ and $\boldsymbol{v}$

$f_{\sigma_b^2}(\boldsymbol{u})$: Density function for spatial random effects with relative covariance form $\boldsymbol{L}$

$f_{\sigma_t^2}(\boldsymbol{v})$: The joint probability density function for temporal random effects with the form of relative covariance $\boldsymbol{\Lambda}$.

Based on (4), the h-likelihood function for this study is:

$$\boldsymbol{h} = \sum_{y=0}\left(-\frac{\mu^{2-p}}{\phi(2-p)}\right) + \sum_{i=1.y_i>0}^{m} \log\left(a(\boldsymbol{y},\phi)\exp\left(\frac{1}{\phi}\left(\frac{\boldsymbol{y}\mu^{1-p}}{(1-p)} - \frac{\mu^{2-p}}{2-p}\right)\right)\right.$$

$$-\tfrac{s}{2}\log(2\pi) - \tfrac{s}{2}\log(\sigma_b^2) - \frac{\boldsymbol{u}'\boldsymbol{u}}{2\sigma_b^2} + -\tfrac{t}{2}\log(2\pi) - \tfrac{t}{2}\log(\ \sigma_t^2) - \frac{\boldsymbol{v}'\boldsymbol{v}}{2\ \sigma_t^2} \qquad (5)$$

The estimation of fixed effect parameters $\boldsymbol{\beta}$, spatial random effect $\boldsymbol{u}$, and temporal random effect $\boldsymbol{v}$ is done by maximizing the $\boldsymbol{h}$ function in equation (5) by finding the first derivative $\boldsymbol{h}$ with respect to $\boldsymbol{\beta}, \boldsymbol{u}$ and $\boldsymbol{v}$ by completing $\frac{\partial \text{h}}{\partial \boldsymbol{\beta}} = 0$, $\frac{\partial h}{\partial \boldsymbol{u}} = 0$, and $\frac{\partial h}{\partial \boldsymbol{v}} = 0$ using the chain rule as in [22] and [23]. The three derivatives are difficult to obtain a closed-form. This will be difficult if the parameter estimation is done manually. This problem can be solved by the iteration method such as the Newton – Raphson method.

2.3. The Estimation of Variance Component Parameters for Spatial and Temporal Random Effect Estimation of Fixed and Random Effect

One of the concerns in random effects modeling is to develop better methods for estimating the component of the variance (dispersion) parameter. Estimating parameters involving random effects will involve complex integration using the likelihood concept. Meanwhile, the Bayes method requires priors in modeling and convergence is difficult to obtain. [12] developed a method of estimating parameter that is the maximum adjusted profile hierarchical likelihood estimator (MAPHLE) that is defined as follows:

$$\boldsymbol{h_A} = \left(\boldsymbol{h} - \frac{1}{2}\log\left(\det\left(\frac{\boldsymbol{H}}{2\pi}\right)\right)\right)_{\beta_0=\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}},\text{u}=\hat{\text{u}},\text{v}=\hat{\text{v}}} \qquad (6)$$

Parameter σ_b^2 and $\sigma_t^2$is estimated by maximizing the function $\boldsymbol{h_A}$ i.e. looking for the first derivative $\boldsymbol{h_A}$ with respect to σ_b^2 and $\sigma_{e_t}^2$by completing the derivative $\frac{\partial \boldsymbol{h_A}}{\partial \sigma_b^2} = 0$ dan $\frac{\partial \boldsymbol{h_A}}{\partial \sigma_t^2} = 0$. The estimated parameters σ_b^2 and $\sigma_t^2$are obtained by iteration method because it is difficult to get the closed form. The Newton-Raphson method is used to estimate σ_b^2 and σ_t^2 parameter. The variance for $\sigma_b^2\ and\ \sigma_t^2$ is the inverse of the Hessian matrix of the MAPHLE. The H matrix has the following sizes:

