<MULT2>

Statistical Downscaling with Gamma Distribution and Elastic Net Regularization

(*Case Study : Monthly Rainfall 1981-2013 at Indramayu*)

Sri Maulidia Permatasari[1,a) , Anik Djuraidah[1,b), Agus M Soleh[1, c)

[1]*Department of Statistics, Faculty of Mathematics and Natural Science,*
Bogor Agricultural University, Indonesia

a)srimaulidia@gmail.com
b)anikdjuraidah@gmail.com
c)agusms@apps.ipb.ac.id

Abstract. Rainfall data are more than or equal zero and can be represented using Gamma distribution. In statistical downscaling the local scale rainfall data are used as the response variable to develop functional relation with global scale precipitation data of Global Circulation Model (GCM) output as the predictor variables. Generally, GCM output are multicollinear and regularization method can solve the problem. This paper develops a statistical downscaling model with the response of Gamma distribution using ridge regularizations and elastic net regularizations. Data are monthly rainfall in Indramayu at 1981-2013 and monthly precipitation data of GCM output in 1981-2013. The result shows that the elastic net (standard deviation of RMSEP value is 22.7 mm/month and standard deviation of correlation between actual and predicted value is 0.20 in four years) is more consistent than ridge regularization (standard deviation of RMSEP value is 35.7 mm/month and standard deviation of correlation between actual dan predict rainfall is 0.22 in four years) in predicting a next year rainfall.

Keywords: Gamma Distribution, Statistical Downscaling, Global Circulation Model, Ridge, Elastic Net

INTRODUCTION

Rainfall is one of natural occurrence which has non-negative value. In statistics, this occurrence is regarded by random variables with range of value ≥ 0. One of the distribution that represents this occurrence is Gamma distribution. According to Das 1955, the gamma distribution was postulated for rainfall (precipitation) because precipitation occurs when water particles can form around dust of sufficient mass and the waiting time for such accumulation of dust is similar to the waiting time aspect implicit in the gamma distribution [1]. Stephenson *et al* in 1999 [2] also used Gamma and Weibull distribution to predict the wet-day rainfall in India which provide good fits.

Topography and the complex interaction between sea, land, and atmosphere make the rainfall prediction in Indonesia so difficult, so that rainfall prediction model that accurate in local scale by considering about global atmosphere circulation information is needed. Statistical downscaling can be used to model and predict rainfall by using the local scale rainfall data as response variable and global scale precipitation data of Global Circulation Model (GCM) output as predictor variables. In statistical downscaling model, rainfall data and GCM output should have a high correlation to explain local climate variability well [3].

The characteristic of GCM output are curse of dimensionality and multicollinearity in each grid [4] It can cause parameter estimation value becomes unstable, so that GCM output can not be used directly before the problem is solved. Principal component analysis (PCA), lasso regularization (L_1), and ridge regularization (L_2) are some of method that usually used to resolve multicollinearity. Beside that, there is elastic net regularization that combine L_1 and L_2 method.

Soleh *et al* [5] used PCA and lasso regularization to resolve multicollinearity in GCM output for statistical downscaling model with Gamma distribution for rainfall prediction in Indramayu. Lasso regularization method predicted the rainfall in Indramayu better than PCA method. This paper explains statistical downscaling model with Gamma distribution by using ridge regularization and elastic net regularization for multicollinearity problem in GCM output, then compare both of these methods.

LITERATURE REVIEW

General Circulation Model and Statistical Downscaling

General circulation model (GCM) is an important tool in the study of climate variability and climate change [6]. This model describe the subsystems from climate on earth, such as the process in atmosphere, sea, land, and simulate the climatic conditions in global scale. GCM simulate global climate variables on each grid (the size is $\pm 2.5^0$ or ± 300 km^2) in every atmospheric layers. However GCM can't give an important information with higher resolution, such as temperatures and rainfall in local scale, but GCM is still possible to be used to get the local scale information by using downscaling method [4].

