ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1, Issue 1, April 2021, pp. 36-52

Generalized Linear Mixed Model and Lasso Regularization for Statistical Downscaling

Ma'rufah Hayati [a,1,*], Agus Muslim [b,c,2]

[a] Department of Statistics, University of Nahdlatul Ulama Lampung, Jl. Mataram Marga Street, Lampung, 34194, Indonesia
[b] Department of Statistics, Bogor Agricultural University, Jl. Raya Dramaga, Bogor, 16680, Indonesia
[c] Statistics of Kepulauan Riau Province, Jl. Ahmad Yani, Tanjung Pinang, 29124, Indonesia
[1] marufahhayatimt1@gmail.com *; [2] agusmuslimkayyasah@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received February 10, 2021
Revised March 29, 2021
Accepted March 29, 2021

Keywords
Statistical Downscaling
GLMM
GLM
LASSO
GLMM-Lasso

ABSTRACT

Rainfall is one of the climatic elements in the tropics which is very influential in agriculture, especially in determining the growing season. Thus, proper rainfall modeling is needed to help determine the best time to start cultivating the soil. Rainfall modeling can be done using the Statistical Downscaling (SDS) method. SDS is a statistical model in the field of climatology to analyze the relationship between large-scale and small-scale climate data. This study uses response variables as a small-scale climate data in the form of rainfall and explanatory variables as a large-scale climate data of the General Circulation Model (GCM) output in the form of precipitation. However, the application of SDS modeling is known to cause several problems, including correlated and not stationary response variables, multi-dimensional explanatory variables, multicollinearity, and spatial correlation between grids. Modeling with some of these problems will cause violations of the assumptions of independence and multicollinearity. This research aims to model the rainfall in Indramayu Regency, West Java Province using a combined regression model between the Generalized linear mixed model (GLMM) and Least Absolute Selection and Shrinkage Operator (LASSO) regulation (L_1). GLMM was used to deal with the problem of independence and Lasso Regulation (L_1) was used to deal with multicollinearity problems or the number of explanatory variables that is greater than the response variable. Several models were formed to find the best model for modeling rainfall. This research used the GLMM-Lasso model with Normal spread compared to the GLMM model with Gamma response (Gamma-GLMM). The results showed that the RMSE and R-square GLMM-Lasso models were smaller than the Gamma-GLMM models. Thus, it can be concluded that GLMM-Lasso model can be used to model statistical downscaling and solve the previously mentioned constraints.

p-ISSN 2798-253X
e-ISSN 2798-3153

1. Introduction

Located in the tropical region, Indonesia becomes one of the world’s major agricultural nations that supplies various agricultural products, including rice as one of the prime products. Rice farming in Indonesia highly depends on climatic elements, especially rainfall. Rainfall is one of the climate elements in the tropics with significant variability and thus unpredictability. This condition urges the need to predict rainfall through statistical modeling to increase rice productivity in Indonesia.

Statistical downscaling (SDS) can be classified as spatio-temporal modeling in the field of climatology because data can be collected from several locations and observed over time. SDS is a technique in climatology that uses statistical modeling to analyze the relationship between large-scale (global) data and small-scale (local) data.

This research used the local scale data of rainfall as a response variable and General Circulation Model (GCM) output in the form of precipitation as an explanatory variable. The GCM output data has several problems, including multiple-dimensional explanatory variables, spatial correlation between grids, and multicollinearity between explanatory variables [14]. This problem can be solved by several methods, such as dimensional reduction, variable selection, and shrinkage in parameter estimation. An example of the dimensional reduction method is the principal component of analysis method. An example of the variable selection method and shrinkage coefficient that is often used is the lasso method. This method has advantage in selecting variables and estimating stable parameters [5].

Rainfall data can be measured periodically based on daily, weekly, and monthly seasons, such as rainy season and dry season, or annually with repeated measurements, such as months, which will result in mutually correlated data. Generalized Linear Model (GLM) is usually used to model rainfall data, but cannot handle some violations that occur both from response and covariates data. Therefore, it is necessary to expand the GLM model that can handle some of these constraints. This can be done by inserting random effects into linear predictors called Generalized Linear Mixed Models (GLMM) (also known as random effects models). Mixed effects models have become a popular approach to periodic and group data analysis emerging in diverse fields. Several studies have been conducted on rainfall, such as [14] analyzing precipitation prediction in the Daqing Mountains using the Multivariate Regression Model, and [6] on prediction of daily rainfall using Gamma and Weibull distributions in India, which gives a good fit.

Research on SDS in Indonesia, among others, was conducted by [5] who carried out rainfall modeling using the distribution of the most compressed gamma and Pareto responses with lasso regularization. Furthermore, [3] conducted rainfall modeling using the gamma response distribution with elastic net regulation, while [6] used a Gaussian response distribution with a fused lasso penalty. Similarly, [5] conducted SDS modeling in the GLMM framework using the Gaussian response and lasso penalties.

All in all, several models that have been developed for rainfall modeling have not considered violations of the independence assumption, while at the same time addressing the multicollinearity problem. Hence, this study proposes a model that is able to deal with the problem of violation of the independence and multicollinearity assumptions simultaneously, through a combined model between GLMM and Lasso penalty (L_1) on fixed effects called GLMM-Lasso. The GLMM-Lasso application is applied to statistical downsaling modeling with the response variable in the form of monthly rainfall in the Indramayu Regency from January 1981 to 2014 and the predictor variable from the GCM output in the form of precipitation obtained from the interpolation of a combination of surface and satellite observation data in the form of a grid from GPCP (Global Precipitation Climatology Project) version 2.2. The use of Lasso in the GLMM (GLMM-Lasso) model will reduce the complexity of the model. This method can be used for data with very large dimensions (High Dimensional Data), especially models that involve a very large number of predictors/independent variables [2]

2. Literature Review

This section provides a brief overview about the exponential family, Gamma distribution, Generalized Linear Models, and their components for the benefit of building a regression model in GLM.

2.1. Exponential Family

A single random variable Y, which has a probability distribution that depends on one parameter θ has an exponential family distribution if written as follows:

$$p(y;\theta,\phi) \left[\frac{y\theta\text{-}b(\theta)}{a(\phi)}+c(y,\phi)\right] \qquad (1)$$

$p(y;\theta,\phi)$: The probability function for a discrete or continuous random variable.

$b(.)$: Canonical parameter functions

$c(.)$: Normalization parameters

$a(\phi)$: The dispersion parameter function

$\theta=g(\mu)$: Canonical parameter (location parameter). The link between the means of the data distribution with explanatory variables.

$\phi>0$: Parameters related to the range of distribution are known as dispersion parameters (scale parameters). If ϕ is known, equation (1) is a family of one parameter exponential distribution. If ϕ is not known, equation (1) is a family of two-parameter exponential distribution [8].

2.2. Generalized Linear Model (GLM)

Linear regression models usually use linearity to describe the relationship between the mean and the response variable and the set of explanatory variables with inference assuming that the distribution of responses is normal. The generalized linear model (GLM) is an extension of the standard linear regression model to deal with the distribution of non-normal responses and possible nonlinear functions for the mean. GLM is defined as the set of independent random variables $Y_1,\dots.,Y_n$ each with the distribution form of the exponential family and has the following 3 components:

1. Random component: the distribution of each Y_i has a canonical form and depends on one parameter θ_i
2. Linear Predictor: the parameter vector $\beta=(\beta_1,\beta_2,\dots..,\beta_n)^T$ and $n\times p$. Matrix Model X which contains the value of p predictor variables and n observation, linear predictor $X\beta$
3. Link Function: g is a monotonous hybrid function, namely: the link function g connects the function $E(Y_i)$ with its predictor.

$$g(\mu_i)=x_i^T\beta \qquad (2)$$

where in,

$$\mu_i=E(Y_i) \qquad (3)$$

where in,

$g(\mu_i)$: A function that connects the expected observational data with a linear predictor

$x_i^T\beta$: Linear predictor

β : Regression parameters

X : Predictor variables

2.3. LASSO (Least Absolute Shrinkage and Selection Operator)

The least absolute shrinkage and selection operator (LASSO) method was introduced by Tibshirani in 1996.This method shrinks the regression coefficient of the predictor with high correlation to error, to almost zero or exactly zero by changing the penalty in the roll regression with the L1 norm (L1 regularization)[12]. The following is the formula to give a penalty on Lasso with constraints:

$$\sum_{i=1}^{p}\left|\beta_j\right| \leq \lambda,\ \lambda \geq 0 \tag{4}$$

The value of λ is the parameter controlling the LASSO coefficient shrinkage $\lambda \geq 0$. If $|\beta_j|$ is the least squares estimator and $\lambda_0=\sum_{i=1}^{p}|\beta_i|$, the value of $\lambda<\lambda_0$ will cause the MKT solution to shrink towards zero and allow some coefficients to shrink to zero. Estimating coefficients using LASSO can be written in the lagrange equation L

$$\hat{\boldsymbol{\beta}}^{lasso}=\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{i=1}^{p}\left|\beta_j\right|\right\},\ \lambda \geq 0 \tag{5}$$

Cross validation (CV) is a method that can be used to select the controlling parameter ("λ ") in regression. One type of cross validation is k-fold (k fold).