$$\boldsymbol{H} = \begin{pmatrix} \left(-\frac{\partial^2 h}{\partial \beta^2}\right)_{(p+1)\times(p+1)} & \left(-\frac{\partial^2 h}{\partial \beta u}\right)_{(p+1)\times s} & \left(-\frac{\partial^2 h}{\partial \beta v}\right)_{(p+1)\times t} \\ \left(-\frac{\partial^2 h}{\partial u\beta}\right)_{\boldsymbol{s\times(p+1)}} & \left(-\frac{\partial^2 h}{\partial u^2}\right)_{\boldsymbol{s\times s}} & \left(-\frac{\partial^2 h}{\partial uv}\right)_{s\times t} \\ \left(-\frac{\partial^2 h}{\partial v\beta}\right)_{s\times(p+1)} & \left(-\frac{\partial^2 h}{\partial vu}\right)_{\boldsymbol{t\times s}} & \left(-\frac{\partial^2 h}{\partial v^2}\right)_{t\times t} \end{pmatrix}_{((p+1)+s+t)\times((p+1)+s+t)}$$

2.4. The Algorithm for Spatio-Temporal GLMM (STGLMM) Models Prediction

nnn Prediction is done to see how close the actual data is to the predicted results and can also be used to predict several future time periods. The following prediction algorithm is used

1. Call the parameters $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{u}}$, $\widehat{\boldsymbol{v}}$ estimation that has been obtained.
2. Use $\widehat{\boldsymbol{v}}$ as much time as needed especially for parameters near the end of the period. For example, the time used in modeling is 348. Then the parameter $\widehat{\boldsymbol{v}}$ is a vector measuring 348 x 1. If we want to make predictions for the next 12 months period then use $\widehat{\boldsymbol{v}}$ between 337 to 348.
3. Create matrix $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ according to the number of locations and times to be predicted. For example, the locations used are 27 locations and 12 time periods, so $\boldsymbol{Z_1}$ is $n_{st} \times 27$ and $\boldsymbol{Z_2}$ is $n_{st} \times 12$. n_{st} is the number of observations multiplied between many locations and times.
4. Call the location correlation matrix in the model according to the parameter range obtained from the actual data. The location correlation matrix is created according to the number of locations to be predicted. If the location is 27 and the parameter range is 151 then the size of the correlation matrix is 27 x 27.
5. Create a time correlation matrix with autoregressive coefficients obtained from actual data. For example, if the coefficient obtained is =0.57, then the time correlation coefficient matrix is 12 × 12. Furthermore, the time correlation matrix used is $\frac{\rho_{t,t'}}{1-\rho^2}$.
6. Create matrix $\boldsymbol{L}\ and\ \boldsymbol{\Lambda}$ as the result of cholesky decomposition of spatial correlation matrix $\boldsymbol{\Sigma}_s = \sigma_b^2 \boldsymbol{LL'}$ and temporal correlation matrix $\boldsymbol{\Sigma_t} = \sigma_t^2 \boldsymbol{\Lambda\Lambda'}$ where $\boldsymbol{LL'} = \boldsymbol{\gamma}(\boldsymbol{d})$ and $\boldsymbol{\Lambda\Lambda'} = \frac{\rho_{t,t'}}{1-\rho^2}$ are cholesky decomposition matrix.
7. Create matrix $\boldsymbol{Z_1^*} = \boldsymbol{Z_1\Lambda}$ and $\boldsymbol{Z_2^*} = Z_2\boldsymbol{L}$.