Downscaling method is a data transformation process from a grid with large scale unit into data on a grid with a smaller scale units. One of downscaling methods is statistical downscaling, which data on large scale grid in a certain period that used as base to determine the data on smaller scale grid. The equation for this method is

$$\mathbf{Y}_{(n\times 1)} = f\left(\mathbf{X}_{(n\times p)}\right) \tag{1}$$

with $\mathbf{Y}_{(n\times 1)}$ is a vector of local climate variable/ response variable, $\mathbf{X}_{(n\times p)}$ is a matrix of GCM output/ predictor variables, n is the number of observation, and p is the number of grid in GCM output.

Generalized Linear Model

A linear model for response variable that comes from an exponential family distribution and has a link function which relates expectation value and systematic component from linear model is known as Generalized Linear Model (GLM) [7]. Gamma distribution is part of exponential family. Probability density function of Gamma distribution with two parameters (v, ξ) for response variable (y) on $(0, \infty)$ area is:

$$f_Y(y; v, \xi) = \frac{v^{\xi}}{\Gamma(\xi)} y^{\xi-1} exp(-vy)\,, \quad y > 0 \tag{2}$$

with v is rate parameter and ξ is shape parameter. Probability density function of Gamma distribution in the form of an exponential family with $\theta = \frac{1}{\mu}$ and $\phi = \frac{1}{\xi}$ is:

$$f_Y(y; \theta, \phi) = \exp\left\{-\xi\left(y\left(\frac{1}{\mu}\right) - \log\left(\frac{1}{\mu}\right)\right) + \xi\log(\xi y) - \log(y) - \log\left(\Gamma(\xi)\right)\right\} \tag{3}$$

The relation of μ with parameter inside of Gamma distribution is $\mu = \frac{\xi}{v}$, with parameter ξ is assumed to be constant [4]. Parameter β_j in systematic component is used to estimate the parameters μ, based on used link function. The canonical link function for the GLM with Gamma distribution is the reciprocal $\frac{1}{\mu}$, so:

$$\mu = \frac{1}{\sum_{i=1}^{p} \beta_i x_i} \tag{4}$$

Parameter estimation of GLM uses maximum likelihood method, which obtained by Iterated Re-Weighted Least Squares (IRWLS).

Regularization

Ridge Regularization (L_2)

Hoerl and Kennard [8] introduced one of the methods to resolve multicollinearity by using ridge regression in 1970. Ridge regression add a penalty in the regression coefficients in L $_2$ norm, that is parameter estimation $\boldsymbol{\beta}$ by minimizing the sum of squared errors of linear regression (least squares method) with constraints $\sum_{j=1}^{p} \beta_j^2 \leq k$. In GLM, parameter estimation solution with ridge regularization is:

$$\widehat{\boldsymbol{\beta}}_{ridge} = argmin_{\beta}\{-log[L(\mathbf{y};\boldsymbol{\beta})]/n + \lambda_{ridge} \textstyle\sum_{j=1}^{p} \beta_j^2\} \quad (5)$$

with $L(\mathbf{y};\boldsymbol{\beta})$ is the likelihood function of exponential family distributions, $\lambda \geq 0$ is control parameter. Estimation of ridge regression coefficients will be depreciated along with the value of λ. The bigger of λ, the more shrinkage of regression coefficient towards zero. Estimation of coefficient in ridge regression is unequivariant because of there is different scale on input data, so that it needs to standardize the variables [9].

LASSO Regularization (L_1)

Robert Tibshirani introduced least absolute shrinkage and selection operator (LASSO) for the first time in 1996. LASSO is also used to resolve multicollinearity and do variable selection of correlated variables automatically. In GLM, parameter estimation solution with lasso regularization is

$$\widehat{\boldsymbol{\beta}}_{lasso} = argmin_{\beta}\{-log[L(\mathbf{y};\boldsymbol{\beta})]/n + \lambda_{lasso} \textstyle\sum_{j=1}^{p} |\beta_j|\} \quad (6)$$

with $\sum_{j=1}^{p} |\beta_j| \leq k$ is penalty on $\boldsymbol{\beta}$, $L(\mathbf{y};\boldsymbol{\beta})$ is the likelihood function of exponential family distributions, $\lambda \geq 0$ is control parameter. This penalty cause the non-linear equations in y, so that to obtain the solution of coefficient estimation have to use quadratic programming [10].