2.4. Generalized Linear Mixed Model (GLMM)

Suppose y_{it} is the observation t in cluster i, i=1, …., n, $t=1,\ldots,T_i$ in $y_i^t=\left(y_{iT_1},\ldots,y_{iT_i}\right)$, $x_{it}^T=1,x_{it_1},\ldots,x_{it_p}$ are covariate vectors, which are related to fixed effects and $z_{it}^T=1,z_{it_1},\ldots,z_{it_p}$ are covariate vectors related to effects random. It is assumed that the conditional independent y_{it} with mean $\mu_{it}=E(b_i,x_{it},z_{it})$ and the variance var $(var(b_i)=\phi v(\mu_{it})$ where v(.) is the known variance function and ϕ is the scale parameter. The GLMM form is as follows:

$$g(\mu_{it})=x_{it}^T\beta+z_{it}^T b_i=\eta_{it}^{par}+\eta_{it}^{rand} \tag{6}$$

g is a monotone function and the link function can be derived continuously $\eta_{it}^{par}=x_{it}^T\beta$ is a linar parametric form with the parameter vector $\beta^T=\left(\beta_0,\beta_1,\ldots,\beta_p\right)$ including the intercept and $\eta_{it}^{rand}=z_{it}^T b_i$ contains a group-specific random effect where $b_i\sim N(0,Q)$ with Q covariance matrix of size qxq. Thus, the alternative form of GLMM:

$$\mu_{it}=h(\eta_{it}),\ \eta_{it}=\eta_{it}^{par}+\eta_{it}^{ran} \tag{7}$$

where $h=g^{-1}$ is the inverse of the link function.

In GLMM, it is assumed that the conditional density of y_{it} after the explanatory variable is given and the random effect b_i and is an exponential family.

$$f(x_{it},b_i)=\exp\left\{\frac{\left(y_{it}\theta_{it}-k(\theta_{it})\right)}{\phi}+c\left(y_{it},\phi\right)\right\} \tag{8}$$

where $\theta_{it}=\theta(\mu_{it})$ is a natural parameter, $k(\theta_{it})$ is a specific function depending on the type of exponential family, c(.) is a constant of log normal, and ϕ is a dispersion parameter.

One method to maximize GLMM is the penalized quasi likelihood (PQL) (Breslow and Clayton (1993), Lin and Breslow (1996), and Breslow and Lin (1995)) [2]. The covariance matrix Q (ϱ) of the random effect b_i depends on the unknown vector ϱ. In the basic penalized concept, the combined likelihood function is defined by the parameter vector of the covariance structure ϱ together with the dispersion parameter ϕ in $\gamma^T=(\phi,\varrho^T)$ and the parameter vector $\delta^T=(\beta^T,b^T)$ with the log likelihood function:

$$l(\delta,\gamma)=\sum_{i=1}^{n}\log\left(\int f\left(y_i|\delta,\gamma\right)p(b_i,\gamma)db_i\right) \tag{9}$$

where $p(b_i,\gamma)$ is density of the random effect. On this basis, Breslow and Clayton(1993) derived the following approach:

$$l^{app}(\delta,\gamma)=\sum_{i=1}^{n}\log\log\left(f(\delta,\gamma)\right)-\frac{1}{2}b^{T}Q(\varrho)^{-1}b \tag{10}$$

In the form of a penalty $b^{T}Q(\varrho)^{-1}b$

2.5. Generalized Linear Mixed Lasso Model (GLMMLASSO)

Development of the GLMMLASSO method lies in the inclusion of a penalty $\lambda\sum_{i=1}^{p}|\beta_i|$ in equation (8), so the form of the penalized likelihood of Breslow and Clayton(1993) is as follows:

$$l^{pen}(\beta,b,\gamma)=l^{pen}(\delta,\gamma)=l^{app}(\delta,\gamma)-\lambda\sum_{i=1}^{p}|\beta_i| \tag{11}$$

with $\hat{\gamma}$ obtained by optimizing the function

$$\hat{\delta}=l^{pen}(\delta,\hat{\gamma})=\left[l^{app}(\delta,\hat{\gamma})-\lambda\sum_{i=1}^{p}|\beta_i|\right] \tag{12}$$

The Penalty used in equation (11) and (12) is considered a partial penalized approach by taking into account all the parameter vectors used $\delta^{T}=\left(\beta^{T},b^{T}\right)$.

3. Generalized Linear Mixed Lasso Model for Statistical Downscaling Modelling

This research used rainfall data of Indramayu Regency from January 1980 to 2014 as a response variable and GMC output in the form of precipitation denoted by pr as the predictor variable. precipitation is taken from the interpolated combination of surface and satellite observation data in the form of a grid from GPCP (Global Precipitation Climatology Project) version 2.2 and abbreviated as GPCP, used as big scale covariate. GPCP data were obtained from the NOAA / OAR / ESRL PSD, Boulder, Colorado, USA, via its website at http://www.esrl.noaa.gov/psd/. The covariate data were taken in the 7 × 7 grid domain (49 covariates) in the coordinate system 101.25◦ - 116.25◦ EL and 13.75◦ LS - 1.25◦ NL with a grid width of 2.5◦ x 2.5◦. In this position, Indramayu Regency is located below the middle grid of the selected area (Figure 1).

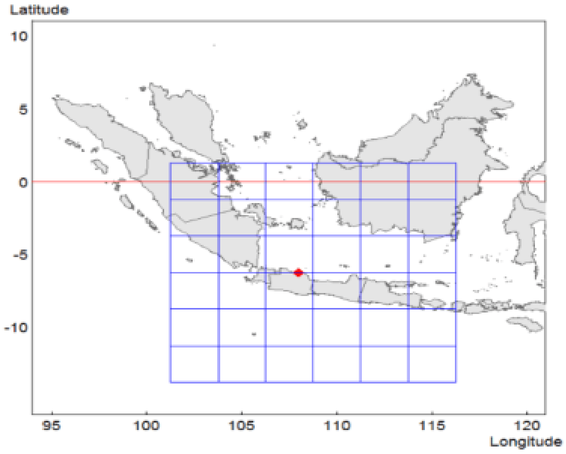


Fig. 1. Covariate Domain

Fig. 1. Shows that the Precipitation variable consists of 49 precipitation variables represented by pr11, pr12, pr13, ... pr77. where each of the precipitation variables to represent the position of precipitation in the grid row-i, and column j with i = 1,2,3,4,5,6,7 and j = 1,2,3,4,5, 6,7. The influence of precipitation on the response variable to monthly rainfall in Indramayu Regency from to April 2014 can be seen in the two SD modeling scenarios that we created, namely:

1. GLMM-Lasso with a Gaussian distributed response, the link function used is identity. Because the form of rainfall generally follows the Gamma distribution pattern without involving a value of 0, but sometimes rainfall does not occur, the Gaussian distribution is considered so that the value of 0 and continuous positive can still be modeled. Rainfall data is time series data which is generally not stationary, so it is necessary to compare it with standardized data modeling.

Generalized Linear Mixed Model Lasso (GLMM-Lasso) in this study uses two cases, namely with standardized and non-standardized responses. Each case has 2 models tested, namely the random effects of seasons and months, different random effects are considered to see better rainfall patterns when viewed based on seasonal or monthly random effects.