8. The number of predictor variables used is the same as the number of principal components used in the modeling. The length of the predictor variable is as many as the location multiplied by the time to be estimated.
9. Predictions are calculated using the following equation:

$$\hat{\boldsymbol{y}} = \boldsymbol{\mu}_{(n_{st}\times\mathbf{1})} = \exp\left(\boldsymbol{\eta}_{(n_{st}\times\mathbf{1})}\right)$$

$$= \exp\left(\boldsymbol{X}_{(n_{st}\times(\boldsymbol{p}+\mathbf{1}))}\boldsymbol{\beta}_{((\boldsymbol{p}+\mathbf{1})\times\mathbf{1})} + \boldsymbol{Z}^{*}_{\mathbf{1}(n_{st}\times s)}\,\boldsymbol{u}_{s\times 1} + \boldsymbol{Z}^{*}_{\mathbf{2}(n_{st}\times t)}\,\boldsymbol{v}_{t\times 1}\right) \tag{7}$$

For example, the number of locations is $S = 27$ and time is $T = 12$, so the number of observations is $n_{st} = n_{27\times 12}$.

3. Application of STGLMM in Statistical Downscaling Using TCPG Response

3.1. Data Source

This study uses monthly rainfall and precipitation data from the GCM with a period of January 1981 to December 2009 as many as 348 months. Rainfall data is used as a response variable consisting of 27 observation stations located in latitude between 7.78^0 to -6.28^0 and longitude 108.40^{o} to 107.87^{o} which was obtained from the Meteorological and Geophysical Agency. GCM data as explanatory variables are obtained from The National Centers for Environmental Prediction (NCEP) in the form of a Climate Forecast System Reanalysis (CSFR) model which can be downloaded on the website https://rda.ncar.edu/. The grid domain used in this study is $2.5^0 \times 2.5^0$ with a size of 8×5 which is equivalent to 40 explanatory variables.

3.2. Rainfall Prediction using STGLMM in Statistical Downscaling

This study aims to build an GLMM model that has a spatio-temporal dependence with a Tweedie Compound Poisson-Gamma response which can be called a model STGLMM. This study uses 3 models to be compared, namely the spatio-temporal GLMM (STGLMM) method with location assumed to be exponentially correlated and time assumed to follow a first-order autoregressive process, spatial GLMM (SGLMM) with location assumed to be exponentially correlated and time assumed to be independent, Temporal GLMM (TGLMM) with location assumed to be independent and time is assumed to follow a first-order autoregressive process. The three models were compared to see which model was better and the effect of random effects on the GLMM model gradually.

Modeling using the TCPG distribution needs to consider the scale of the data used. Large-scale data need to be divided by a certain value so that modeling can be done. Large-scale data make modeling difficult, such as singular Hessian matrices, inconsistent results, and predictive results are not obtained because they are infinite. Some of the STGLMM model parameters that will be estimated include: fixed effect parameters, region and time random effects, spatial and time random effects variance components. The goodness of the model is measured by using the root mean square error prediction (RMSEP), and the correlation between the actual data and the prediction. The data exploration is shown in Figure 1.

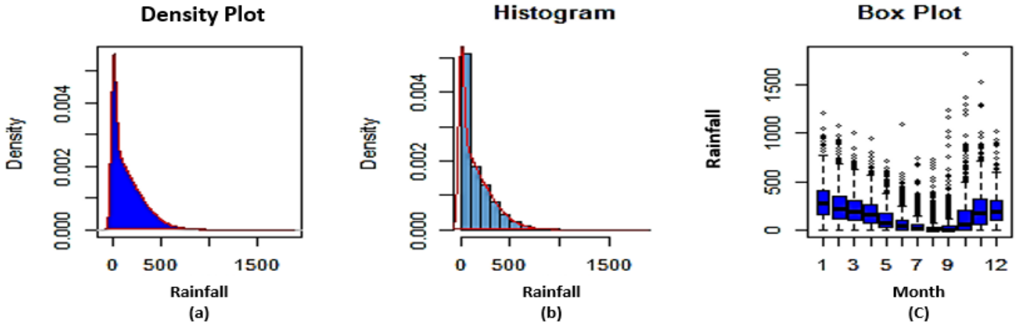

FIGURE 1. (a) Plot of rainfall density, (b) Rainfall histogram, (c) Box-plot of rainfall for all Rain stations from January-December 1981-2009

