Continuum Regression Modeling with LASSO to Estimate Rainfall

Arwini Arisandi[*], Aji Hamim Wigena[*], Agus Mohamad Soleh[*]

[*] Department of Statistics, IPB University

DOI: 10.29322/IJSRP.10.10.2020.p10651
http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10651

Abstract- Statistical downscaling (SDS) is a method to relate functionally global scale to local scale climate data. The global scale data are from the Global Climate Models (GCM) output while the local data are from a rainfall station. Generally, the GCM output data are available in the form of contiguous grids which commonly causes the multicollinearity problem. The problem can be overcome by a method such as principal component analysis (PCA), LASSO, forward selection. A SDS modeling can use the continuum regression with PCA. The research aims to develop SDS model using LASSO in the continuum regression to predict the local scale rainfall. The SDS model development uses the monthly GCM precipitation data at 9×9 grids as predictor variables and the local rainfall data as the response variable from January 2011 to December 2019 at West Java province. The model evaluation is based on the values of RMSEP and correlation. The results showed that the continuum regression with LASSO was better than the LASSO regression and the regression with forward selection in the case of rainfall prediction.

Index Terms- statistical downscaling, continuum regression, LASSO, GCM

I. Introduction

Indonesia is located in a tropical region with high rainfall intensity, especially in the highlands. The frequency of rainfall can cause various extreme phenomenon, such as rain, flooding, and drought. These extreme phenomenon will have an impact on the quality and quantity of agricultural products. Therefore, it is very important to estimate the potential of rainfall, especially in the agricultural sector. Rainfall is one of the sources of water supply for plants. The data of Global Climate Models (GCM) is used to get some information about rainfall.

Currently, the data of GCM relates to the climate system. However, the results are inadequate because of its global scale. It means, it is difficult to obtain the local scale information to predict rainfall. One of the efforts to overcome this problem is the technique of statistical downscaling (SDS). The SDS method is a method for estimating rainfall by linking global climate elements obtained from the output data of GCM with the local-scale climatic elements through climatology stations. The output data of GCM is generally available in the form of grids which is located within the domain. The SDS model is composed of the covariate variables on a large scale, and the variables are dependent each other [1].

The problem in SDS method is determining the domain [2]. The domain is used as a predictor with many dimensions. In these dimensions, there is a possibility of the curse of dimensionality, the spatial correlation between grids in the domain, and the multicollinearity between variables [3]. These cases need to be resolved to avoid biased allegations. Currently, the SDS evolving models to cope the multicollinearity is the principal component regression (PCR), the continuum regression (CR) and the Least Absolute Shrinkage and Selection Operator (LASSO) regression. According to [4], CR is a generalization of the least squares regression, PCR, and the partial least squares regression. The results show that CR generates better predictions than least squares and partial least squares regression.

The problem in CR is the number of observations which is smaller than the number of predictor variables ($n \ll p$). So that, a pre-processing method is needed in the form of reducing dimensions of the predictor variables. The pre-processing methods which are often used are the principal component analysis (PCA), the LASSO selection and the forward selection. [2] used CR with PCA pre-processing to generate multicollinearity. In addition to PCA pre-processing, LASSO pre-processing shrinks the estimator coefficient to zero in order to allow the variable selection [5]. [6] provided an alternative solution for the variable selection and the coefficients shrinkage of linear regression model by using LASSO. The results show that LASSO provides an alternative for variable selection. The LASSO method can provide excellent predictive accuracy and improve the interpretability of the model [7]. The objectives of this research are modeling the statistical downscaling by using the continuum regression with LASSO selection for estimating the rainfall and comparing the models with LASSO regression models, the continuum regression with PCA, and the continuum regression with forward selection.

II. METHOD

A. Data

The data used in this study is the monthly rainfall data of GCM released by the Climate Forecast System Reanalysis (CFSR) as a predictor variable. The GCM domain used is a number of 9×9 grid squares (0.5°×0.5° for each grid). The local rainfall data as a response variable issued by the Badan Meterologi, Klimatologi dan Geofisika (BMKG) for the period of January 2011 to December 2019 with the observation rain stations in West Java.