2. GLMM models with random effects of months and seasons on different models. The distribution used is Gamma.

The two model scenarios above will be compared to see the best model. the best model is selected by selecting the smallest RMSE value and the largest R-square

Table 1. Research Variable

Variable	Name of Variable	Information
Response	Rainfall | Seasonal(M=m) Rainfall | Monthly(B=b)	Consist of 347 rainfall observations with a continuous measurement scale
Covariate	Precipitation P_i i=1,2,….,49	Consist of 49 covariates (precipitation) came from pr11,pr12, p17,pr21,....pr77
Random effect 1	Seasonal M_j j=1,2	Consist of 2 season levels 1=Summer (April-August) 2=Rainy (September-March)
Random effect 2	Monthly B_k k=1,2,…..,12	Consist of 12 month levels, namely January- December

Note : Precipitation can be abbreviated as “pr”

4. Results and Discussion

GLMM is applied to rainfall data divided into 2 cases, namely the rainfall response variable without standardization and the variable rainfall response with standardization. In each case there are two models for the model with seasonal and month random effects. The purpose of this study is to determine the precipitation that has an influence on rainfall in Indramayu district.

GLMMLasso Model in the case of 1 and 2 (M=m) ~ $N(\mu,\sigma^2)$

Random effect of seasonal

$$\eta_{ij}=\mu+P_i+M_j \qquad (13)$$

Random effect of monthly

$$\eta_{ij}=\mu+P_i+B_k \qquad (14)$$

with :

$M\sim N(0,\sigma^2_M)$, $B\sim N(0,\sigma^2_B)$

Where in, P_i is precipitation , M_j is seasonal random effect, B_k is monthly random effect.

4.1. Statistical Downscaling Modelling with GLMM-Lasso with several cases

4.1.1.Case 1 GLMM-Lasso Model Rainfall Response variable without Standardization

The first case is divided into two models to be compared, namely a model with seasonal random effect and month random effect. In each model, analysis was carried out with different lambda values, namely 30,500,1372,2000, 5000 and 10000. So that:

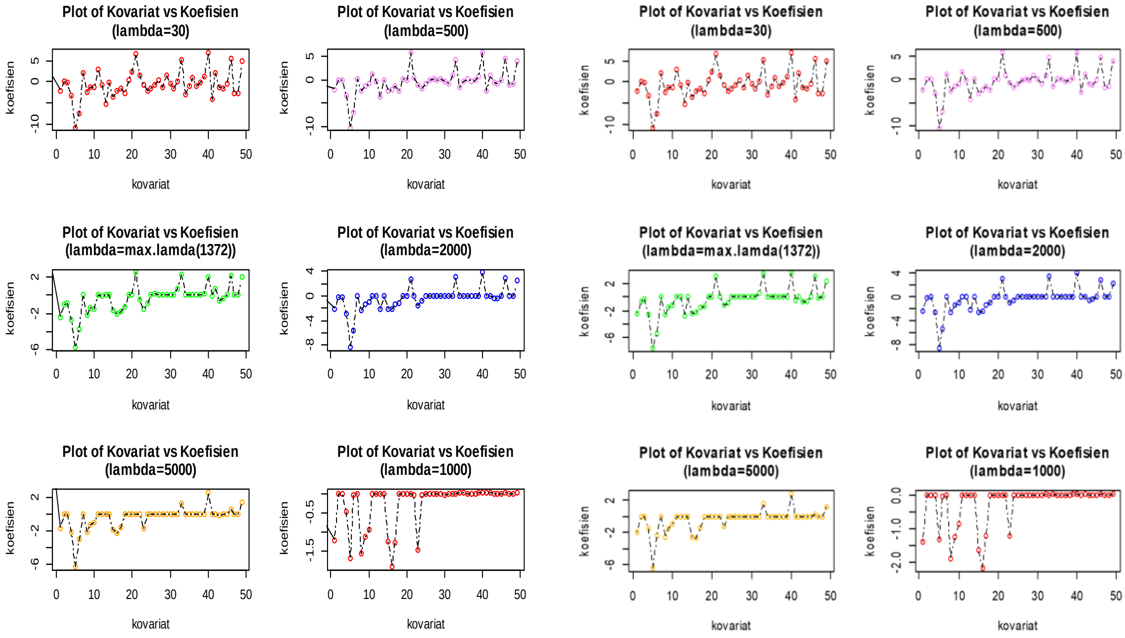
Table 2. RMSE dan R Square of Rainfall data

λ	RMSE |R Square Y_i|seasonal_j	RMSE |R Square Y_i|monthly_j
30	124.796 (42.493)	124.796(42.490)
500	129.233(38.332)	126.781 (40.649)

1372	135.097 (32.608)	137.308 (30.383)
2000	133.261(34.427)	132.404 (35.268)
5000	137.939(29.743)	135.826 (31.878)
10000	150.752(16.084)	150.933 (15.882)

In Table 2, it can be seen that the smallest RMSE and the largest R Square value for both models occurs when the lambda is 30, namely 124.796 and 42.49%, respectively. This value means that all the covariates in models 1 and 2 can explain 42.49% of the diversity of monthly rainfall in Indramayu Regency collectively.

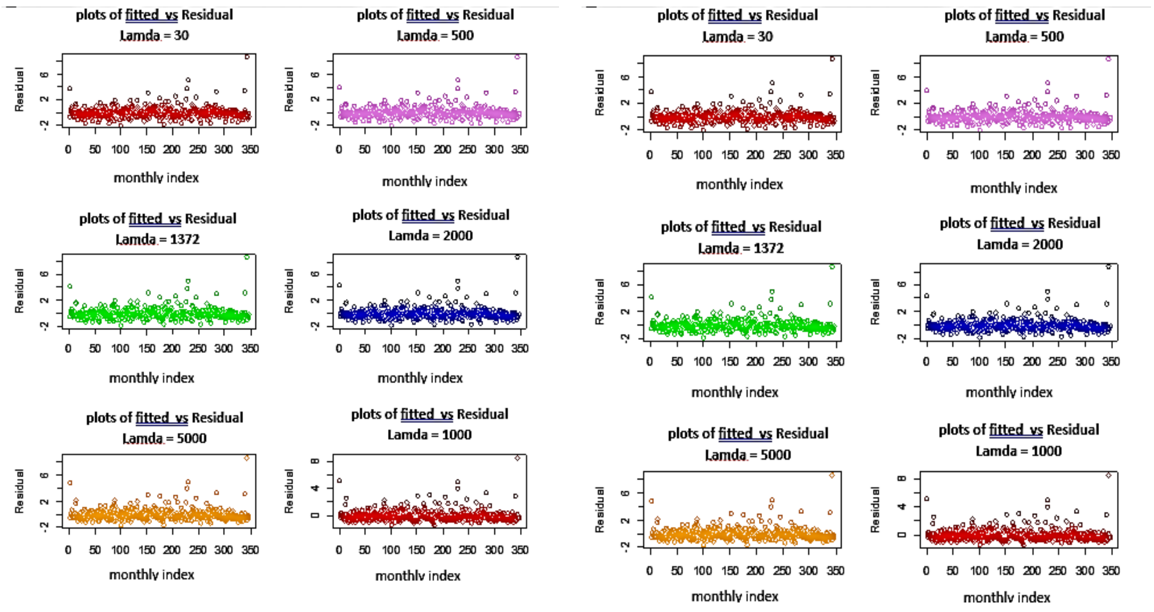
The effect of adding the constraint l_1 to the fixed effect parameter to the likelihood function equation in each model can be seen in the estimation results of the regression coefficient which shrinks to zero. The results of the estimated regression coefficient parameters are presented in Figure 1. Fig. 1 shows that there is no significant difference between the two models for the several values of λ used. The greater the value of λ, the smaller the regression coefficient to zero and vice versa. The lambda value specified is λ = 10000, which is the largest λ value, while λ = 30 is the smallest λ value.