Figure 1 shows the Box-plot rainfall data for all stations from January – December 1981-2009. The box plots have the form of a monsoon pattern where the rainfall follows the pattern of the letter U. The lowest rainfall is between June and September. The histogram plot of the observation data contains a value of zero and the data has positive continuous and exact zero characteristics, so it is suspected that the data follows the TCPG distribution. The next step is to estimate the index parameter as a determinant of the TCPG distribution if the estimated index parameter has a value between $1 < p < 2$. Some of the important parameters that are suspected for modeling are listed in Table 1.

TABLE 1. The estimated parameters p, dispersion (ϕ), range (α), autoregressive coefficient (ρ) and number of principal components

Value of parameter estimate	Estimation
$\boldsymbol{p}$	1.5
$\boldsymbol{\phi}$	17
$\boldsymbol{\alpha}$	1
Range	151
Autoregresive $(\boldsymbol{\rho})$	0.58
PCA	6

Table 1 provides information that the index parameter value$p$ is around$1 < p < 2$, the dispersion parameter ϕ =17 and α =1. While the parameter of variance for the spatial component is 151 and the autoregressive parameter is 0.58. The GCM output has a problem, namely between variables that are correlated with each other so that it is handled by PCA which is determined by selecting the root of the trait with a value of more than one.

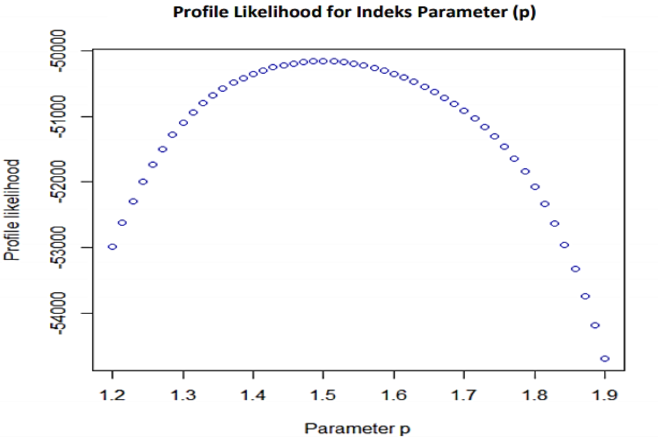

FIGURE 2. Profile Likelihood plot for the index parameter (p)

The index parameter estimates (p) are selected based on the smallest profile likelihood value. Figure 2 shows that the index parameter with the smallest profile likelihood is 1.5 which is then used in modeling. Table 2 is the parameter estimates for the fixed effects, and the variance components for STGLMM model.

TABLE 2. Fixed effect and Variance Component Estimator

STGLMM	Estimation	Confidence Interval	P-Value
$\hat{\boldsymbol{\beta}}$	0.94	(0.87 ; 0.99)	0.000
	1.14	(1.13 ; 1.15)	0.000
	0.78	(0.75 ; 0.81)	0.000
	0.92	(0.87 ; 0.96)	0.000
	0.78	(0.68 ; 0.89)	0.000
	0.66	(0.34 ; 0.67)	0.000
	0.98	(0.85 ; 1.11)	0.000
$\hat{\boldsymbol{\sigma}}_b^2$	0.18		0.000
$\hat{\boldsymbol{\sigma}}_t^2$	2.30		0.000

There are seven parameters of fixed effect obtained including intercept and five of them are significant with indicated confidence interval and p-value obtained is less than 0.05. Variance component for location random effect is 0.18 and variance component for time random effect is 2.03.

