Extreme Rainfall Prediction Using Bayesian Binary Quantile Regression

Putri Maulidina Fadilah*, Aji Hamim Wigena*, Anik Djuraidah*

* Department of Statistics, IPB University, Bogor, Indonesia

DOI: 10.29322/IJSRP.10.11.2020.p10702
http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10702

Abstract- Extreme rainfall can cause flood and bring bad impact to the agricultural sector. Indonesia has extreme rainfall in several region such as West Java, where the lowland area is at the northern region and the highland area is at the southern. The Global Circulation Models (GCM) output data can be used to get informations about rainfall. Statistical downscaling (SD) is the technique which can be used to analyze the functional relationship between local scale data (rainfall) and global scale data (GCM output) using statistical method. The extreme rainfall can be analyze using quantile regression in SD to measure the effect of explanatory variables not only at the center but also at the two tails of the data distribution. Bayesian Binary Quantile Regression (BBQR) is an extension of quantile regression where the dependent variable scale is binary. BBQR is based on asymmetric Laplace distribution, so that the parameter estimation of the posterior distribution uses the Markov Chain Monte Carlo (MCMC) method. The multicollinearity in GCM output data can be overcome by adding the Least Absolute Shrinkage and Selection Operator (LASSO) penalty in the model. The BBQR model resulted in the good extreme rainfall prediction and the prediction at lowland area was more accurate than that at the other land area.

Index Terms- Bayesian Binary Quantile Regression, LASSO, MCMC, Statistical Downscaling

I. Introduction

Rainfall is a climate factor that can affect for human life. Extreme rainfall can cause flood and bring bad impact to the agricultural sector. Indonesia has extreme rainfall in some areas such as West Java, where the northern region of West Java is a lowland area while the southern part is a highland area. The difference in altitude resulted in a difference rainfall intensity [9], so accurate analysis was needed for predicting extreme rainfall to reduce the negative impact on the agricultural sector. Global Circulation Models (GCM) output data is used to obtain information about rainfall. But GCM output data is still a global-scale data that has large dimensions so statistical downscaling techniques can be used to analyze the functional relationship between local-scale data (rainfall) and global-scale data (GCM output) using statistical method.

Extreme rainfall can cause the probability distribution of the rainfall to be far from Gaussian [12]. Quantile regression is a method that can be used to analyze asymmetric data [7]. Quantile regression can measure the effect of explanatory variables not only at the center but also at the two tails of the data distribution. Yu and Moyeed (2001) popularized the Bayesian method on quantile regression using sample information and prior distribution to obtain posterior distribution. They suggest that Bayesian quartile regression is based on asymmetric laplace distribution, so that to obtain estimated paramater from posterior distribution using Markov Chain Monte Carlo (MCMC) method. The famous MCMC methods are Metropolis-Hasting [14] and Gibbs sampling [8].

Benoit and Van den Poel (2012) developed quantile regression for dichotomus/binary response variables. They developed the regression using the Bayesian approach through the Metropolis-Hasting algorithm to estimate the parameters of binary quantile regression [2]. Alhamzawi et al. (2013) made improvements to previous journals by developing quantile binary regression with the addition of LASSO penalties using Bayesian analysis through gibbs sampling algorithms [1].

Many studies of statistical downscaling using quantile regression have been done. Wigena and Djuraidah (2014) estimates extreme rainfall using quantile regression with Principle Component Analysis as dimensional reduction [13]. Based on Zaikarina et al. (2016), quantile regression with LASSO penalty was better to predict extreme rainfall than quantile regression with ridge penalty [15]. Hendri et al. (2019) estimated extreme rainfall in West Java using Bayesian quantile regression. The results of the study showed that Bayesian quantile regression models are accurate enough to predict rainfall in West Java compared to the quantile regression model [6]. Previous research has used continuous response as dependent variable, so that this study will predict extreme rainfall using quantile regression with binary responses namely extreme rainfall ($y_i = 1$) and not extreme ($y_i = 0$) with LASSO penalty with the Bayesian approach. The purpose of this study is statistical downscaling modeling to predict extreme rainfall using Bayesian binary quantile regression with LASSO penalty.