B. Data Analysis Procedure

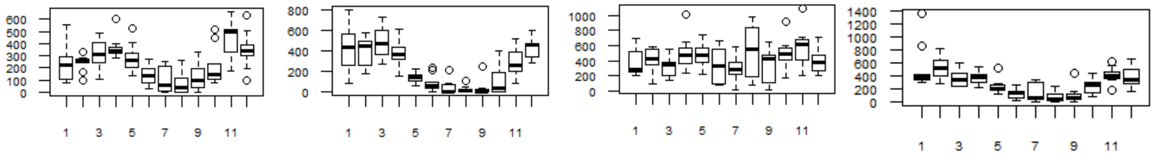
The data analysis is performed by using the R 3.5.1 software. The stages carried out in this study are as follows.

1. Preparing and exploring the data
2. Performing the continuum regression modeling with LASSO selection.
   a. Selecting the predictor variables by using the LASSO method. This algorithm is available in the glmnet package. The selection of predictor variables is based on the selection of the optimum lambda value with the smallest cross validation value. The procedure of k- fold cross validation is as follows [8].
      1) Dividing the data randomly into k sections or folds with k=1,2,…,K, so it is formed $F_1,F_2,\ldots,F_k$.
      2) Using folds K-1 to build a model (training).
      3) The model obtained from (2) is presumed as the value of the response variable.
      4) Calculating the value of the mean squared error prediction (MSEP) of the fold response variable to -k ,
      $$\mathrm{MSEP_k}=\sum\nolimits_{i\in N_k}\frac{(y_i-\hat{y}_i)^2}{n_k}$$
      with $\hat{y}_i$ is the estimated value of y and $\mathcal{N}_k$ is a data set of fold to - k.
      5) Repeating the steps (2) to (4) as much as K, so it is obtained $\mathrm{MSEP_1,MSEP_2,\ldots,MSEP_K}$. Each candidate model generates the measure of estimator performance,
      $$\mathrm{CV}=\sum\nolimits_{k=1}^{K}\frac{\mathrm{MSEP_k}}{K}$$
      in which the optimal model is the model with the smallest CV value.
      6) Modeling the rainfall data and GCM output data with the continuum regression modeling based on the LASSO method selection results. The continuum regression model is formulated in Equation (1) below:
      $$\mathbf{y}=\mathbf{T}_h\xi+\boldsymbol{\varepsilon} \tag{1}$$
      with the matrix of $\mathbf{T}_h\mathbf{=XW}_h$ and $\mathbf{W}_h\mathbf{=(w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_h)$ contain the variable columns h with h<p and it is called a weighting matrix. [9] formulated the vector $\mathbf{w}_i(i=1,2,\ldots,h)$ as in Equation (2) below:
      $$\mathbf{w}_i=\arg\max_w\left\{\mathrm{Cov}(\mathbf{x_w},\mathbf{y})^2\mathrm{Var}(\mathbf{x_w})^{[\delta/(1-\delta)]-1}\right\} \tag{2}$$
      with constraints of $\|\mathbf{w}_i\|=1$ and $\mathrm{Cov}\left(\mathbf{x}_{\mathbf{w_i}},\mathbf{x}_{\mathbf{w_j}}\right)=0$ for $i<j=1,2,\ldots,h$. Meanwhile, the δ adjustment parameter are the real numbers of $0\leq\delta<1$. The Estimation of parameters ξ in Equation (1) is carried out using the least squares method and formulated as in Equation (3), and the predictive value can be calculated through Equation (4) below.
      $$\hat{\boldsymbol{\xi}}=\left(\mathbf{T}_h^T\mathbf{T}_h\right)^{-1}\mathbf{T}_h^T\mathbf{Y} \tag{3}$$
      $$\hat{\boldsymbol{\beta}}=\mathbf{W}\left(\mathbf{T}_h^T\mathbf{T}_h\right)^{-1}\mathbf{T}_h^T\mathbf{Y}$$
      $$\hat{\mathbf{y}}=\mathbf{XW}_h\hat{\boldsymbol{\xi}} \tag{4}$$
3. Predicting and evaluating the model is conducted by calculating the model accuracy through the root mean squared error prediction (RMSEP) with the formula
   $$\mathrm{RMSEP}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$
   and the correlation (r) between the actual rainfall (y) with the predicted rainfall (ŷ) with formula
   $$r_{y_i,\hat{y}_i}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$$
4. Comparing the results of modeling evaluation by using the LASSO regression model, the continuum regression with PCA and the continuum regression with forward selection through RMSEP values and correlation.
5. Examining the model consistency based on the RMSEP value and the correlation by carrying out the validation four times using four different validation data of the last four years, namely 2019, 2018, 2017 and 2016.