Table 3 shows that the STGLMM method has the smallest RMSEP value compared to the SGLMM and TGLMM models. The highest correlation value between observation and prediction data is owned by the TGLMM model, followed by the SGLMM and SGLMM models. This shows that the three models are equally good at modeling rainfall data with spatial and temporal dependencies. However, the SGLMM model produces a model that is able to reduce the variability due to spatial and temporal dependencies and has the smallest RMSE value compared to the SGLMM and TGLMM models. The modeling results show that the STGLMM model has a good ability to predict two components of rainfall simultaneously.

TABLE 3. Comparison of RMSEP and Correlation Methods of SGLMM, TGLMM, and SGLMM

	Goodness of Fit	
Models	RMSEP	Correlation
SGLMM	71.94	0.67
TGLMM	71.13	0.68
STGLMM	65.19	0.64

Predictions regarding the following year are shown in Figure 3. Based on the plot in Figures 3(a) to 3(i) show that the plots of the three models have almost the same and close patterns, but the STGLMM method is close to the actual data. The nine rain stations from each image have the same predictive pattern, namely the monsoon pattern. The monsoon pattern is a type of rainfall that is unimodal, namely one peak of the rainy season between December-January-February and the dry season between June-July-August.

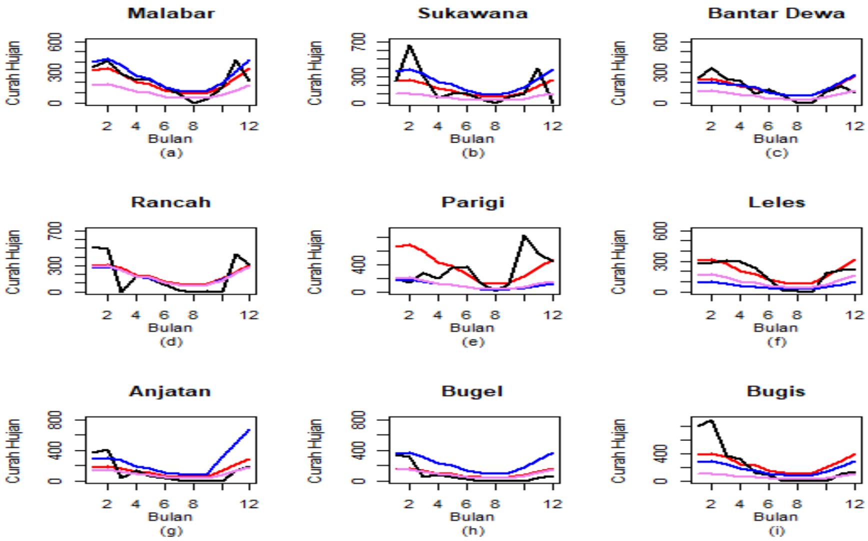

FIGURE 3. (a) – (i) rainfall prediction plot from SGLMM model (━━), TGLMM model (━━) and STGLMM model (━━) with actual data (━━) for Malabar – Bugis rain station

TCPG distribution is good to use in rainfall modeling because it can model two rain components simultaneously in one distribution. Some information about the components of rain can be obtained such as the intensity of rain, the average number of rain events per month λ, the average rainfall per incident $\alpha\gamma$, the probability of no rain events per month $\pi = \exp(-\lambda)$, many events no rain $(N\pi)$. Predicted rainfall characteristics are described in Table 4 for the Malabar rain station.

TABLE 4. Estimated Parameters $\hat{\lambda}$, $\hat{\alpha}\hat{\gamma}$, $\pi = exp(\hat{\lambda})$ for Malabar Station

Month	Actual	Prediction (μ)	λ	$\alpha\gamma$	π	$N\pi$
1	348	314	2.1	151	0.21	1.49
2	415	333	2.1	155	0.12	1.40
3	284	290	2.0	144	0.13	1.61
4	228	207	1.7	122	0.18	2.20
5	222	182	1.6	114	0.20	2.44
6	151	119	1.3	93	0.27	3.31
7	86	95	1.1	83	0.31	3.80
8	0	84	1.1	78	0.34	4.07
9	39	97	1.2	84	0.31	3.75
10	150	156	1.5	106	0.22	2.75
11	426	246	1.8	133	0.15	1.89
12	211	334	2.2	155	0.11	1.39

To make it easier to see the predicted pattern of rainfall characteristics for other regions, the plot of the estimated parameters for γ, αγ, π, Nπ of the three models for several regions can be seen in Figures 4(a) to 4(l).