II. Method

A. Data

The dependent data in this study is secondary data from 1981 to 2009. Monthly rainfall data in West Java from BMKG are grouped based on type of land, low, medium, and high. The rainfall data in the lowlands are from 12 stations, the medium land are from 3 stations, and the highlands are from 3 stations.

The independent variables in this study are GCM output data which is climate forecast system reanalysis (CFSR) monthly precipitation data with a grid size of 2.5º×2.5º of domain 5×8 grid based on the best domain [5]. The CFSR is a mathematical model that describes the global interaction between lands, oceans, and air issued by The National Centers for Environmental Prediction (NCEP) (https://rda.ucar.edu).

B. Procedure of Analysis

1. Grouping regions in West Java into three parts, lowland (0-200 masl), medium land (201-500 masl) and highland (>500 masl).
2. Grouping monthly rainfall in West Java becomes two types, extreme rainfall $y_i = 1$ and not extreme $y_i = 0$ using the measure of surprise based on the study by Manurung et al. (2018) to determine the threshold [10].
3. Binary rainfall data exploration.
4. Split the monthly rainfall data into two parts, which the 1981-2008 data as modeling data and 2009 data as validation data.
5. Predict monthly extreme rainfall using Bayesian binary quantile regression method [1] with the following models:

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \omega v_i + \psi\sqrt{\sigma v_i}u_i$$

Where $= \frac{1-2\tau}{\tau(1-\tau)}$, $\psi^2 = \frac{2}{\tau(1-\tau)}$, $\boldsymbol{v} = [v_1 \dots v_n]'$, $\boldsymbol{v} \sim \exp(\sigma)$, $u_i \sim \mathrm{N}(0,1)$, and $y_i = \begin{cases} 1, & \text{if } y_i^* \geq threshold \\ 0, & \text{else} \end{cases}$.

The LASSO penalty in quantile 0.70, 0.75, 0.80, 0.85, 0.90, 0.925, 0.95, 0.975, and 0.99 are used to obtain variables that are not multicollinearity with the following formula:

$$\min_{\beta \epsilon R} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$$

with $\lambda \geq 0$ is the lagrange multiplier, and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is l_1 LASSO penalty.

The estimate of the parameters is using the Bayes method with the MCMC (Markov Chain Monte Carlo) algorithm, gibbs sampling as follows:

a. Suppose the initiation value for $y_i^*, \beta_j, \eta^2, \zeta, \sigma, s_j, v_i, \delta$.

b. For the first iteration, do
   - Generate $y_i^{*(1)}, y_i^{*(1)} \sim \pi(y_i^* | \boldsymbol{\beta}, \eta^2, \zeta, \sigma, \boldsymbol{s}, \boldsymbol{v}, \delta, \boldsymbol{y})$
   - Generate $\beta_j^{(1)}, \beta_j^{(1)} \sim \pi(\beta_j | y^*, \eta^2, \zeta, \sigma, \boldsymbol{s}, \boldsymbol{v}, \delta, \boldsymbol{y})$
   - Generate $\sigma^{(1)}, \sigma^{(1)} \sim \pi(\sigma | y^*, \boldsymbol{\beta}, \eta^2, \zeta, \boldsymbol{s}, \boldsymbol{v}, \delta, \boldsymbol{y})$
   - Generate $\eta^{2(1)}, \eta^{2(1)} \sim \pi(\eta^2 | y^*, \boldsymbol{\beta}, \zeta, \sigma, \boldsymbol{s}, \boldsymbol{v}, \delta, \boldsymbol{y})$
   - Generate $s_j^{(1)}, s_j^{(1)} \sim \pi(s_j | y^*, \boldsymbol{\beta}, \eta^2, \zeta, \sigma, v, \delta, \boldsymbol{y})$
   - Generate $v_i^{(1)}, v_i^{(1)} \sim \pi(v_i | y^*, \boldsymbol{\beta}, \eta^2, \zeta, \sigma, s, \delta, \boldsymbol{y})$
   - Generate $\zeta^{(1)}, \zeta^{(1)} \sim \pi(\zeta | y^*, \boldsymbol{\beta}, \eta^2, \sigma, \boldsymbol{s}, \boldsymbol{v}, \delta, \boldsymbol{y})$
   - Generate $\delta^{(1)}, \delta^{(1)} \sim \pi(\delta | y^*, \boldsymbol{\beta}, \eta^2, \zeta, \sigma, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{y})$

c. Repeat steps 5b as many as m iterations.