III. RESULTS

A. Data Exploration

The distribution of rainfall in each region in Indonesia is diverse due to various factors. [10] divided the pattern of rainfall in Indonesia into three regions, they are: Region A (monsoonal type), Region B (equatorial type) and Region C (local type). West Java Region has monsoonal type of rainfall pattern. This region experiences the average of the highest monthly rainfall one time and the average of the lowest monthly rainfall one time. It happens because of the effects of west monsoon which is occurred on November until March and east monsoon which is occurred on May until September. The region of monsoonal rainy type normally get the maximum rainfall in the period December-January-February (DJF). However, in the transition monsoon, the monthly data of rainfall shows the second maximum value during the period of March-April-May (MAM). [11] stated that the maximum rainfall in Java island can't be defined equally occur during the period of DJF. It because some areas also have maximum rainfall during the transition monsoon in the period of MAM. The data exploration of West Java rainfall uses the monthly data in the period of 2011 to 2019 for each observation station. The data can be seen in Figure 1.



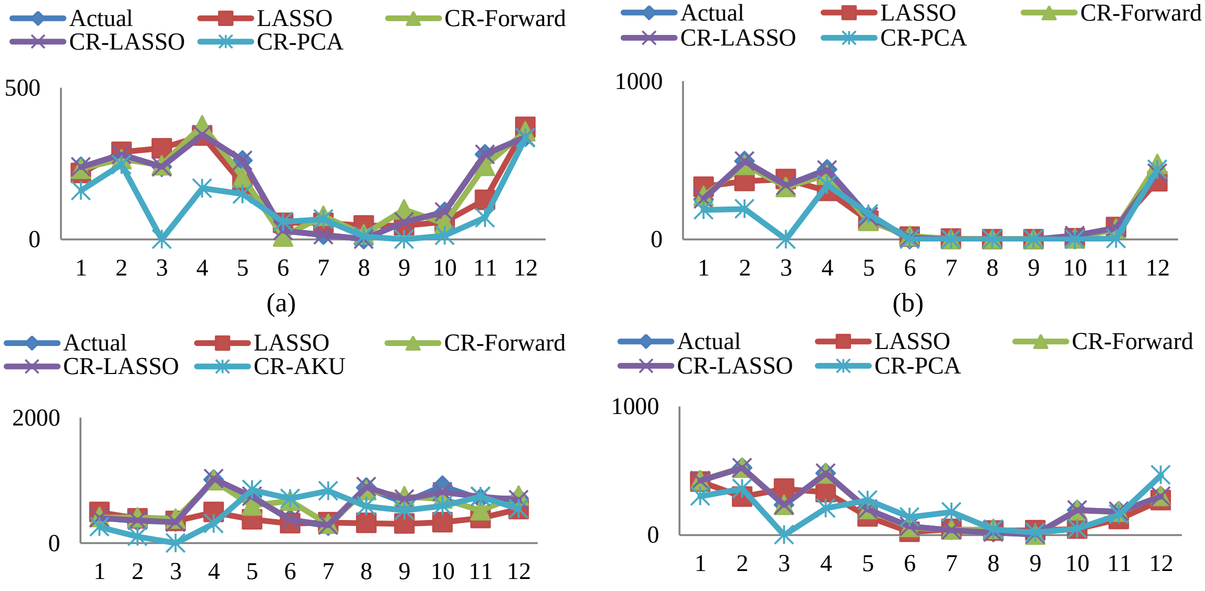
(a) (b) (c) (d)

Figure 1 Boxplot of monthly rainfall station (a) Bandung, (b) Jatiwangi, (c) Bogor dan (d) Citeko

Figure 1(a) shows that the highest rainfall occurs on November until April, with an average amount of rainfall of 200 mm/month to 435 mm/month. The lowest rainfall occurs on June until October with an average amount of rainfall of 85 mm/month to 140 mm/month for Bandung station. Figure 1(b) shows the highest rainfall occurs on November until April, with an average amount of rainfall of 290 mm/month up to 485 mm/month. The lowest rainfall occurs on June until October with an average amount of rainfall ranges from 10 mm/month to 125 mm/month for Jatiwangi station. Figure 1(c) shows that the highest rainfall occurs in every month with the average amount of monthly rainfall >300 mm/month for the station Bogor. It happens because the morphological conditions of Bogor are mostly in the highlands, hills and mountains; furthermore, the climatological conditions including the very wet tropical climate. Figure 1(d) shows that the highest rainfall occurs on November until April, with an average amount of rainfall of 360 mm/month up to 530 mm/month, and the lowest rainfall is on May until September with an average amount of rainfall <300 mm/month for Citeko station.

