Two-Stage Statistical Downscaling Modeling with Multi-Class Random Forest on Rainfall Prediction

Riana Hadiana[1], Agus Mohamad Soleh[2], Bagus Sartono[3]
{riana_cato@apps.ipb.ac.id[1], agusms@apps.ipb.ac.id[2], bagusco@gmail.com[3]}

Department of Statistics, IPB University, Bogor, 42176, Indonesia[1]
Department of Statistics, IPB University, Bogor, 42176, Indonesia[2]
Department of Statistics, IPB University, Bogor, 42176, Indonesia[3]

Abstract. Statistical downscaling (SD) modeling to predict rainfall has been widely used using the General Circulation Model (GCM) output. Based on the previous study, SD modeling to predict rainfall by rainfall grouping (two-stages) gives a smaller Root Mean Squares Error of Prediction (RMSEP) than SD modeling without rainfall grouping (one-stage). In this study, the daily and monthly rainfall were divided into three groups based on their intensity (volume) and two-stages SD modeling was applied to predict rainfall. The first stage was rainfall groups classification using random forest. The second stage was rainfall prediction using Partial Least Squares Regression (PLSR). The accuracy obtained by random forest for daily and monthly rainfall lied between 62%-84%. The RMSEP obtained from two-stages SD modeling for daily rainfall was similar to one-stage SD modeling, where the Coefficient of Variation (CV) was above 100%. The different results happened when two-stages SD modeling was applied to monthly data. The RMSEP obtained was better than one-stage SD modeling, where the CV lied between 30%-50%.

Keywords: Multi-class, partial least square regression, random forest, statistical downscaling.

1 Introduction

Rainfall modeling has been widely used using the GlobalClimate Model (GCM) output. Data provided by GCM is data from climate parameter modeling on a global scale by utilizing satellite data and local climate data from climate observation stations on land. GCM output can be used as a tool to predict climate and weather numerically as well as a source of primary information to assess climate change[1].

GCM output can be used to estimate climate parameters on a local scale using downscaling techniques. One of the downscaling techniques that can be used to obtain local-scale information from GCM output data is statistical downscaling (SD) [2]. SD uses a statistical model to connect functionally between global climate parameters obtained from GCM output with local climate parameters obtained from climatology observation stations.

The GCM output for the specific grids located in a domain above the target location is used as the predictor in SD. The grids at a GCM domain are strongly correlated, so they cannot be directly used as predictors because there is multicollinearity. One solution to the problem of multicollinearity in SD is the transformation of variables (Principal Components Regression (PCR) [3]and Partial Least Square Regression (PLSR)[4].

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290525

The one-stage SD modeling using PCR and PLSR is not good enough to predict rainfall, especially at observation stations that have extreme rainfall intensity. Adding rainfall groups to the model as a dummy variable provides better rainfall predictions[5]. Unfortunately, this method cannot be applied to predict future rainfall because of rainfall group information is not available. The two-stage SD modeling with classification modeling of rainfall groups is expected to increase the precision of rainfall predictions.

Nadya R [6]conducted two-stage SD modeling using Global Precipitation Climatology Project (GPCP) data as the predictor where the first stage was classification modeling using logistic regression and classification tree to predict rainfall groups. The results of the rainfall group prediction were used as dummy variables on the second stage using the PCR model. Khairunisa [7]conducted two-stage SD modeling using Climate Forecast System (CFS) data as the predictor where monthly rainfall groups divided into four groups based on quartiles. The first stage was classification modeling using ordinal logistic regression to predict rainfall groups. The second stage was rainfall prediction using PCR and PLSR model where these model were applied to each rainfall group. The classification model used produces a low accuracy so that the RMSEP obtained was not significantly better than the one-stage SD modeling.

This study aimed to find the best daily and monthly rainfall models using two-stage SD modeling, where local rainfall are grouped into three rainfall groups. The first step is classification modeling using the Random Forest (RF). RF algorithm is chosen because in many cases, RF algorithm produces better accuracy than logistic regression[8]. RF algorithm also supports multi-class classification. Besides that, multi-class classification modeling using RF can be done using binarization approach. The second stage is modeling using RKTP to predict rainfall after rainfall group prediction is obtained by classification modeling.