d. Obtained examples that have a joint posterior distribution $\pi(y_i^*, \boldsymbol{\beta}, \eta^2, \zeta, \sigma, \boldsymbol{s}, \boldsymbol{v}, \delta | \boldsymbol{y})$

6. The model is evaluated based on accuracy values.

Predicted Values	Actual Values	
	Extreme (1)	Not Extreme (0)
Extreme (1)	A	B
Not Extreme (0)	C	D

Accuracy is the overall accuracy of the prediction that can be calculated with the following formula:

$$\text{Accuracy} = \frac{(A+D)}{(A+B+C+D)}$$

where,

A : Extreme rainfall predicted to be extreme
B : Not Extreme rainfall predicted to be extreme
C : Extreme rainfall predicted to be not extreme
D : Not extreme rainfall predicted to be not extreme

III. Result

A. Data Description

a. Rainfall

A description of rainfall data in West Java on low, medium, and highlands is presented in Figure 1.

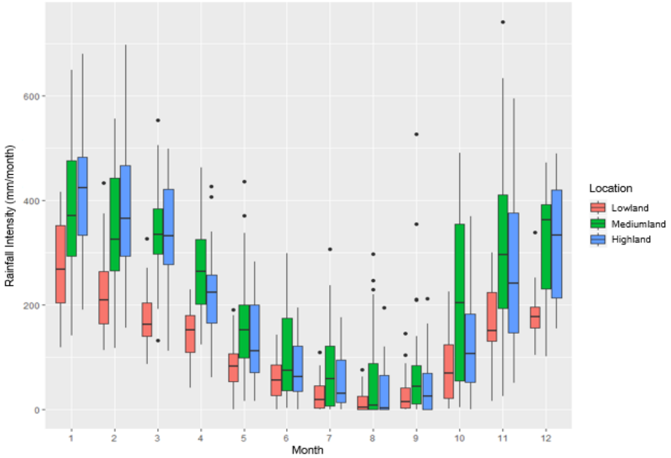


Figure 1: Boxplot of rainfall in West Java from 1981 to 2009

The pattern of rainfall of each land is U shaped. In the rainy season, the intensity of rainfall has an average greater than 150 mm/month [11]. The rainy season occurs between October - May in the lowlands of West Java, while in the medium and highland of West Java the rainy season occurs every month. The intensity of rainfall on medium and highland is higher than the intensity of rainfall in lowland. Extreme rainfall has a monthly rainfall intensity greater than 400 mm/month [4]. Based on that, extreme rainfall occurs in January and February in the lowland. In medium and highland, extreme rainfall occurs in October - April. The difference of altitude results in a difference in rainfall intensity [9], so the extreme values for extreme rainfall intensity in each land are different. Therefore, to categorize rainfall into extreme and not extreme categories is using the measure of surprise (MOS) based on the study by Manurung et al. (2018) to determine the threshold [10].

b. GCM Output

GCM output data are high dimension that can be multicollinearity between the variables, so it checked by looking at the variance inflation factor (VIF) value. Based on Table 1, there are variables that have a VIF value greater than 10. It indicates that there is a multicollinearity problem.