(a) (b)

Fig. 2. Plot of Covariate Coefficient of Rainfall Data for Model Seasonal (a) and Monthly Random Effect (b)

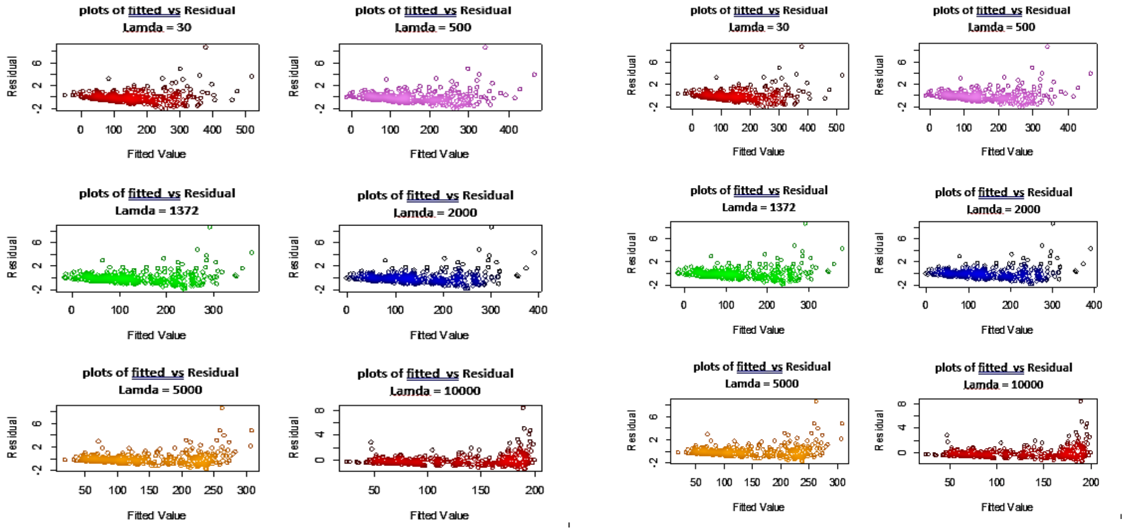
To check the error of the model whether it meets the assumptions, independent error, homoscedasticity or constant variance, the normal distribution error is presented in Fig 2, 3 and 4. We checked the assumption whether the constant variance error was met or not, and the results is shown in Fig. 2, namely the plot between the residual values versus the lunar index of model 1.



(a) (b)

Fig. 3. Plot of Residual Vs Monthly Index of Rainfall Data with Seasonal Random Effect (a) and Monthly Random effect (b)

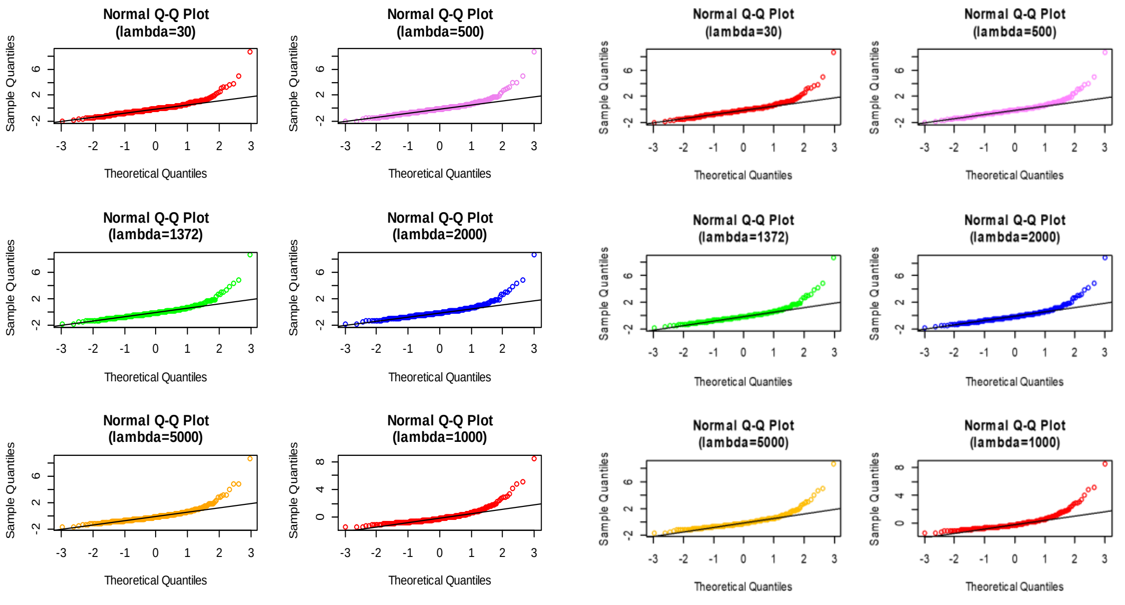
Fig 3 shows that the residuals spread around zero and have a random pattern even though there are very high and small residual values. This result indicates that the independent error assumption can be satisfied by the residual GLMM-Lasso. In addition, because the error is also assumed to be independent and have a normal distribution, we checked whether these assumptions were met by constructing a residual versus monthly index plot and a normal Q-Q plot of the residuals of models 1 and 2.



(a) (b)

Fig. 4. Estimating Y plot Vs Residual Data for Model with Seasonal Random Effect (a) and Monthly Random Effect (b)

Based on Fig 4, there is no significant difference between the two methodsIn other words, there is no certain trend or pattern so that the assumption of constant error variance can be fulfilled by the two models. In Fig 5, it can be seen that the assumption of normally distributed errors is not fulfilled, especially for residual values above the normal line.



(a) (b)

Fig. 5.Q-Q Norm Rainfall Data for Model with Seasonal Random Effect (a) and Monthly Random Effect (b)

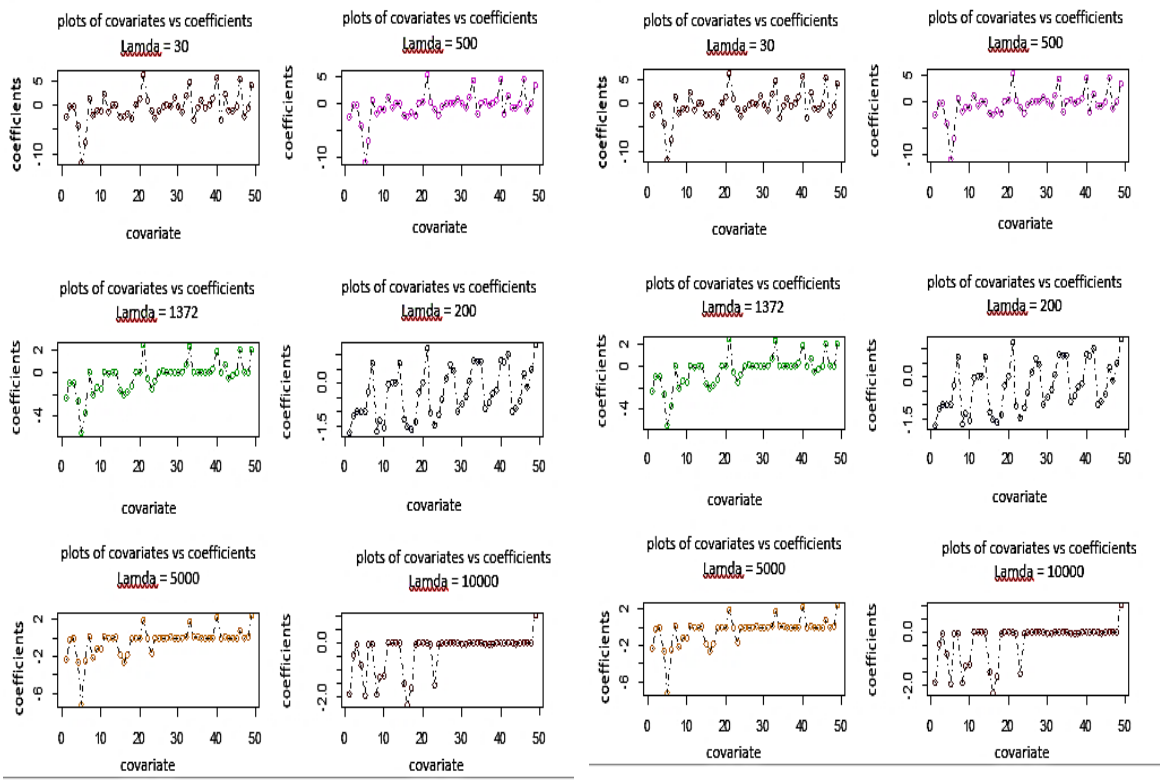
4.1.2. GLMMLasso Model with Standardized Rainfall Response Variable

Case 2 is different from case 1, since the latter is without standardization of the response variable. Thus, in case 2, a standardized response variable was performed. There are 2 models to be compared, namely a model with seasonal random effect and moon random effect. For each model, analysis was carried out with different lambda values, namely 30,500,1372,2000, 5000 and 1000.