2 Materials

The data used in this study were daily rainfall data from eight rain observation stations from 2011 to 2018 provided by the Meteorological, Climatological, and Geophysical Agency (BMKG) as the response variable. The chosen rain observation station represented the rainfall zone in Indonesia. The monsoonal rainfall pattern was represented by Bogor, Citeko, Jatiwangi, Bandung (West Java), and Serang (Banten) rain observation stations. The equatorial rainfall pattern was represented by Syarif Kasim II (Riau) and Mempawah (West Kalimantan) rain observation stations. The local rainfall pattern was represented by the Pattimura (Ambon) rain observation station. The predictor was the daily precipitation rate from GCM output, version 2 of Climate Forecast System (CFS) with the area of the grids was $0.5^{o} \times 0.5^{o}$. The domain area of observation was 6×6 grids at each station location. Daily rainfall data was downloaded from http://dataonline.bmkg.go.id and GCM output data was downloaded from https://rda.ucar.edu [9].

3Methods

The steps of data analysis in this study are as follows.

1. Data preprocessing
   a. The CFS data domain for each rain station was formed in the size of 6×6 grids.

b. Monthly rainfall was formed from BMKG daily rainfall data by calculating the average daily rainfall in each month, then multiplying by the number of days in each corresponding month.
c. CFS monthly rainfall data was formed by aggregating daily data into monthly.
d. Daily and monthly rainfall data at each rain observation station were combined with CFS data.

2. One-stage SD modeling was applied for daily and monthly rainfall using PLSR. The RMSEP of each model were calculated using 5-fold cross-validation.
3. Two-stage SD modeling for daily and monthly rainfall
   a. Daily rainfall was grouped into three groups with the following conditions.
      - The first group (K1) was the group with no rain.
      - The second group (K2) was the group with rainfall more than 0 and less than equal with B mm/day.
      - The third group (K3) was the group with rainfall more than B mm/day.
      - The B values used in this studywereP_1, P_2, P_3, …, P_{99} from daily rainfall greater than zero, where P_i was the i-th percentile.
   b. Monthly rainfall was grouped into three groups with the following conditions.
      - The first group (K1) was the group with rainfall less than or equal with B_1 mm/month.
      - The second group (K2) was the group with rainfall more than B_1 mm/month and less than equal with B_2 mm/month.
      - The third group (K3) was the group with rainfall more than B_2 mm/month.

      The pairs of B_1 and B_2 used in this study were all combinations of P_1, P_2, P_3, …, P_{99} from the monthly rainfall where $B_1 < B_2$.
   c. Data were divided into training data and testing data using 5-fold cross-validation.
   d. Classification modeling using RF algorithm on training data was used to predict groups of the testing data.

RF algorithm is the development of CART (Classification and Regression Tree). RF applies the bootstrap aggregating (bagging) method and the random features selection[10]. In a RF, many trees are grown so that a forest is formed. Then the analysis is carried out on the tree collection so that it can be used to classify binary responses (two classes).

Suppose a group of data consists of N observations with M explanatory variables. The RF algorithm consists of the following stages.

- Drawing an N-sized random sampling with replacement from N-sized datasets. This stage is called bootstrapping.
- Form a classification tree without pruning based on the bootstrap example in step 1.At each node, m variables ($m \ll M$) are selected randomly, and the best splitter is chosen among the selected m variables.
- Repeat the previous step k times to form a group of trees or forest. Response to the observation is predicted by aggregating the predicted results from k trees. In the classification problem, the response is made based on the majority vote.

RF algorithm is usually used in classification problems with binary responses. However, the RF algorithm can also be applied to classification problems with multi-responses (multi-class). It is possible because of the RF algorithm base for classification is the classification tree, where the classification tree can be used to classify data with multi-class response variables. Each leaf node of a classification

tree produces the probability values for each class. The class prediction at each leaf node of a classification tree is the class with the highest probability value.

e. PLSR modeling for K1, K2, and K3 on training data was used to predict rainfall of the testing data.

PLSR combines the principal components analysis with the multiple regression analysis. Each PLSR components are obtained by maximizing variation between response variables (*Y*) and explanatory variables (*X*) to get the components that explain Y more than the components obtained from the principal components analysis.

Suppose$\boldsymbol{X}$ is an $n \times p$ matrix and$\boldsymbol{Y}$ is an $n \times q$matrix where n is the number of observations, $p$is the number of explanatory variables, and$q$ is the number of response variables. $\boldsymbol{X}$ consist of vectors$\boldsymbol{x}_j$, $j = 1, 2, 3, \ldots, p$and$\boldsymbol{Y}$consist of vectors$\boldsymbol{y}_k$, $k = 1, 2, 3, \ldots, q$. The PLSR method produces several new components that will model $\boldsymbol{X}$ against$\boldsymbol{Y}$, so that the relationship between $\boldsymbol{X}$ and$\boldsymbol{Y}$is obtained. These new components are referred to as$\boldsymbol{X}$ scores, can be written ast_a, $a = 1, 2, 3, \ldots, A$.