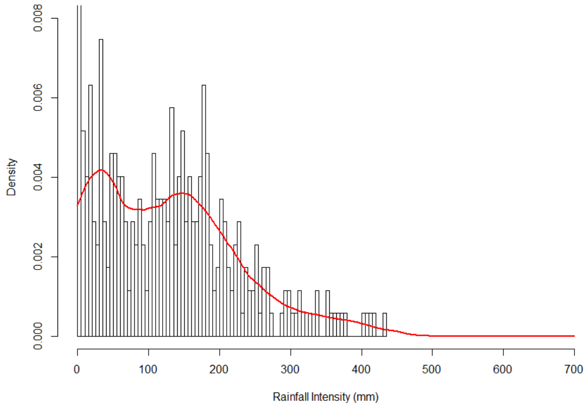
http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10702

Table 1: The VIF value of the GCM output data

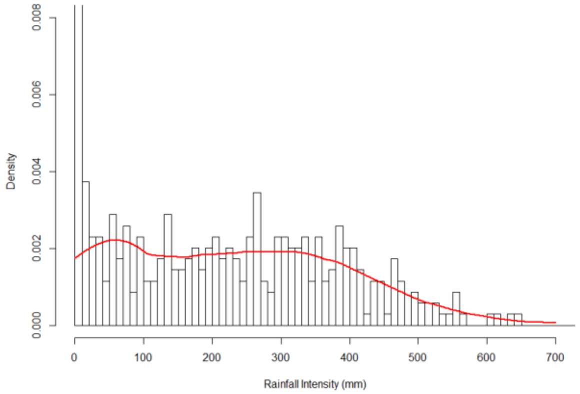
Variable	VIF	Variable	VIF	Variable	VIF	Variable	VIF
X1	2.22	X11	5.04	X21	13.99	X31	14.14
X2	3.10	X12	5.20	X22	12.33	X32	12.54
X3	3.09	X13	5.70	X23	8.32	X33	4.61
X4	2.58	X14	4.02	X24	5.52	X34	7.66
X5	2.65	X15	4.37	X25	5.41	X35	8.73
X6	3.29	X16	4.66	X26	8.29	X36	7.45
X7	2.98	X17	4.20	X27	8.00	X37	8.18
X8	3.44	X18	4.48	X28	10.20	X38	14.81
X9	3.80	X19	5.75	X29	10.27	X39	18.34
X10	4.63	X20	8.06	X30	7.68	X40	11.12

B. Measure of Surprise

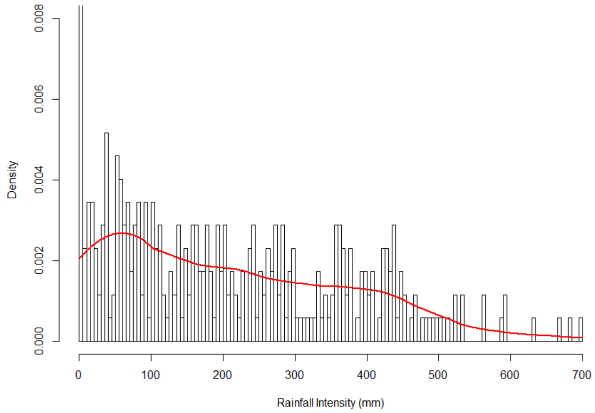
Extreme rainfall can cause the probability distribution of the rainfall to be far from Gaussian [12]. The shape of distribution below the threshold is difficult to determine from graph exploration. Histograms and plot density rainfall data for average rainfall at several stations on lands in West Java are presented in Figure 2.



(a)



(b)

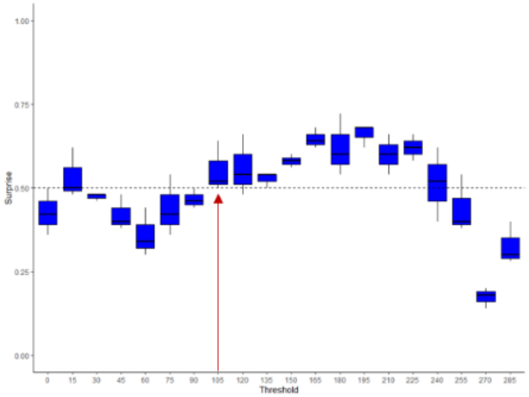


(c)