Table 3. RMSE and Rsquare of Rainfall data

λ	**RMSE |Rsquare**	**RMSE |Rsquare**
	$Y_i \mid$seasonal$_j$	**$Y_i \mid$monthly$_k$**
30	121.3515 (44.61005)	119.992 (45.8442)
500	122.7617 (43.31526)	121.6191 44.36549)
1670	137.2613 (29.13411)	130.5941 (35.85127)
200	182.6518 (25.48407)	189.0333 (34.40557)
5000	129.2655 (37.1499)	129.4643 (36.95642)
10000	140.8905 (25.33727)	140.8905 (25.33727)

In Table 3, we can see that the RMSE value of the two models has a value similar to the 30 lambda, and the smallest RMSE values are 121.3515 and 119.992. For the R Squared value of the two models, it is found that the lambda 30, the R Squared value of the two models, is the greatest value of all lambda, namely 44.61005% and 45.8442%. This result means that all covariates in models 1 and 2 can explain the 44.61005% and 45.8442% variations of the monthly rainfall in Indramayu Regency. Thus, from the two models, it is found that model 2, the model with the moon random effect, has the smallest RMSE and the largest Rsquare.



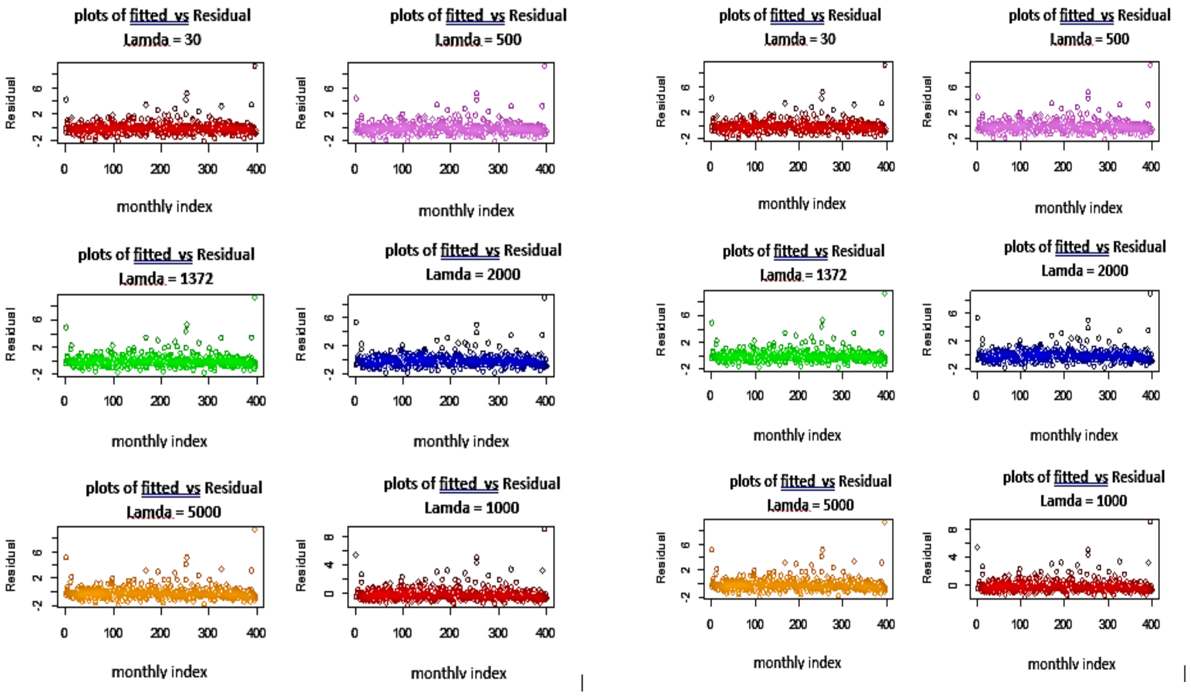
(a) (b)

Fig. 6. Plot of Covariate Vs Coefficient of Rainfall Data with Seasonal Random Effect (a) and Monthly Random Effect (a)

The likelihood function in each model is with the expectation that most of the regression coefficient shrinks to the value of Zero. Thus, it can be seen in Figure 1 that there is no significant difference between the two models for some differences in the value of λ. The greater the value of λ, the more shrinkage the regression coefficient to a value of 0 and vice versa with value of λ = 10000 as the largest specified value of λ and λ = 30 as the smallest lambda value. In contrast to case 1, in case 2, the value of λ = 2000 was not able to make the regression coefficient shrinkage.

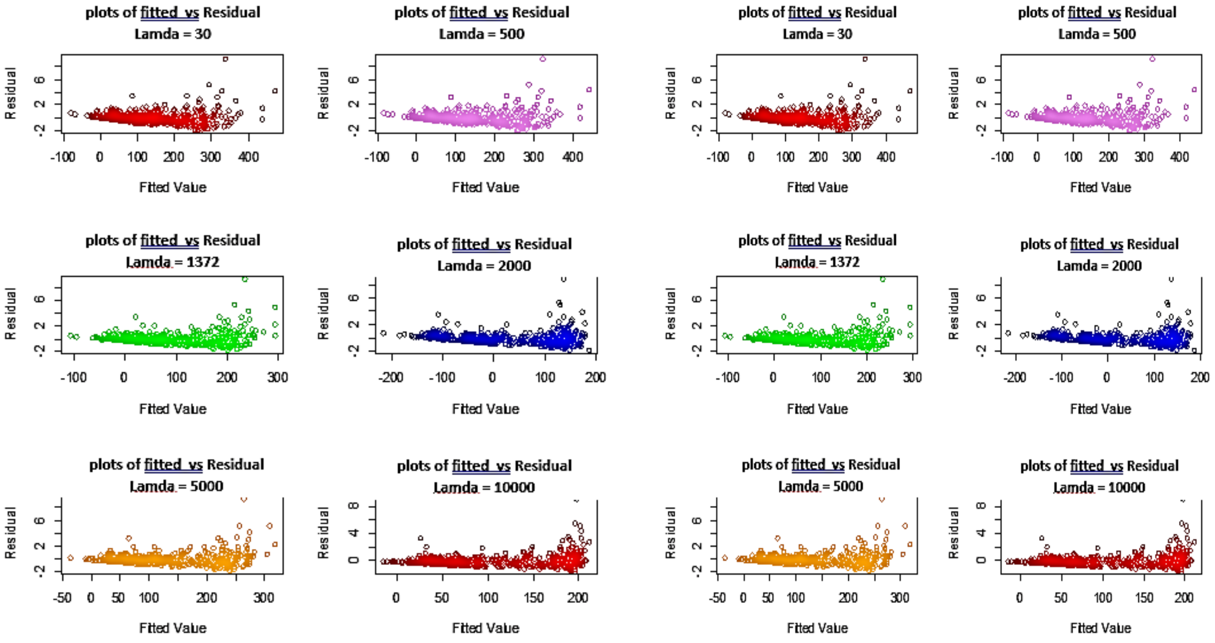
To check the error of the model, whether it meets the assumptions, independent error, homoscedasticity or constant variance, the normal distribution error is depicted in Figures 6, 7 and 8. We checked the assumption whether the constant variance error is met or not, and the result is presented in Figure 2, namely the estimation of plot between the y values versus residuals of models 1 and 2. In Figure 2, the residuals spread around 0 and already have a random pattern even though there are very high and small residual values. This result indicates that the independent error assumption can be satisfied by the residuals by both models. In addition, because the error is also assumed to be independent and have a normal distribution, we checked whether these assumptions are met, by constructing a residual versus monthly index plot and a normal Q-Q plot of the residuals of models 1 and 2 in Fig 7 and 8.

Based on Fig 7, the two methods do not have a significant difference. In other words, there is no certain trend or pattern so that the assumption of constant error variance can be met by the two models. In Figure 8, it can be seen that the normal distribution error assumption is not fulfilled, especially for residual values above the normal line.