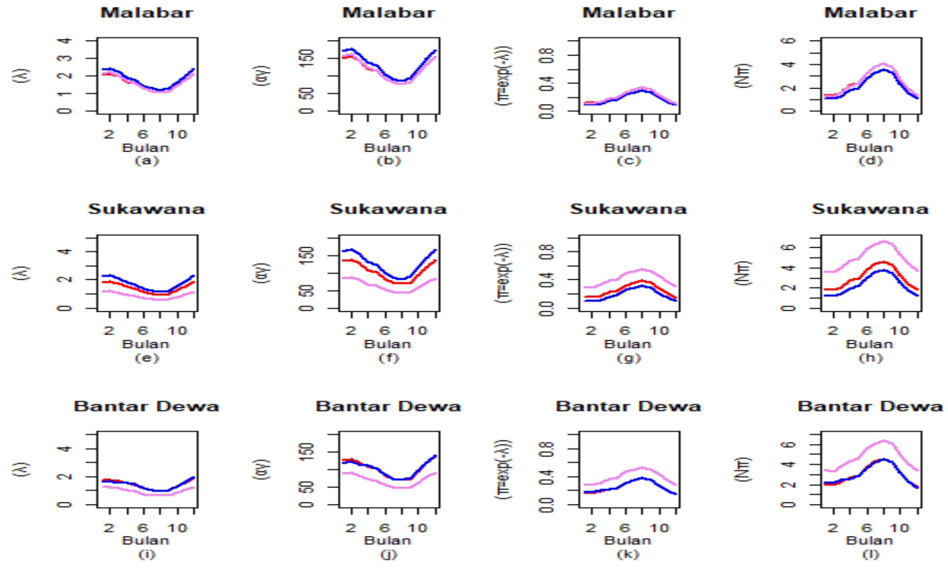

FIGURE 4. (a) – (i) are Plots of Estimated Parameters for λ,αγ,π, Nπ rainfall prediction plot from SGLMM model (▬), TGLMM model (▬) and STGLMM model (▬) with actual data (▬) for Malabar, Sukawana and Bantar Dewa rain station

Table 4 can be interpreted that the average daily rainfall events per month (λ) in January are twice, the average daily rainfall events per month (αγ) in January are 151, the probability of no rain events per month (π) for January is 0.21, the number of events without rain (Nπ) January is once.

4. CONCLUSIONS

Based on the description above, it can be concluded that The Spatio–Temporal Generalized Linear Mixed Model (STGLMM) is good for rainfall modeling which can be seen from the RMSEP value obtained, which is the smallest compared to the SGLMM and TGLMM models. The TCPG distribution is not only able to predict the intensity of rainfall but is also able to predict the number of rain events, and the probability will not rain in a certain month.

ACKNOWLEDGE

The authors acknowledge with thanks to both the referees and the editor for their fruitful suggestions and comments, which led to the overall improvement in the manuscript.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interest.

REFERENCES

[1] N.C. Dzupire, P. Ngare, L. Odongo, A Poisson-gamma model for zero inflated rainfall data, J. Probab. Stat. 2018 (2018), 1012647.

[2] F. Novkaniza, M Hayati, B Sartono, KA Notodiputro, Fused lasso ffor modeling monthly rainfall in Indramayu sub distric West Java Indonesia. IOP Conf. Ser.: Earth Environ. Sci. 187 (2018), 012046.

[3] P.K. Dunn, Occurrence and quantity of precipitation can be modelled simultaneously. Int. J. Climatol. 24 (2004), 1231–1239.