Elastic-Net Regularization

According to Zou and Hastie [11] LASSO has some limitation, there are three scenarios:

1. In the $p > n$ case, the LASSO selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method.
2. If there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to select only one variable from the group and does not care which one is selected.
3. For usual $n > p$ situations, if there exist high correlations among predictors, it has been empirically observed that the prediction performance of the LASSO is dominated by ridge regression (Tibshirani 1996).

Hui Zou and Trevor Hastie (2005) introduced elastic-net regularization (EN) that can solve those problems. Elastic-net regularization is a combination of L $_1$ and L$_2$ and depend on penalty, $(1-\alpha)\sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \leq k$. If $\alpha = 1$ then EN becomes penalty on LASSO and if $\alpha = 0$ then EN becomes penalty on ridge regularization. In EN regularization, there are shrinkage of coefficient simultaneously from correlated predictor variables and selection variables. Parameter estimation solution in GLM with elastic-net regularization is:

$$\widehat{\boldsymbol{\beta}}_{EN} = argmin_{\beta}\{-log[L(\mathbf{y};\boldsymbol{\beta})]/n + \lambda_{EN} \textstyle\sum_{j=1}^{p} [(1-\alpha)\beta_j^2 + \alpha|\beta_j|]\} \quad (7)$$

DATA AND METHODS

Data

This study uses two kinds of data, that is monthly local rainfall (precipitation) data in Indramayu regency as response variable (Y) and data of GCM output as predictor variables (X) from 1981 to 2013 (33 years). The monthly local rainfall data is in ZOM 79 that consist of four rainfall stations in Indramayu (Krangkeng, Sukadana, Karangkendal, and Gegesik). Data of GCM output is monthly precipitation data of Coupled Model Intercomparison Project Phase 5 (CMIP5) from http://www.climatexp.knmi.nl/ with coordinate: 18.75^0 LS – 1.25^0 LS and 101.25^0 BT – 118.75^0 BT. Size of each of grids is 2.5^0. In this area there is 8x8 grids data of GCM output, so that there are 64 predictor variables.

Methods

The following steps of data analysis in this study are:

1. Exploration of monthly rainfall data in Indramayu with descriptive statistics and check data distribution, check correlation of each grid on GCM output data, and calculate the correlation between rainfall data and GCM output data.
2. Determine the time lag of GCM output data to monthly rainfall data by using cross correlation function (CCF) with formula:

   $$r_{xy}(l) = \frac{c_{xy}(l)}{S_x S_y} \qquad (7)$$

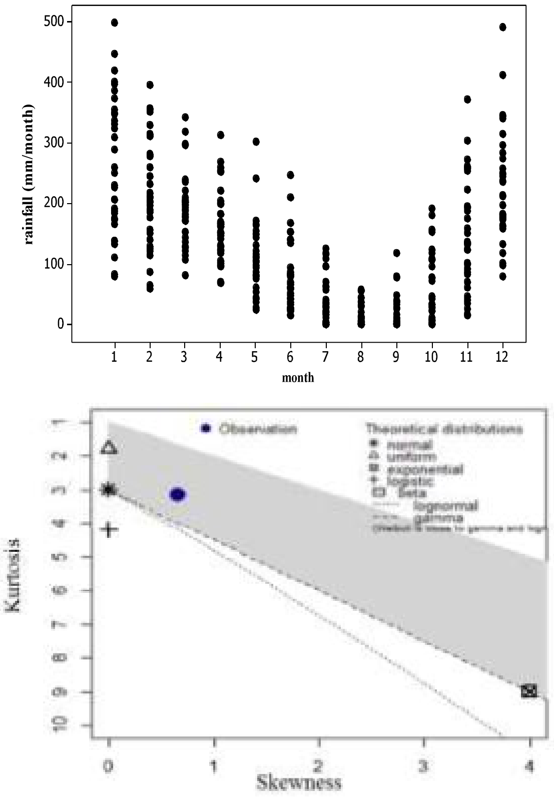
   with $c_{xy}(l)$ is covariance of x and y, S_x is standard deviation of x, and S_y is standard deviation of y. In this case, time lag is in the interval of -12 to 12. Thus data of lag GCM ouput can have a strong correlation with monthly rainfall data.
3. Data is divided into two parts, data from 1981-2012 as training data to make model and data in the year of 2013 as testing data for validation.
4. Analyze statistical downscaling by using training data set by assuming distributed Gamma response variable with penalty L$_2$ and elastic net. The following steps are :
   a. Select the optimum value of the penalty coefficient (λ) by using cross validation (CV). In CV, data modelling are subdivided into two groups, they are training data and testing data for randomly validation. Geisser (1975) generally describe that CV is taking the average of some predictor that generated from the retained data [12]. The general form of a CV is k-fold CV, the data is divided into k sections with equal size. First modelling is done by retaining the first part of data for validation and (k - 1) the next part as training data. Then, mean square error (MSE) is obtained, that calculated from validation data set. The second modelling is retaining the second part of data as validation and other (k - 1) as training data. Then, second MSE is obtained, that also calculated from validation data set. And so on until the k-data is retained for validation and MSE of k are obtained. Then all values of MSE averaged and the average result is a value of cross validation error (CVE) for a value of the regularization parameter (λ). CV Procedure is used for another λ value. The best parameter (optimum) in the model is the regularization parameter (λ) which gives the smallest value of CVE. This paper use k =5.
   b. Modelling the rainfall data with time lag GCM output data by using the value of optimum λ.
5. After getting model for each penalty L$_2$ and elastic net, calculate the forecast of rainfall in Indramayu regency at 2013 by using prediction data.
6. Measuring the goodness of fit model by using root mean square error of prediction (RMSEP), with formula $RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ in each regularization (L$_2$ and EN). The best model has the smallest RMSEP value.
7. Checking consistency model by calculating the prediction value in 2010, 2011, and 2012.

RESULT AND DISCUSSION

Data Exploration

Data of monthly rainfall at Indramayu in 1981-2013 have monsoon pattern or U-shaped (Figure 1(a)), with the average of rainfall is 127.19 mm/month and standard deviation is 107.47 mm/month which means that diversity of rainfall is quite high. The highest intensity of rainfall occurs in January with 498 mm/month and the lowest intensity is 0 mm/month. The average of the lowest rainfall is in August with 14.50 mm/month. The highest standard deviation also occurs in January, that is 112.92 mm/month which indicates rainfall in January is quite diverse. Based on Cullen and Frey graph (Figure 1 (b)), monthly rainfall data in Indramayu regency can spread by distribution Lognormal, Weibull and Gamma. Based on Table 1, Gamma distribution has smallest AIC and BIC value, so that Gamma distribution approach can be used to modelling the monthly rainfall in Indramayu Regency.

Precipitation data of GCM output in different grids have many pairs of variable with high correlation (more than 0.7), that is about 84.28%. This indicates the occurrence of multicollinearity among GCM grid output. The correlation between monthly rainfall data and GCM output data in each grid that more than 0.7 is about 9.38%.