The $\boldsymbol{X}$ score is a linear combination of the original variables$\boldsymbol{x}_j$ with a coefficient, "weights",$\boldsymbol{w}_{ja}$ .The process is formulated as follows[11].

$$\begin{cases} t_{ia} = \sum_j w_{ja}\, x_{ij}, i = 1,2, \ldots, n \\ \boldsymbol{T} = \boldsymbol{XW} \end{cases} \#(1)$$

The $\boldsymbol{X}$ score, t_a, has the following properties.

- The $\boldsymbol{X}$ score is multiplied by m_{aj}, so the $\boldsymbol{X}$-residuals(e_{ij})are small.

$$\begin{cases} x_{ij} = \sum_a t_{ia} m_{aj} + e_{ij} \\ \boldsymbol{X} = \boldsymbol{TM}' + \boldsymbol{E} \end{cases} \#(2)$$

With multivariate $\boldsymbol{Y}$ (when $k > 1$), the corresponding $\boldsymbol{Y}$-scores ($\boldsymbol{u}_a$) are multiplied by the weights c_{ak}, so the residuals (g_{ik}) are small.

$$\begin{cases} y_{ik} = \sum_a u_{ia} c_{ak} + g_{ik} \\ \boldsymbol{Y} = \boldsymbol{UC}' + \boldsymbol{G} \end{cases} \#(3)$$

- the $\boldsymbol{X}$-scores are good predictors of Y, i.e.

$$\begin{cases} y_{ik} = \sum_a c_{ak}\, t_{ia} + f_{ik} \\ \boldsymbol{Y} = \boldsymbol{TC}' + \boldsymbol{F} \end{cases} \#(4)$$

The $\boldsymbol{Y}$-residuals, f_{ik}, express the deviations between the observed and the modelled responses. Based on equation (1) and equation (4) can be rewritten to look as multiple regression models as follow.

$$\begin{cases} y_{ik} = \sum_a c_{ak} \sum_j w_{ja} x_{ij} + f_{ik} = \sum_j b_{kj} x_{ij} + f_{ik} \\ \boldsymbol{Y} = \boldsymbol{XWC}' + \boldsymbol{F} = \boldsymbol{XB} + \boldsymbol{F} \end{cases} \#(5)$$

The PLSR coefficients, b_{kj}, can be written as:

$$\begin{cases} b_{kj} = \sum_a c_{ak} w_{ja} \\ \boldsymbol{B} = \boldsymbol{WC}' \end{cases} \#(6)$$

Predictions for the new observation data were obtained based on data $\boldsymbol{X}$ and the coefficient matrix $\boldsymbol{B}$.

f. Group prediction for testing data
g. Rainfall prediction for testing data using PLSR model correspondence with its group prediction results. Daily rainfall predictions for observations with K1 classification results was 0 mm or no rain.
h. Daily and monthly rainfall prediction for each observation on testing data using each PLSR model taken from part 3e ($\hat{Y}_{ik}$), then $\hat{Y}_{ik}$ multiplied with the probability for each group obtained from 3d (P_{ik}). The rainfall prediction for the *i*-th observation is

$$\hat{Y}_i = \sum_{k=1}^{3} \hat{Y}_{ik} P_{ik} \#(7)$$

where $i = 1, 2, 3, \ldots n$, n is the amount of testing data, k is the amount of group, and daily rainfall prediction for K1 is $\hat{Y}_{i1} = 0$.
i. RMSEP RF was calculated using rainfall prediction taken from 3g. RMSEP RFWt was calculated using rainfall prediction taken from 3h.
j. The selected *B* was *B* with the smallest RMSEP RF or RMSEP RFWt.

4. The best model was determined by comparing the RMSEP (Root Mean Squares of Error Prediction) and CV (Coefficient of Variation) among the models. The best model is the model with the smallest RMSEP and CV values. RMSEP and CV are calculated using the following formula.

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}} \#(8)$$

$$\text{CV} = \frac{\text{RMSEP}}{\bar{Y}} \#(9)$$

where Y_i is observed value, $\hat{Y}_i$ is predicted value, n is the number of observations, and $\bar{Y}$ is the mean of observed value (Y_i).