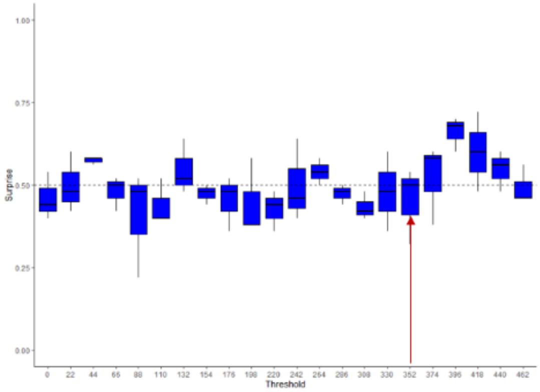
Figure 2: Histogram of rainfall (a) lowland, (b) medium land, (c) highland

The three histrograms have the same of a global turning point around 50-70 mm. The difference is at the local turning point on each land. Rainfall in the lowland has local turning point at 110 mm and 150 mm. In the medium land at 250 mm and 330 mm. In the highland at 210, 350 and 410 mm. These points potentially to be threshold values that will be used to categorize rainfall into extreme and not extreme in each land in West Java.

The measure of surprise (MOS) plot of rainfall for each land in West Java is in Figure 3. The MOS plot consists of several boxplots representing the number of threshold candidates on each lands. The dotted line indicates the lowest limit when the surprise starts to stabilize at 0.5.

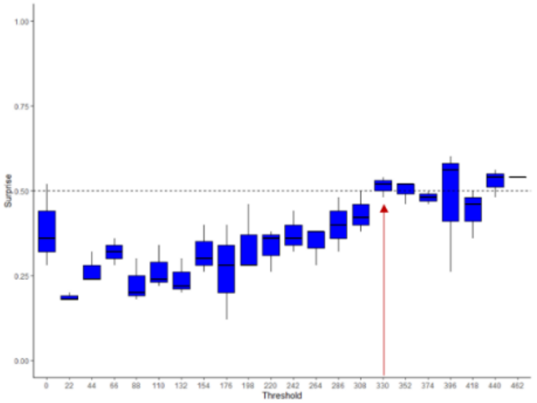


(a)



(b)

http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10702



(c)

Figure 3: Measure of Surprise plot (a) lowland, (b) medium land, (c) highland

The estimated threshold reaches 285 mm in lowlands, while in medium and highland 462 mm. The Generalized Pareto Distribution (GPD) in the lowlands starts at 105 mm, in a medium land 330 mm, and in highland 352 mm. The frequency of rainfall that has been converted into binary data above the threshold is 181 (52%) in the lowland, 97 (28%) in the medium land, and 81 (23%) in the highland. Then, this binary data is used to predict extreme rainfall in West Java using Bayesian binary quantile regression method with LASSO penalty.

C. Bayesian Binary Quantile Regression

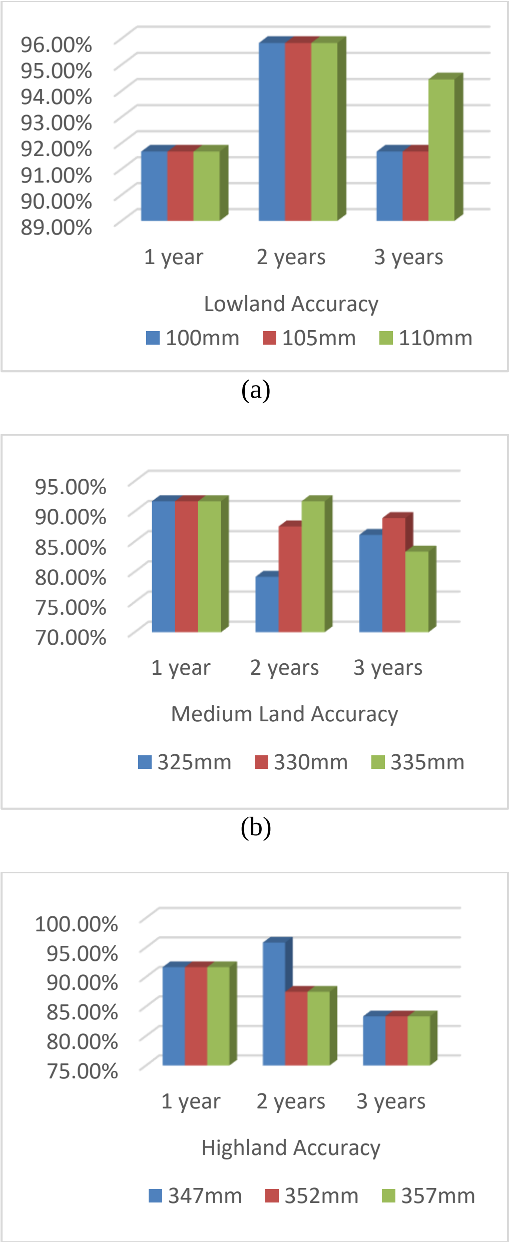
The Bayesian binary quantile regression models is on quantile 0.70, 0.75, 0.80, 0.85, 0.90, 0.925, 0.95, 0.975, and 0.99 in each land. The 9 quantile points of Bayesian binary quantile regression models is intended to obtain accuracy values. For a meaningful interpretation of the predicted probabilities, the results of at least 9 quantiles should be used [3]. In general, there are several variables eleminated from the model due to LASSO penalty. Table 2 is an example of the LASSO coefficients and variables selected for the lowland at quantile 0.75.