(a) (b)

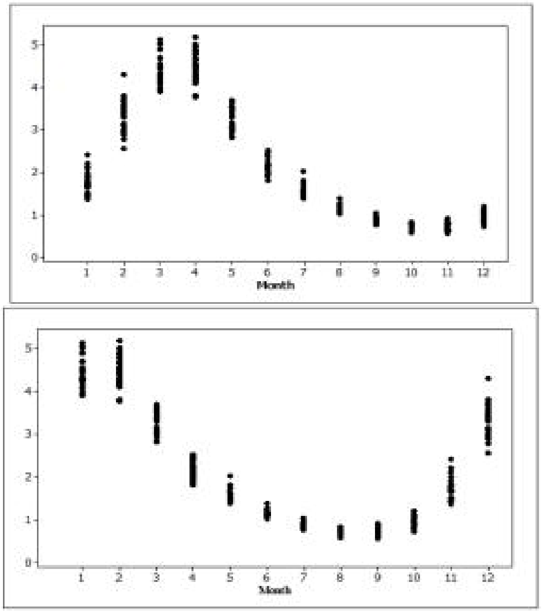
FIGURE 1. (a) Monthly rainfall pattern at Indramayu 1981-2013, (b) Cullen and Frey graph

TABLE 1. The comparison of RMSEP value and correlation

Distribution	AIC	BIC
Lognormal	4794.18	4802.14
Weibull	4636.05	4644.02
Gamma	4625.44	4633.41

Time Lag of GCM output

Data of GCM output as predictor variables (X) should have a high correlation with rainfall data as response variable (Y) in order to get the better estimate value. Time lag of GCM output data determined from the highest cross correlation value between GCM output and rainfall data, which is calculated by CCF. If the *k*-time lag value is positive, then predictor variables move backwards as much as *k* months. Otherwise, if the *k*-time lag value is negative, then predictor variables move towards as much as *k* months. For example, the precipitation of X_1 variable has the highest cross correlation value in the second time lag, so that precipitation data of X_1 variable is postponed as 2 months/ move 2 months backwards. Thus the correlation between precipitation data of X_1 variable and rainfall data increased from 0.30 to 0.72. After doing time lag process in GCM output, the correlation between each grid of GCM output and rainfall data that more than 0.7, increased to 70.31%. Thus, moving based on time lag in GCM output data has followed the pattern of rainfall.



(b) (b)

FIGURE 2. (a) Precipitation pattern of X_1, (b) Precipitation pattern of X_1 after time lag

Monthly Rainfall Estimation Model and Prediction

Data from 1981-2012 are used to do cross validation process with 5-folds to select the optimum control parameter value (λ) and also to make model. Statistical downscaling model with gamma distribution by using ridge regularization to resolve multicollinearity, obtain the optimum control parameter value (λ) of 817.23 in cross validation process. Modelling by using optimum λ value reveals model RMSEP value of 76.36 mm/month. Furthermore statistical downscaling model with gamma distribution by using elastic net regularization to resolve multicollinearity, obtain the optimum control parameter value (λ) by smallest CVE value is of 81.72 with $\alpha = 0.1$. Modelling by using optimum λ value reveals model RMSEP value of 73.11 mm/month.

Data 2013 as testing data are used for validating the model, which data of GCM output are used to predict monthly rainfall in 2013. Comparison of monthly rainfall prediction in 2013 at Indramayu by using ridge regularization and elastic net are shown in Figure 3.

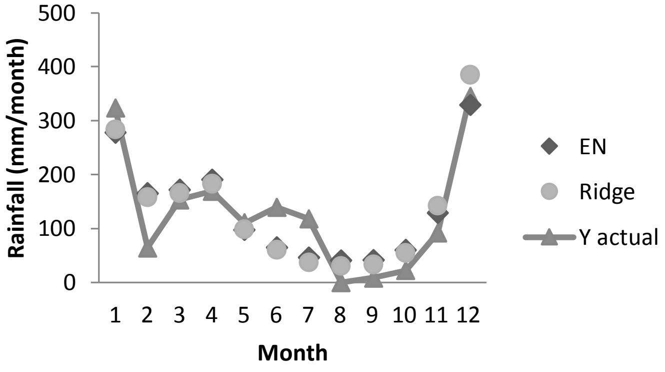
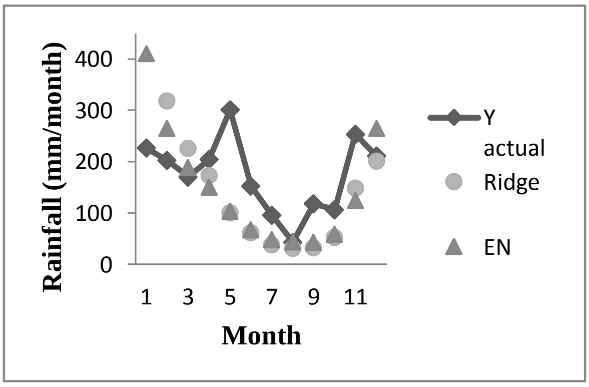