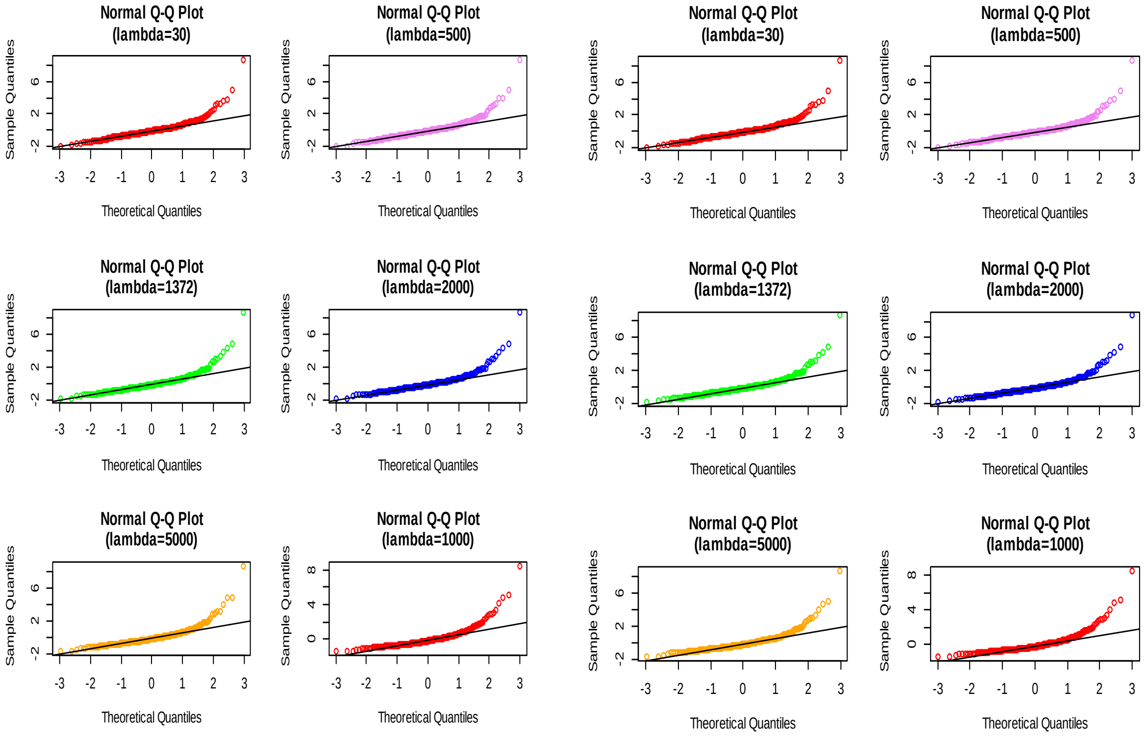
(a) (b)

Fig. 7. Residual Vs Montly Index of Rainfall Data for The model with Seasonal Random Effect (a) and Monthly Random Effect (b)



(a) (b)

Fig. 8. Estimate Y Plot Vs Residual Rainfall Data for The model with Seasonal Random Effect (a) and Monthly Random Effect (b)

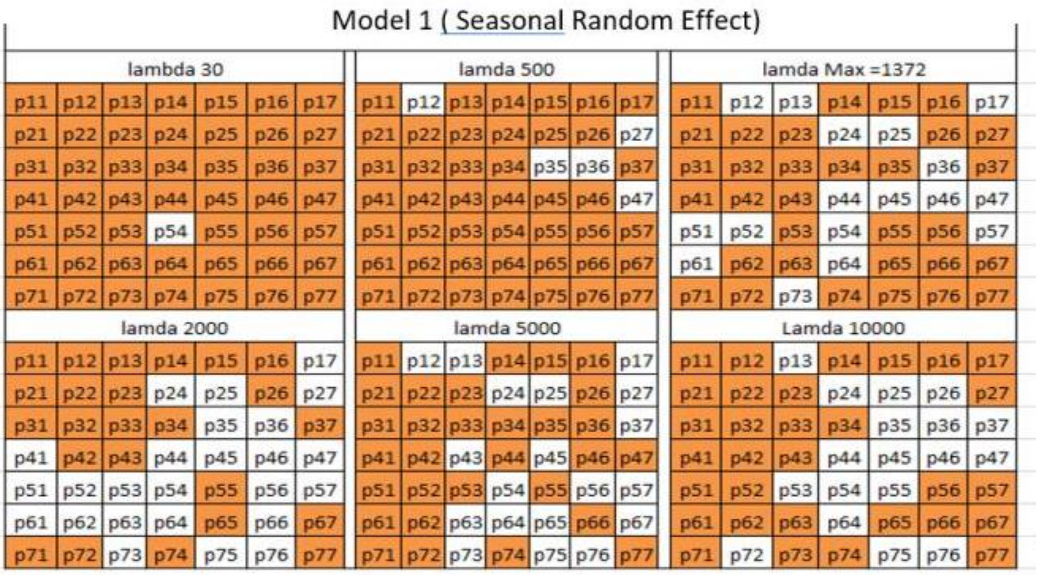


(a) (b)

Fig. 9.Q-Q Norm of Rainfall Data for Seasonal and Monthly Random Effect

3.1. Model-Based Precipitation Grid Map

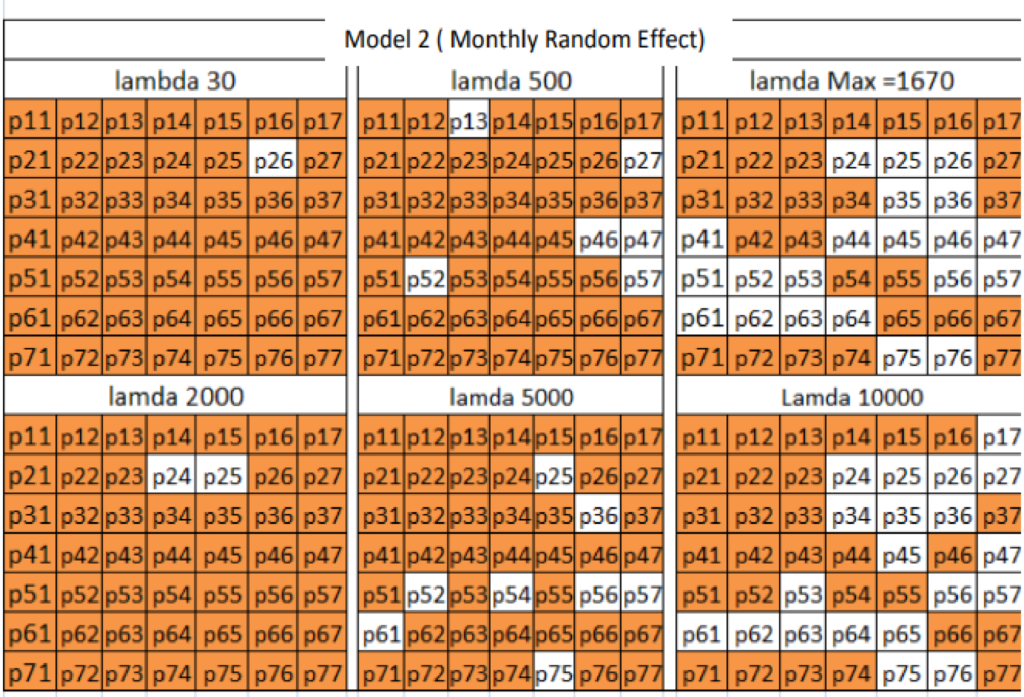
The following is a map of the allowance grid based on models 1 and 2 to see the performance of models with different lambda values.