B. Continuum Regression with LASSO Selection

LASSO method is a pre-processing stage which is conducted before continuum regression modeling. This method is able to reduce the coefficient estimator of zero exact, so that it can select the variables. It can improve the accuracy and interpretability of the model by eliminating predictor variables which are irrelevant to the response variables. The variables which are selected by LASSO are used in the continuum regression modeling. The model obtained is used in predicting monthly rainfall. The plot of the prediction results for the first validation or 2019 can be seen in Figure 2 below.



(a) (b)

(c) (d)

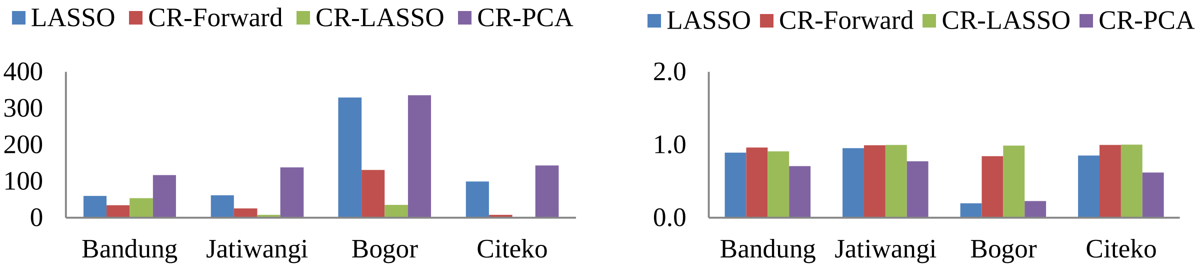
Figure 2 Actual plot and prediction model of station (a) Bandung, (b) Jatiwangi, (c) Bogor dan (d) Citeko

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10651

Figure 2 is a comparison plot of the actual rainfall value in the prediction results of the LASSO regression, the continuum regression with PCA, the continuum regression with forward selection and the continuum regression with LASSO selection for each station. The results show that the prediction value of the continuum regression model with LASSO selection and the continuum regression with forward selection tend to have patterns which are relatively the same as the pattern of actual rainfall, so the error between the actual value and the predicted rainfall is relatively small. The differences between the actual rainfall value and the predicted value will be evaluated through the value of RMSEP and its correlation.

C. Evaluation of Regression Model

The evaluation of regression models is performed by calculating the RMSEP value and the correlation between actual rainfall and predicted rainfall. The regression models compared are the LASSO regression model, the continuum regression with PCA, and the continuum regression with forward selection. The results of the regression model evaluation for the first validation or 2019 can be seen in Figure 3 below.



(a) (b)

Figure 3 Regression model evaluation for 2019 (a) RMSEP and (b) Correlation

Figure 3 is the evaluation result charts of several regression models for each station. Figure 3(a) shows the value of RMSEP and Figure 3(b) shows the correlation value of RMSEP. The best model for Bandung station is the continuum regression with forward selection with a minimum RMSEP value of 33.8603 and a maximum correlation of 0.9633. The best model for Jatiwangi station is the continuum regression with LASSO selection with a minimum RMSEP value of 7.7552 and a maximum correlation of 0.9993. The best model for Bogor station is the continuum regression with LASSO selection with a minimum RMSEP value of 34.8461 and a maximum correlation of 0.9897. The best model for Citeko station is the continuum regression with LASSO selection with a minimum RMSEP value of 0.0397 and a maximum correlation of 1.000.

D. Model Validation and Consistency

The validation is performed by using four different validation data based on the last four years data, namely 2019, 2018, 2017 and 2016. The consistency of the model based on the value RMSEP and the correlations which can be seen in Table 1-4 below.