FIGURE 3. Monthly rainfall prediction in 2013

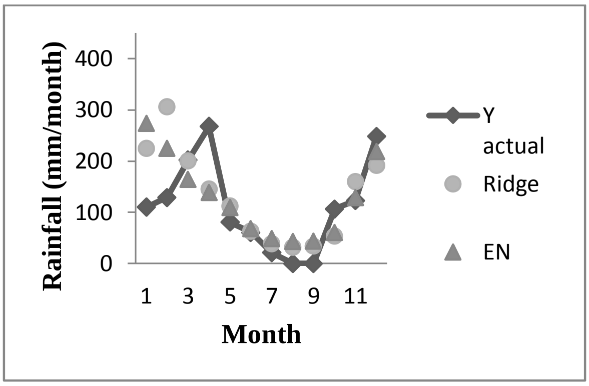
The plot in Figure 3 shows that there are a few of estimated value nearing the actual value as in March, April, May, November, and December. In January of 2013, rainfall in Indramayu included into the extreme category because of the intensity of more than 200 mm / month [13]. But in February, rainfall dropped dramatically and it characteristic included into below normal category. Then rainfall in June and July of 2013 had a slightly different pattern. It had higher intensity than in May, so the difference of rainfall estimate is high enough. The estimated value of rainfall is obtained by using ridge regularization method is not much different from the elastic net regularization method, with each validation RMSEP value is 50.37 mm / month and 49.82 mm/month, and each correlation value is 0.89.

Selection of The Best Model

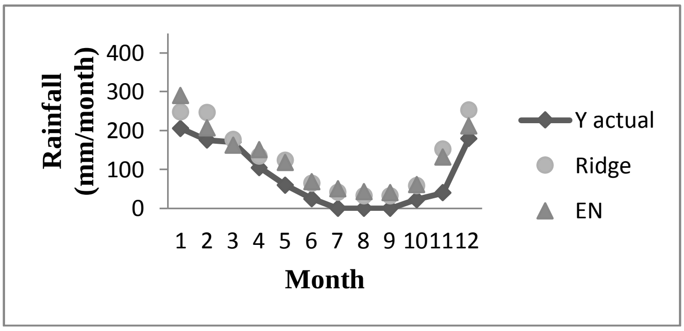
Ridge regularization generates RMSEP value is 50.37 mm/month and elastic net regularization method generates RMSEP value is 49.82 mm / month, which means the RMSEP value with elastic net regularization method is smaller than RMSEP value of ridge regularization method, with the correlation value is same. Statistical downscaling model will give good results if the correlation between the response variable and predictor variables is not changed by the change of time and keep same even though there is climate change, or a model of statistical downscaling stay consistent in it's prediction at different times [4], so as to see the consistency, it can be made by building a model for prediction of a year rainfall in 2010 (the training of data for modeling between 1981 to 2009), in 2011 the training of data for modeling between 1981 to 2010), and 2012 (the training of data for modeling between 1981 to 2011).