[4] W.H. Bonat, C.C. Kokonendji, Flexible Tweedie regression models for continuous data, J. Stat. Comput. Simul. 87 (2017), 2138–2152.

[5] M.M.Hasan, P.K. Dunn, A simple Poisson-gamma model for modelling rainfall occurrence and amount simultaneously, Agric. Forest Meteorol. 150(10) (2010), 1319–1330.

[6] R.N. Rachmawati, A. Djuraidah, A.H. WIgena, I.W. Mangku, Additive Bayes spatio-temporal model with INLA for West Java ainfall prediction, Procedia Computer Sci. 157 (2019), 414–419.

[7] A. Djuraidah, R.N.Rachmawati, A.H. WIgena, I.W. Mangku, Extream data analysis using spatio-temporal Bayes regression with INLA in statistical downscaling model, Int. J. Innov. Comput. Inform. Control. 7 (1) (2021), 259–273.

[8] M. Hayati, A.H. Wigena, A. Djuraidah, A. Kurnia, A new approach to statistical downscaling using Tweedie compound Poisson gamma response and lasso regularization, Commun. Math. Biol. Neurosci. 2021 (2021), Article ID 60.

[9] C.E. McCulloch, S.R. Searle, Generalized, linear, and mixed models, Wiley and Sons, New York, 2000.

[10] B.M. Bolker, M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, J.S.S White, Generalized linear mixed models: a practical guide for ecology and evolution, Trends Ecol. Evol. 24(3) (2009), 127–135.

[11] K. Lekdee, L. Ingsriawang, Generalized linear mixed model with spatial random effect for spatio-temporal data: an application to dengue fever mapping, J. Math. Stat. 9(2) (2013), 137-143

[12] Y. Lee, J.A. Nelder, Hierarchical generalized linear models, J. R. Stat. Soc.: Ser. B (Methodol.). 58 (1996), 619–656.

[13] Y. Lee, J.A. Nelder, Y. Pawitan, Generalized linear models with random effect unified analisis via h-likelihood, monograph on statistics and applied probability 106, Chapman & Hall/CRC, 2006.

[14] M. Noh, L. Wu, Y. Lee. Hierarchical likelihood methods for nonlinear and generalized linear mixed model with missing data and measurement error in covariates, J. Multivar. Anal. 109 (2012), 42-51.

[15] A. Muslim, M. Hayati, B. Sartono, K.A. Notodiputro, A combined modeling of generalized mixed model and LASSO technique for analizing monthly rainfall data, IOP Conf. Ser.: Earth Environ. Sci. 187 (2018), 012044.

[16] M. Hayati, A. Muslim, Generalized linear mixed model and lasso regularization for statistical downscaling, Enthuastic Int. J. Stat. Data Sci. 1 (1) (2021), 36-53.

[17] Y. Zhang. Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models, Stat. Comput. 23(6) (2012), 743–757.

[18] W. Qian, Y. Yang, H. Zou, Tweedie's compound Poisson model with grouped elastic net, J. Comput. Graph. Stat. 25(2) (2016), 606–625.

[19] L.R. Dietz, S. Chateterjee, Logit-normal mixed model for Indian monsoon precipitation, Nonlinear Proc. Geophys. 21 (2014), 939-953.

[20] M. Cameletti, R. Ignaccolo, S. Bande. Comparing spatio-temporal models for particulate matter in Piemonte, Environmetrics, 22 (2011), 985–996.

[21] Y. Lee, J.A. Nelder, Modelling and analysing correlated non-normal data, Stat. Model. 1 (2001), 3-16.

[22] P. McCullagh, J.A. Nelder, Generalized linear models, Second Edition, 1989.

[23] J. Jiang, On maximum hierarchical likelihood estimators, Commun. Statist.-Theory Meth. 28 (1999), 1769-1775.