Table 2: LASSO coefficient and variables of BBQR in Q (0.75) in the lowland

Variable	Coefficient	Interval Credibility	
		Lower	Upper
X1	-11.45	-15.70	-2.18
X5	0.46	0.03	0.89
X9	7.81	2.16	9.68
X10	3.27	1.00	5.06
X12	2.48	1.49	3.11
X14	4.56	1.27	5.70
X16	2.70	1.52	3.72
X17	3.10	1.03	4.15
X19	3.61	1.70	4.37
X21	3.87	1.47	4.95
X26	3.06	1.26	3.60
X27	1.73	0.78	2.39
X29	3.51	1.57	3.96
X35	3.44	1.79	4.10
X37	2.96	0.94	5.44
X38	3.00	1.50	4.06
X40	4.68	1.50	6.18

D. Validation and Consistency of the Model

Validation is an important step because it reflects the accuracy of the prediction of the model. The consistency of Bayesian binary quantile regression on each land can be known from consistent predictions at different times. A comparison of accuracy values from the model on the low, medium and highland for 1, 2, and 3 years at some threshold points is presented in Figure 4.



(a)

(b)

(c)

Figure 4: Comparison of accuracy values from the model (a) lowland, (b) medium land, (c) highland

The diagram on lowland show that the accuracy values above 90% and seems consistent for the last 1, 2, and 3 years predictions, on medium land the accuracy values above 80% and seems good enough for the last 1, 2 and 3 years predictions, while on the highland the accuracy values above 83% and seems consistent enough for the last 1, 2, and 3 years predictions. The accuracy of the Bayesian binary quantile regression prediction for 1 year validation with different times is in Table 3.

Table 3: Accuracy value of BBQR Prediction

Training Data	Testing Data	Accuracy		
		Lowlands	Medium Land	Highlands
1981-2008	2009	91.67%	91.67%	91.67%
1981-2007	2008	100%	83.33%	83.33%
1981-2006	2007	91.67%	75%	83.33%
Mean		94.45%	83.33%	86.11%
Standard Deviation		0.048	0.083	0.048

Based on the validation and consistency above, the Bayesian binary quantile regression model for all lands has good accuracy. The accuracy value of predictions in lowland is higher than in medium and highland. The standard deviation value of the accuracy of the model is quite small, that is 0.048 for lowland and highland, and 0.083 for medium land. This shows that Bayesian binary quantile regression models on all lands are consistent enough to predict extreme rainfall for the next 1 year.

IV. Conclusion

Bayesian binary quantile regression model is accurate to predict extreme rainfall in West Java because it has average accuracy value above 83%. The threshold values from the MOS method is good enough to determine the category of extreme rainfall. Bayesian binary quantile regression models in lowlands are better than medium and highland. Bayesian binary quantile regression models are more accurate and consistent to predict the next 1 year.