4Results and Discussion

Daily and monthly local rainfall as response variables were modeled and predicted directly using PLSR with precipitation rates in 36 selected grids around the rain observation stations as the predictor. The modeling used 5-fold cross-validation. Model performance was measured using RMSEP and CV. The CV is the ratio between the standard deviation (RMSEP) and the average of the response variable. The CV is expressed as a percentage. The model with a smaller CV would produce a better prediction. However, there is no standard rule that states the specific size of a good CV. The average RMSEP and CV for daily and monthly rainfall data at each rain observationstations are presented in Table 1.

Table 1. The Average RMSEP and CV one-stage SD modeling using PLSR on daily and monthly rainfall

No	Observation Stations	Daily		Monthly	
		RMSEP	CV	RMSEP	CV
1	Bogor	19.88	155%	185.87	46%
2	Citeko	16.42	162%	122.26	40%
3	Jatiwangi	16.81	212%	115.52	47%
4	Bandung	13.56	164%	96.58	39%
5	Serang	10.88	200%	71.87	43%
6	Mempawah	16.64	190%	130.16	48%
7	Syarif Kasim II	17.54	201%	108.03	41%
8	Pattimura	22.62	198%	129.63	39%

Daily rainfall modeling using one-stage SD modeling using PLSR at each rain observation stations produced RMSEP that bigger than the average of the rainfall. It caused the CV obtained was more than 100%. Daily rainfall modeling resulted has the smallest RMSEP at the Serang observation station. However, the Bogor observation station has the smallest CV. The one-stage SD modeling on monthly rainfall data produced RMSEP that was smaller than the average of the rainfall at all rain observation stations. It can be seen from the CV in all observation stations were smaller than 50%. The lowest RMSEP value was obtained at the Serang observation station with the RMSEP around 71.87 mm/month. The largest RMSEP value was obtained at the Bogor observation post with the RMSEP around 185.87 mm/month. The smallest CV was obtained at Bandung and Pattimura observation station even though the RMSEP in these two observation stations was not the smallest.

The results of the two-stage rainfall modeling in this study can be seen in Table 2 and Table 3. Two-stage SD modeling on daily rainfall produced *B* values that varied at each observation post. The value of *B* for each observation post in Table 2 was chosen when it produced the smallest RMSEP RF or RMSEP RFWt value. If the classification modeling can produce 100% accuracy, then the two-stage SD modeling will produce the RMSEP value as in the RMSEP column in Table 2. However, the resulted accuracy value did not reach 100%. The RF accuracy values produced at each observation post varied from 62% at the Syarif Kasim II observation post to 77% at the Bogor observation post. The RF RMSEP value, which has

decreased from one-stage RMSEP, only occurred at the Bogor observation station, while the RMSEP RFWt value has decreased at all observation station except the Serang observation post. Although there was a decrease in the value of RMSEP, the decrease that occurred was not significant compared to the RMSEP one-stage SD modeling.

Table 2. Two-stage SD modeling results for daily rainfall

No	Observation Stations	*B*	RMSEP	Accu-racy	RMSEP		CV (%)	
					RF	RFWt	RF	RFWt
1	Bogor	73.0	14.46	77%	19.76	19.81	155	155
2	Citeko	22.2	9.64	65%	17.24	16.20	170	160
3	Jatiwangi	66.1	10.96	75%	16.83	16.61	213	210
4	Bandung	30.7	7.32	74%	13.60	13.33	165	162
5	Serang	48.3	7.84	70%	11.13	10.92	205	201
6	Mempawah	62.2	11.09	71%	16.72	16.61	190	189
7	Syarif Kasim II	39.3	9.81	62%	17.68	17.52	203	201
8	Pattimura	0.9	21.44	70%	22.70	22.12	199	194

Table 3. Two-stage SD modeling results for monthly rainfall

No	Observation Stations	B_1	B_2	RMSEP	Accu-racy	RMSEP		CV (%)	
						RF	RFWt	RF	RFWt
1	Bogor	121.8	605.2	127.90	73%	185.75	193.53	46.2	48.1
2	Citeko	385.5	469.5	88.01	75%	136.86	110.09	44.3	35.7
3	Jatiwangi	0.4	492.6	79.12	77%	121.69	102.07	49.6	41.6
4	Bandung	79.7	205	78.70	76%	95.04	91.37	38.3	36.8
5	Serang	71.5	109.6	57.98	81%	70.14	66.30	42.2	39.9
6	Mempawah	120.2	342.1	76.18	63%	117.74	119.48	43.5	44.2
7	Syarif Kasim II	116.6	475	69.95	78%	108.92	105.20	40.8	39.5
8	Pattimura	292.5	668.3	86.78	84%	117.25	133.48	35.3	40.1