Table 1 Validation model of Bandung station

Validation	Evaluation	LASSO	CR-Forward	CR-LASSO	CR-PCA
2019	RMSEP	59.5023	**33.8603**	53.0569	116.9794
	Correlation	0.8901	**0.9633**	0.9072	0.7067
2018	RMSEP	81.8103	**38.9886**	41.7510	173.0552
	Correlation	0.9355	**0.9653**	0.9607	0.6903
2017	RMSEP	137.4381	**106.1387**	110.6066	279.9951
	Correlation	0.5359	**0.7381**	0.7120	0.1657
2016	RMSEP	110.8518	**48.9254**	88.8541	296.5294
	Correlation	0.7396	**0.9522**	0.8322	0.7933
Average	RMSEP	97.4006	**56.9783**	73.5672	216.6398
	Correlation	0.7753	**0.9047**	0.8530	0.5890

Table 1 shows the RMSEP value and the correlation for each model on the four validation data in Bandung station. The consistency of the model for each validation data which provides the minimum RMSEP value and the maximum correlation value is a continuum regression model with forward selection. The results show that this model consistently has smaller RMSEP values and greater correlation than the LASSO regression model, the continuum regression with PCA and the continuum regression with LASSO selection. It shows that the average value of the smallest RMSEP and the highest correlation value is the continuum regression model with forward selection, namely 56.9783 and 0.9047.

Table 2 Validation model of Jatiwangi station

Validation	Evaluation	LASSO	CR-Forward	CR-LASSO	CR-PCA
2019	RMSEP	61.1119	25.3991	**7.7552**	137.8141
	Correlation	0.9542	0.9909	**0.9993**	0.7748
2018	RMSEP	149.7064	94.7170	**47.3918**	277.0608
	Correlation	0.7694	0.9142	**0.9793**	0.3474
2017	RMSEP	141.3708	57.1089	**16.1061**	350.1397
	Correlation	0.8443	0.9736	**0.9979**	0.5224
2016	RMSEP	94.2279	50.0336	**21.4112**	200.4308
	Correlation	0.8226	0.9524	**0.9915**	0.6266
Average	RMSEP	111.6042	56.8146	**23.1661**	241.3614
	Correlation	0.8476	0.9578	**0.9920**	0.5678

Table 2 shows the RMSEP value and the correlation for each model on the four validation data in Jatiwangi station. The consistency of the model for each validation data which gives the minimum RMSEP value and the maximum correlation value is the continuum regression model with LASSO selection. The results show that this model consistently has a smaller RMSEP value and greater correlation than the LASSO regression model, the continuum regression with PCA and the continuum regression with forward selection. It shows that the smallest average RMSEP value and the highest correlation value is the continuum regression model with LASSO selection, namely 23.1661 and 0.9920.

Table 3 Validation model of Bogor station

Validation	Evaluation	LASSO	CR-Forward	CR-LASSO	CR-PCA
2019	RMSEP	329.4534	131.0199	**34.8461**	335.3703
	Correlation	0.1955	0.8425	**0.9897**	0.2285
2018	RMSEP	134.0708	62.9845	**33.6528**	386.2071
	Correlation	0.8128	0.9173	**0.9771**	-0.4714
2017	RMSEP	194.3788	171.9880	**153.4617**	390.9704
	Correlation	-0.1591	0.4247	**0.3567**	0.1242
2016	RMSEP	197.4429	120.7508	**0.0172**	460.8324
	Correlation	0.3492	0.2016	**1.0000**	-0.3385
Average	RMSEP	213.8365	121.6858	**55.4944**	393.3450
	Correlation	0.2996	0.5965	**0.8309**	-0.1143

Table 3 shows the RMSEP value and the correlation for each model on the four validation data in Bogor station. The consistency of the model for each validation data which provides the minimum RMSEP value and the maximum correlation value is the continuum regression model with LASSO selection. The results show that this model consistently has a smaller RMSEP value and greater correlation than the LASSO regression model, the continuum regression with PCA and the continuum regression with forward selection. It shows that the smallest average RMSEP value and the highest correlation value is the continuum regression model with LASSO selection, namely 55.4944 and 0.8309.