(a)



(b)



(c)

FIGURE 4 Monthly rainfall prediction in (a) 2010, (b) 2011, (c) 2012

Figure 4 show that monthly rainfall in 2010 has different pattern that intensity is relatively low in January and February, as well as with the rainfall pattern in 2011. While rainfall in 2012 has a pattern like U-shaped, but the rainfall intensity is also relatively low in January and February. Based on Table 2, the RMSEP value of elastic net are lower than ridge regularization and the correlation value between actual and predicted value of elastic net are also higher than ridge regularization in every years. Standard deviation of obtained RMSEP and correlation value by elastic net regularization is 22.65 mm / month and 0.20, it less than the standard deviation of obtained RMSEP and correlation value by ridge regularization (35.7 mm/month and 0.22).

TABLE 2. The comparison of RMSEP value and correlation

	Ridge Regularization		**Elastic Net**	
	RMSEP	**correlation**	**RMSEP**	**correlation**
Prediction in 2010	128.48	0.49	98.75	0.52
Prediction in 2012	55.36	0.95	52.26	0.95
Prediction in 2011	76.61	0.62	71.98	0.63
Prediction in 2013	50.37	0.89	49.82	0.89
Mean	77.70	0.74	68.20	0.75
Standard deviation	35.71	0.22	22.65	0.20

CONCLUSION

This research shows that statistical downscaling model in rainfall prediction with gamma distribution responses valriable and using ridge and elastic net regularization method in solving the problem of GCM data, generates RMSEP average value is quite small and the average correlation is quite large. After testing the consistency of model from year to year, the standard deviation value of RMSEP and the correlation between actual and predicted value that are generated by elastic net regularization method are smaller than ride regularization, so the model of statistical downscaling by elastic net regularization method is more consistent than ridge regularization in predicting next year rainfall.

REFERENCES

1. Krishnamoorthy K. 2006. *Handbook of Statistical Distributions with Applications.* New York (USA): Chapman&Hall/CRC.
2. 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.
3. Busuioc A, Chen D, Hellstrom C. 2001. Performance of Statistical Downscalling Models in GCM Validation and Regional Climate Change Estimates: Application For Swedish Precipitation. *Int J Climatol*. 21:557-578
4. Wigena AH. 2006. Pemodelan *Statistical Downscaling* Dengan Regresi *Projection Pursuit* Untuk Peramalan Curah Hujan Bulanan: Kasus Curah Hujan Bulanan Di Indramayu [Disertasi]. Bogor (Id): Institut Pertanian Bogor.
5. Soleh AM, Wigena AH, Djuraidah A, Saefuddin A. 2015. Statistical Downscaling to Predict Monthly Rainfall Using Linear Regression with L_1 Regularization (LASSO). Hikari Journal of Applied Mathematical Sciences, Vol. 9, No. 108:5361-5369.
6. Zorita E, Storch HV. 1999. The Analog Method as a Simple Statistical Downscalling Technique: Comparison with More Complicated Methods. Journal of Climate Vol. 12: 2474-2489.
7. McCullagh P, Nelder JA. 1989. *Generalized Linier Models.* Ed ke-2. London (UK):Chapman & Hall/CRC.
8. Hoerl AE, Kennard RW. 1970. Ridge Regression. Biased Estimation for Nonorthogonal Problems. Technometrics, Vol. 12, No. 1: 55-67.
9. Hastie T, Tibshirani R, Friedman J. 2008. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. *Springer*: Stanford, California.
10. Tibshirani R. 1996. Regression Shringkage and Selection Via The Lasso. *Journal of The Royal Statistical Society. Series B (Methodological),* Vol 58, Issue 1:267-288.
11. Zou H, Hastie T. 2005. Regularization and Variable Selection Via the Elastic-Net. *J.R.Statist. Soc. B* 67, Part 2:301-320.
12. Arlot S, Celisse A. 2010. A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys* Vol. 4: 40-79.
13. [BMKG] Badan Meteorologi dan Geofisika Stasiun Klimatologi Klas 1 Dramaga Bogor. 2013. Buletin Analisis Hujan Bulan Februari 2013 dan Prakiraan Hujan Bulan April, Mei, dan Juni 2013. [downloaded at June 2016[14]. Available on: http://www.depok.go.id/berkas-unggah/2013/05/Prak-Jabar-Juni-2013-B.pdf