Values B_1 and B_2, which produced the smallest RMSEP RF or RMSEP RFWt values using the two-stage monthly rainfall modeling in each observation stations, are presented in columns B_1 and B_2 of Table 3. If the classification modeling can produce 100% accuracy, then the two-stage SD modeling will produce the RMSEP value as in the RMSEP column of Table 3. RF accuracy obtained at each observation stations varied from 63% at Mempawah observation station to 86% at Pattimura observation station. RMSEP RF that have decreased from one-stage SD modeling RMSEP occurred at the Bogor observation post, Bandung, Serang, Mempawah, and Pattimura. While RMSEP RFWt has decreased from one-stage RMSEP at all observation posts except the Bogor and Pattimura observation stations.

Based on Table 2 and Table 3, it can be seen that the RMSEP and CV produced by two-stage SD modeling on daily rainfall was not significantly different from the one-stage SD modeling results, where the CV value was higher than 100%. Two-stage SD modeling on monthly rainfall produced better RMSEP and CV than one-stage SD modeling. The CV values obtained from two-stage SD modeling ranged from 30% -50%.Because the coefficient of diversity obtained was quite small, the variation between monthly rainfall and the predicted value was small too. It could be seen from the plot between monthly rainfall and its predictions in **Figure 1**. The plot at **Figure 1** were generated using the best model of two-stage SD modeling for monthly rainfall where the characteristics of the model taken from Table 3. The predicted value was calculated using 5–fold cross validation to figure out the

ability of the model to predict the future monthly rainfall. The plot showed that the rainfall prediction able to follow the rainfall pattern, although there was still a deviation between the observed rainfall and the predicted value, especially at Bogor observation station (**Figure 1 (a)**).