Model 1 (Seasonal Random Effect)

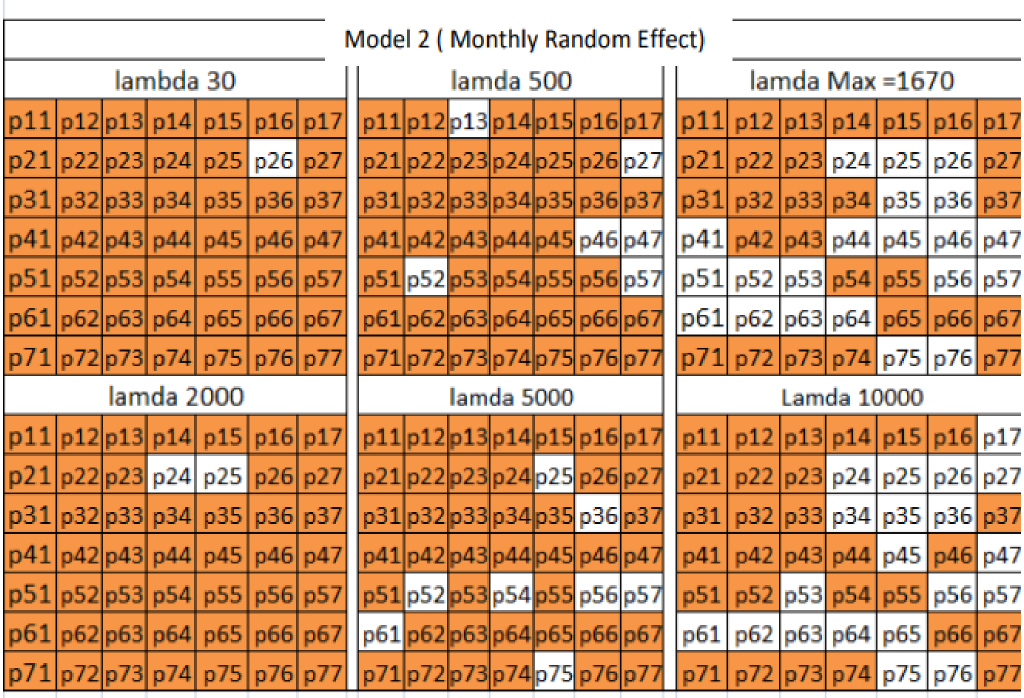
lambda 30							lamda 500							lamda Max =1372						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77
lamda 2000							lamda 5000							Lamda 10000						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77

Fig. 10. Grid Map for Seasonal Random Effect of Model 1



Model 2 (Monthly Random Effect)

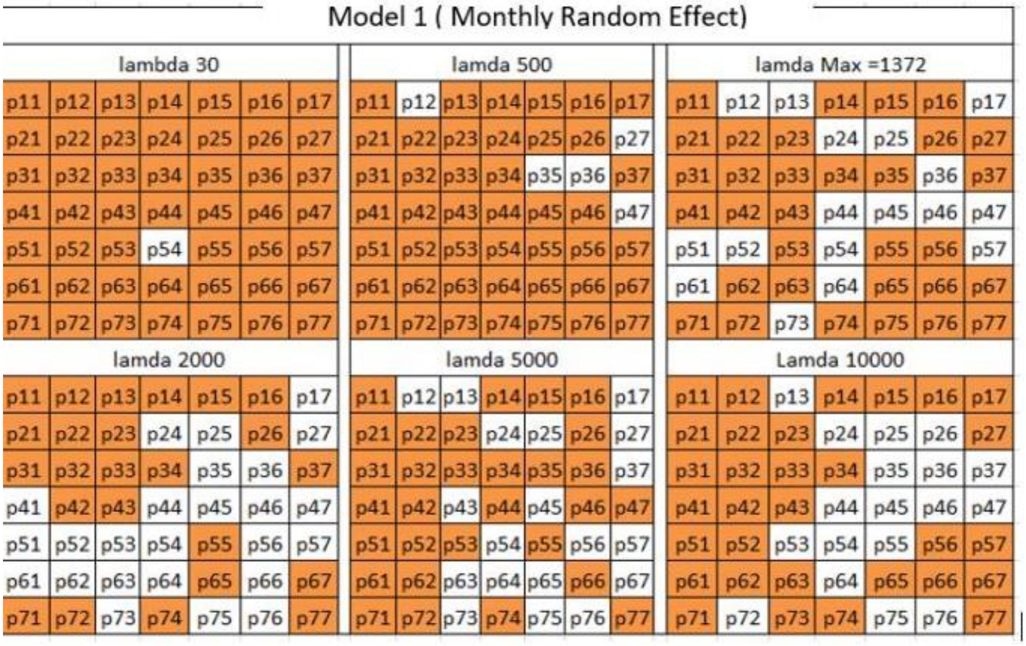
lambda 30							lamda 500							lamda Max =1670						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77
lamda 2000							lamda 5000							Lamda 10000						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77



Model 2 (Monthly Random Effect)

lambda 30							lamda 500							lamda Max =1670						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77
lamda 2000							lamda 5000							Lamda 10000						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77

Fig. 11. Grid Map for Monthly Random Effect of Model 2



Model 1 (Monthly Random Effect)

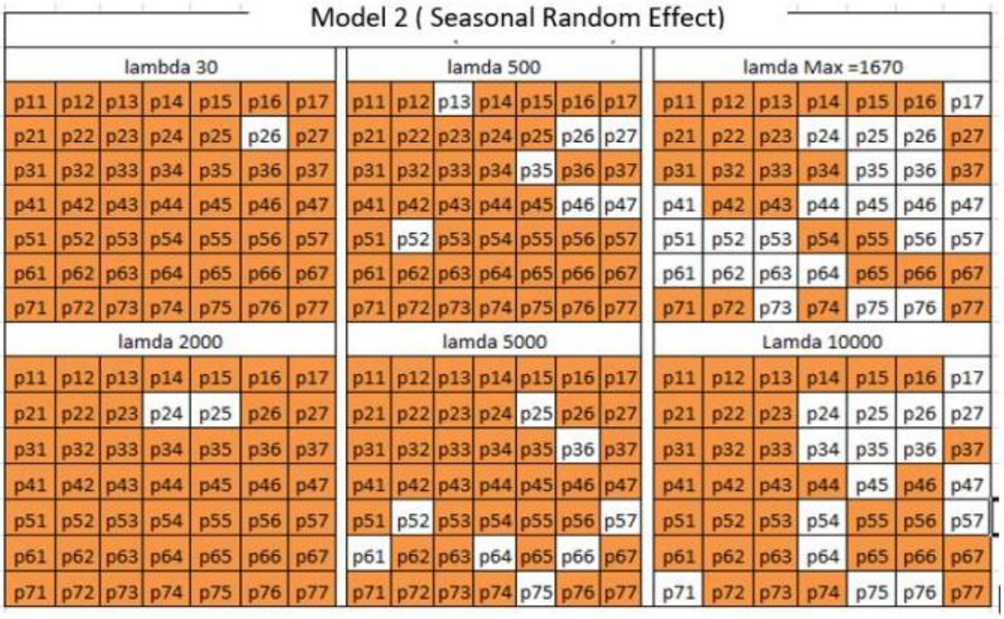
lambda 30							lamda 500							lamda Max =1372						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77
lamda 2000							lamda 5000							Lamda 10000						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77

Information:

🟧 Active

□ Non active

Fig. 12. Grid Map for Monthly Random Effect of Model 1



Model 2 (Seasonal Random Effect)

lambda 30							lamda 500							lamda Max =1670						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77
lamda 2000							lamda 5000							Lamda 10000						
p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17	p11	p12	p13	p14	p15	p16	p17
p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27	p21	p22	p23	p24	p25	p26	p27
p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37	p31	p32	p33	p34	p35	p36	p37
p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47	p41	p42	p43	p44	p45	p46	p47
p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57	p51	p52	p53	p54	p55	p56	p57
p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67	p61	p62	p63	p64	p65	p66	p67
p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77	p71	p72	p73	p74	p75	p76	p77