Table 4 Validation model of Citeko station

Validation	Evaluation	LASSO	CR-Forward	CR-LASSO	CR-PCA
2019	RMSEP	99.4824	7.9739	**0.0397**	143.5886
	Correlation	0.8543	0.9990	**1.0000**	0.6197
2018	RMSEP	141.8329	27.2546	**8.5903**	221.0699
	Correlation	0.7615	0.9919	**0.9992**	0.6493
2017	RMSEP	93.4154	**16.4557**	119.9355	290.3297
	Correlation	0.8824	**0.9959**	0.8349	0.4828
2016	RMSEP	103.8974	**17.4487**	78.3643	232.2325
	Correlation	0.7380	**0.9937**	0.8534	0.5832
Average	RMSEP	109.6570	**17.2832**	51.7324	221.8052
	Correlation	0.8090	**0.9951**	0.9219	0.5837

Table 4 shows the RMSEP value and the correlation for each model on the four validation data in Bandung station. The consistency of the model for the validation data in 2019 and 2018 which provides the minimum RMSEP value and the maximum correlation value is the continuum regression model with LASSO selection. Meanwhile, the validation data in 2017 and 2016 is shown by the continuum regression model with forward selection. The results show that the smallest RMSEP value and the highest correlation value on average is the continuum regression model with forward selection, namely 17.2832 and 0.9951.

IV. CONCLUSION

The modeling of statistical downscaling using a continuum regression model with LASSO selection and a continuum regression model with forward selection can be used to predict rainfall in West Java province. The results show that the continuum regression model with LASSO selection provides fairly accurate prediction results for Jatiwangi and Bogor stations compared to the LASSO regression model, the continuum regression with PCA and the continuum regression with forward selection. The continuum regression model with forward selection also provides fairly accurate prediction results in predicting rainfall at Bandung and Citeko stations.

REFERENCES

[1] A.M Soleh, A.H. Wigena, A. Djuraidah, A. Saefuddin. 2015. “Statistical Downscaling to Predict Monthly Rainfall Using Linear Regression with L1 Regularization (LASSO)”. Applied Mathematical Sciences, vol. 9(108), pp. 5361-5369.

[2] Sutikno, Setiawan, H. Purnomoadi. 2010. “Statistical downscaling output GCM modeling with continuum regression and pre-processing PCA approach”. The Journal for Technology and Science, vol. 21(3), pp. 109-112.

[3] A.H. Wigena . “Pemodelan Statistical Downscaling Dengan Regresi Projection Pursuit Untuk Peramalan Curah Hujan Bulanan: Kasus Curah Hujan Bulanan Di Indramayu”, disertasi. Institut Pertanian Bogor, Indonesia, 2006.

[4] Setiawan. “Pendekatan Regresi Kontinum dalam Model Kalibrasi”, disertasi. Institut Pertanian Bogor, Indonesia, 2007.

[5] T. Hastie, R. Tibshirani, J. Friedman. 2008. “The Elements of Statistical Learning. Data mining, Inference, and Prediction, 2nd ed”. London (GB): Springer.

[6] Soleh AM, Aunuddin. 2013. “LASSO: Solusi Alternatif Seleksi Peubah dan Penyusutan Koefisien Model Regresi Linier”. Forum Statistika dan Komputasi: Indonesian Journal of Statistics, vol. 18(1), pp. 21-27.

[7] V. Fonti. 2017. “Feature Selection using LASSO”. Vrije Universiteit Amsterdam.

[8] Y. Jung , J. Hu. 2015. “A K-fold Averaging Cross-validation Procedure”. Journal of Nonparametric Statistics, vol. 27(2), pp. 1-13.

[9] M. Stone, R. J. Brooks. 1990. “Continuum Regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal component regression (with discussion)”. Journal of the Royal Statistical Society Series B, vol. 52, pp. 237-269.

[10] E. Aldrian, R. D. Susanto. 2003. “Identification of Three Dominant Rainfall Region within Indonesia and Their Relationship to Sea Surface Temperature”. International Journal of Climatology, vol. 23, pp. 1435-1452.

[11] M. A. Fauzi. 2014. “Pengaruh Transisi Monsun Terhadap Siklus Tahunan Curah Hujan di Pulau Jawa”, skripsi. Institut Teknologi Bandung, Indonesia, 2014.

AUTHORS

First Author – Arwini Arisandi, S.Si., college student, Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, Jawa Barat, 16680, Indonesia, email: arwini_arisandi@apps.ipb.ac.id.
Second Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc., lecturer, Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, Jawa Barat, 16680, Indonesia, email: aji_hw@apps.ipb.ac.id.
Third Author – Dr. Agus Mohamad Soleh, M.T, lecturer, Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, Jawa Barat, 16680, Indonesia, email: agusms@apps.ipb.ac.id.

Correspondence Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc., lecturer, Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor, Jawa Barat, 16680, Indonesia, email: aji_hw@apps.ipb.ac.id.