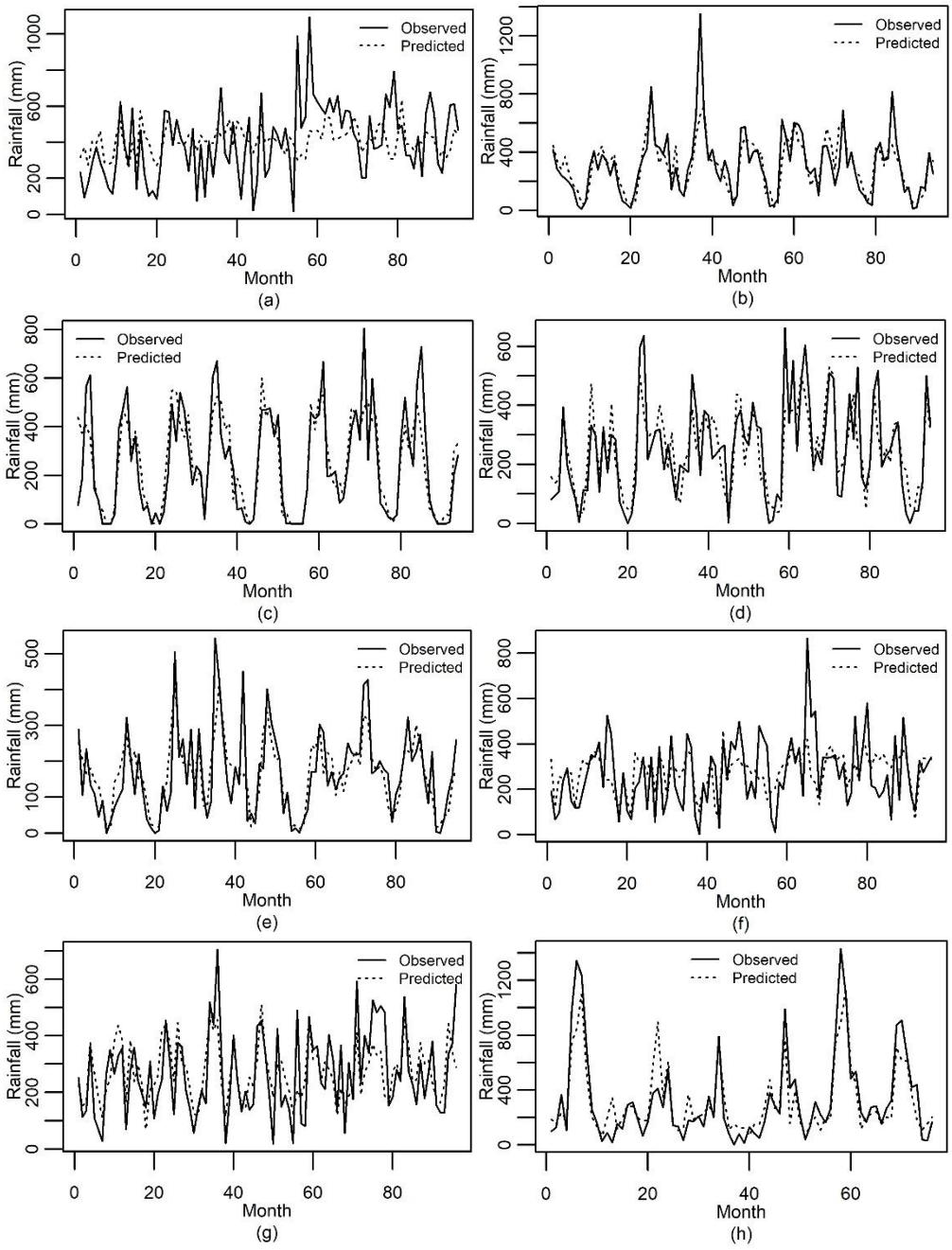

Fig. 1. Plot between observed and predicted rainfall at (a) Bogor, (b) Citeko, (c) Jatiwangi, (d) Bandung, (e) Serang, (f) Mempawah, (g) Syarif Kasim II, and (h) Pattimura observation stations.

5Conclusion

Based on the results and discussion, it can be concluded that the daily and monthly rainfall group boundaries that produce the smallest RMSEP vary at each observation stations. The RMSEP of two-stage SD modelling using the best model was not significantly different from one-stage SD modelling because the model failed to get high accuracy in classifying the class of daily rainfall. The two-stage SD modelling for monthly rainfall classification got better accuracy than daily rainfall classification, so the prediction of monthly rainfall using two-stage SD modelling was better than one-stage SD modelling.

References

[1]Wigena, A. H.: Pemodelan statistical downscaling dengan regresi projection pursuit untuk peramalan curah hujan bulanan : kasus curah hujan bulanan di Indramayu. Institut Pertanian Bogor.(2006)

[2] Wigena, A. H., Djuraidah, A., and Rizki, A.: Semiparametric modeling in statistical downscaling to predict rainfall.*Appl. Math. Sci.*, vol. 9, no. 88, pp. 4371–4382.(2015)

[3]Soleh, A. M., Wigena,A. H., Djuraidah,A., and Saefuddin,A.: Statistical downscaling to predict monthly rainfall using linear regression with L_1 regularization (LASSO).*Appl. Math. Sci.*, vol. 9, no. 108, pp. 5361–5369.(2015)

[4] Wigena,A. H.: Regresi kuadrat terkecil parsial untuk statistical downscaling.*Sci. J. Club BMKG Pros.*, vol. 10, no. 6, pp. 10–13.(2011)

[5]Permatasari,S. M., Djuraidah, A., and Soleh,A. M.: Statistical downscaling with gamma distribution and elastic net regularization (case study : monthly rainfall 1981-2013 at Indramayu).*Proceeding 2nd Int. Conf. Appl. Stat.*, pp. 121–129.(2016)

[6]Nadya R,A.: Pemodelan statistical downscaling untuk menduga curah hujan dengan regresi linear gerombol dan pemodelan dua tahap. Institut Pertanian Bogor.(2018)

[7]Khairunisa.: Pemodelan curah hujan bulanan dengan teknik statistical downscaling melalui pemodelan dua tahap. Institut Pertanian Bogor.(2019)

[8]Couronné, R., Probst, P., and Boulesteix,A. L.: Random forest versus logistic regression: A large-scale benchmark experiment.*BMC Bioinformatics*, vol. 19, no. 1, pp. 1–14.(2018)

[9]Saha, S., Moorthi, S., Pan,H.-L., Wu,X., and Coauthors.: The NCEP climate forecast system reanalysis.*Am. Meteorol. Soc.*, no. August, pp. 1015–1058.(2010)

[10] Breiman,L. E. O.:Random forest.*Mach. Learn.*, vol. 45, no. 1, pp. 5–32, (2001)

[11]Wold, S., Sjostrom,M.,and Eriksson,L.: PLS-regression: a basic tool of chemometrics.*Chemom. Intell. Lab. Syst.*, vol. 58, pp. 109–130.(2001)