Fig. 13. Grid Map for Seasonal Random Effect of Model 2

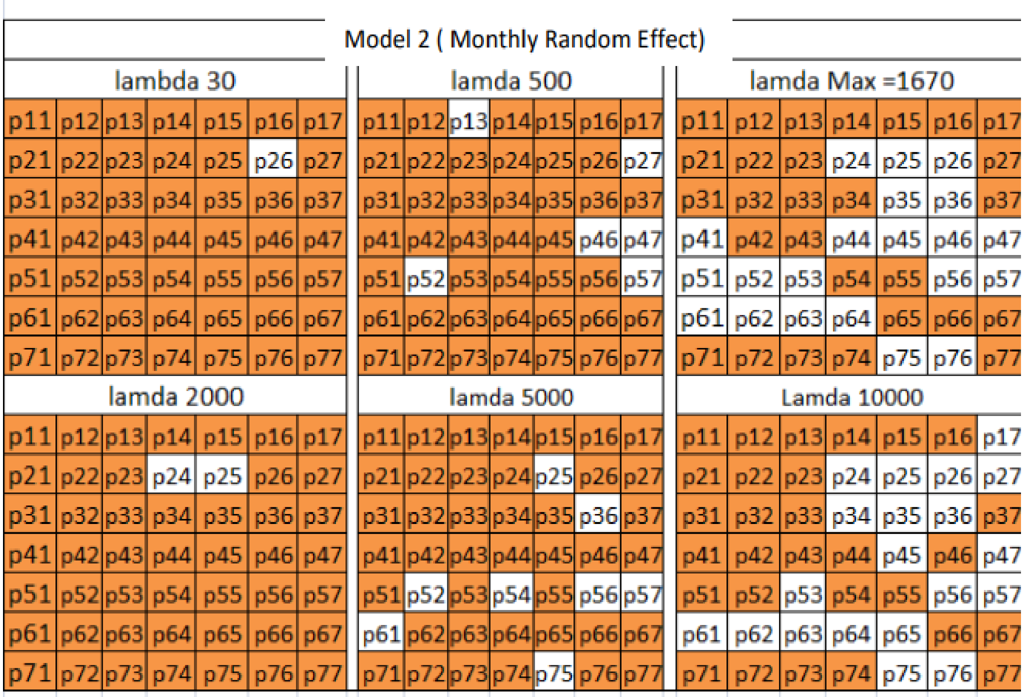

Model 2 (Monthly Random Effect)
lambda 30
lamda 500
lamda Max =1670
lamda 2000
lamda 5000
Lamda 10000
p11 p12 p13 p14 p15 p16 p17
p21 p22 p23 p24 p25 p26 p27
p31 p32 p33 p34 p35 p36 p37
p41 p42 p43 p44 p45 p46 p47
p51 p52 p53 p54 p55 p56 p57
p61 p62 p63 p64 p65 p66 p67
p71 p72 p73 p74 p75 p76 p77

Fig. 14. Grid Map for Monthly Random Effect of Model 2

From Table 4, and 5, it can be seen that the greater the lambda value used, the more likely the regression parameter will to shrink to zero. The white mark of the number of grid maps indicates that the regression coefficient tends to be 0 or close to 0. For table 6, and 7, lambda can make the shrinkage regression coefficient to a value of zero or close to that when the maximum lambda used is 1670.

Table 4. Percentage of Active Regression Coeficients

Model 1 (Seasonal Random effect) (%)						
Covariate	30	500	1372	2000	5000	10000
Non-active	2.04	6.12	36.73	46.9	26.53	48.98
Active	97.96	93.88	63.27	53.1	73.47	51.02
Model 1 (Monthly Random effect) (%)						
Covariate	30	500	1372	2000	5000	10000
Active	2.04	10.2	34.69	51	38.78	36.735
Non-active	97.96	89.8	65.31	49	61.22	63.265
Model 2 (Seasonal Random effect) (%)						
Covariate	30	500	1670	2000	5000	10000
Active	97.96	85.71	53.06	95.9	83.67	67.347
Non-active	2.041	14.29	46.94	4.08	16.33	32.653
Model 2 (Monthly Random effect) (%)						
Covariate	30	500	1670	2000	5000	10000
Active	97.96	87.76	57.14	95.9	83.67	59.184
Non-active	2.041	12.24	42.86	4.08	16.33	40.816

Table 4 shows that the percentage of active regression coefficients denotes that the coefficient is valued other than zero, while non-active means that the regression coefficient is zero or close to this value. Each data analysis using different lambda Model 1 on different random effects also indicates an increase. If the lambda value is used, the regression coefficient tends to make the regression coefficient zero or close to 0. Of the two models, when the maximum lambda is used, the inactive regression coefficient looks a lot, almost proportional to the value of lambda = 10000.

Next, we ummarized the output results for the smallest RMSE and the largest R-square, then compared the result to the GLMM Gamma response model.

Table 5. Comparison of RMSE and Rsquare Across Models

Model	GLMM-Lasso Model1 (Seasonal/monthly)	GLMM-Lasso Model2 (Monthly)	Gamma-GLMM (Seasonal)	Gamma-GLMM (Monthly)
RMSE	124.796	119.992	133.113	130.450
Rsquare	42.492	45.844	34.572	37.164

Table 5 is a summary of the RMSE and R Square of all models. For models 1 and 2, we used a regression model with the smallest RMSE value and the largest Rsquare. From the four models, it can be seen that the GLMMLASSO model is better than the usual GLMM. This is because GLMM is less stable in handling regression models with too large value of covariates. When it is applied using GLMM, there will be a warning that the data has a covariate variable of > 12.

5. Conclusion

From the explanation and model application in the previous sub chapters, it can be concluded that GLMMLasso models generally perform better than regular GLMM models. The GLMMLasso model can be used for Statistical Downscaling Modelling since it can overcome some of the constraints faced in the application, such as correlated response variables and violations of independence assumption.

References

[1] Gad A M, Kholy R S, 2012, Generalized Linear Mixed Models for Longitudinal Data, International Journal of Probability and Statistics 2012, 1(3): 67-73 DOI:10.5923/j.ijps. 20120103.03, Cairo.

[2] Groll A, Tutz, 2012. Variable selection for generalized linear mixed models by L1-penalized estimation. Stat Comput DOI 10.1007/s11222-012-9359-z, Springer Science+Business Media New York 2012.

[3] Jaiswal, R.K., Lohani, A.K., Tiwari, H.L., 2015. Statistical analysis for change detection and trend assessment in climatological parameters. Environ. Proc. 2 (4), 729–749.

[4] Krishnamoorthy K. 2006. Handbook of Statistical Distributions with Applications. New York (USA): Chapman Hall/CRC.

[5] Muslim A, Hayati M, Sartono B, Notodiputro KA. 2018. A Combined Modeling of Generalized Mixed Model and LASSO Technique for Analyzing Monthly Rainfall Data, Iop Conference Series: Earth and Environmental Science.

[6] Novkaniza f , Hayati M, Sartono B, Notodiputro KA. 2018. Fused Lasso for Modeling Monthly Rainfall In Indramayu Sub Distric West Java Indonesia, Iop Conference Series: Earth and Environmental Science

[7] Permatasari SM, Dzuraidah A, Soleh AM. 2016. Statistical Downscaling with Gamma Distribution and Elastic Net Regularization (Case Study: Monthly Rainfall 1981-2013 at Indramayu). Proceeding of The 2nd International Conference on Applied Statistics 2016 ISSN: 2579-4361. Indonesia

[8] Ranhao S, Baiping Z, Jing T. 2008. A Multivariate Regression Model for Predicting Precipitation in the Daqing Mountains. Mountain Research Development. 23(3): 318-325.

[9] Sholeh A M. 2015. Pemodelan Linier Sebaran Gamma dan Pareto Terampat dengan Regularisasi L1 pada Statistical Downscaling untuk Pendugaan Curah Hujan Bulanan

[10] Stephenson DB, Kumar KR, Doblas-Reyes FJ, Royer JF, Chauvin E, Pezzulli S. 1999. Extreme Daily Rainfall Events and Their Impact on Ensemble Forecast of the Indian Monsoon. Monthly Weather Review 127:1954-1966.

[11] Stroup W W, 2012, Generalized Linear Mixed Model modern Concepts, Methods and Applications, 1st Edition, CRC Press

[12] Tibshirani R. 1996. Regression Shringkage and Selection via the Lasso. R. Stat. Soc. Ser. B. 58(1) 267–288. doi:10.1111/j.1467-9868.2011.00771.x

[13] Tibshirani, R., Saunders, M., Rosset S, Zhu, J., Knight, K . 2005. Sparsity and Smoothness via The Fused LASSO, Journal Royal Statistical Soc. B, 67, Part 1, pp 91-108.

[14] Wigena AH. 2006. Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Peramalan Curah Hujan Bulanan. [Disertasi]. Bogor (ID): Institut Pertanian Bogor.