References

[1] Alhamzawi R, Benoit DF, Yu K, “Bayesian Lasso Binary Quantile Regression,” *Journal of Computational Statistics*, vol. 28, pp. 2861–2873, 2013.

[2] Benoit DF, Van den Poel D, “Binary Quantile Regression: A Bayesian Approach Based on the Asymmetric Laplace Distribution,” *Journal of Applied Econometrics*, vol. 27, pp. 1174-1188, 2012.

[3] Benoit DF, Van den Poel D, “bayesQR: A Bayesian Approach to Quantile Regression,” *Journal of Statistical Software*, vol. 76, issue 7, pp. 1-32, 2017.

[4] Badan Meteorologi dan Geofisika [BMKG], Laporan Meterologi, Klimatologi dan Geofisika: Jakarta , 2008.

[5] Fadli N, Wigena AH, Djuraidah A, “Determination of General Circulation Model Grid Resolution to Improve Accuracy of Rainfall Prediction in West Java,” *International Journal of Scientific and Research Publication,* vol. 9(7), pp. 835–838, 2019.

[6] Hendri EP, Wigena AH, Djuraidah A, “Statistical Downscaling with Bayesian Quantile Regression to Estimate Extreme Rainfall in West Java,” *International Journal of Science: Basic and Applied Research (IJSBAR)*, vol. 47(2), pp. 142–151, 2019.

[7] Koenker R, Bassett G, “Regression Quantiles,” *Econometrica*, vol. 46(1), pp. 33–50, 1978.

[8] Kozumi H, Kobayashi G, “Gibbs Sampling Methods for Bayesian Quantile Regression,” *Journal of Statistical Computation and Simulation*, vol. 81(11), pp. 1565–1578, 2009.

[9] Manik TK, Rosadi B, Nurhayati E, “Mengkaji Dampak Perubahan Iklim Terhadap Distribusi Curah Hujan Lokal di Provinsi Lampung,” *Forum Geografi*, vol. 28(1), pp. 73–86, 2014.

[10] Manurung AM, Wigena AH, Djuraidah A, “GDP Threshold Estimation Using Measure of Surprise,” *International Journal of Science: Basic and Applied Research (IJSBAR)*, vol. 42(3), pp. 16–25, 2018.

[11] Pribadi HY, “Variabilitas Curah Hujan dan Pergeseran Musim di Wilayah Banten Sehubungan dengan Variasi Suhu Muka Laut Perairan Indonesia, Samudra Pasifik dan Samudra Hindia,” Tesis, 2012.

[12] Stephenson DB, Kumar KR, Doblas-Reyes FJ, Royer JF, Hauvin FC, “Extreme Daily Rainfall Events and Their Impact on Ensemble Forecasts of the Indian Monsoon,” *Monthly Weather Review*, vol. 127(9), pp. 1954–1966, 1999.

[13] Wigena AH, Djuraidah A, “Quantile Regression in Statistical Downscaling to Estimate Extreme Monthly Rainfall,” *Science Journal of Applied Mathematics and Statistics*, vol. 2(3), pp. 66-70, 2014.

[14] Yu K, Moyeed RA, “Bayesian Quantile Regresion,” *Statistics & Probability Latter*, vol. 54(4), pp. 437–447, 2001.

[15] Zaikarina H, Djuraidah A, Wigena AH, “Lasso and Ridge Quantile Regression using Cross Validation to Estimate Extreme Rainfall,” *Global Journal of Pure and Applied Mathematics*, vol. 12(3), pp. 3305–3314, 2016.

Authors

First Author – Putri Maulidina Fadilah, S.Si, college student, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor, 16680, Indonesia, putri_maulidina@apps.ipb.ac.id.
Second Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor, 16680, Indonesia, aji_hw@apps.ipb.ac.id.
Third Author – Dr. Ir. Anik Djuraidah, MS, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor, 16680, Indonesia, anikdjuraidah@apps.ipb.ac.id.

Correspondence Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor, 16680, Indonesia, aji_hw@apps.ipb.ac.id.

http://dx.doi.org/10.29322/IJSRP.10.11.2020.p10702