International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Statistical Downscaling with Bayesian Quantile Regression to Estimate Extreme Rainfall in West Java

Eko Primadi Hendri[a]*, Aji Hamim Wigena[b], Anik Djuraidah[c]

[a,b,c]*Department of Statistics, IPB University, Bogor, 16680, Indonesia*

[a]*Email: eko_primadihendri@apps.ipb.ac.id*

[b]*Email: ajiwigena@ymail.com*

[c]*Email: anikdjuraidah@apps.ipb.ac.id*

Abstract

West Java is one of the largest regions producing rice in Indonesia. Information on rainfall is very important for farmers to anticipate extreme events that can cause losses in agriculture. Extreme rainfall patterns can be modeled by Bayesian quantile regression. Parameters of the model are estimated by MCMC. This study used statistical downscaling to obtain relationship between global scale data and local scale data. The data were monthly rainfall data in West Java based on three type of land, low, medium, high land, and GCM output data. LASSO regularization was used to overcome multicollinearity problem in GCM output data. The purpose of this study was to model Bayesian quantile regression in each type of land. The Bayesian quantile regression model in the lowlands can predict extreme rainfall accurately and consistently in one year ahead.

Keywords: Statistical Downscaling; Bayesian Quantile Regression; LASSO; MCMC.

1. Introduction

West Java is one of the largest rice producers in Indonesia. In 2018, West Java produces 9.5 million tons of rice and the average monthly rainfall of 191.4 mm/month [2]. The productions can decrease due to flooding that occurs due to extreme rainfall [11].

---

* Corresponding author.

Extreme rainfall causes losses in agriculture [13]. Extreme rainfall is a condition of rainfall above the normal average or rainfall above 400 mm/month [1]. Therefore, rainfall prediction information is very useful for farmers to anticipate the possibility of extreme events, so the analysis is needed to obtain accurate rainfall predictions. Statistical downscaling (SD) is a technique that uses statistical methods to see global-scale data relationships with local-scale data. Global-scale data are represented by the global circulation model (GCM) output data. The characteristics of the GCM output data are high dimension and multicollinearity. Multicollinearity causes the estimated parameters for each model to be biased so that it can be overcome with the least absolute shrinkage and selection operator (LASSO). Quantile regression is a statistical technique is used to predict the relationship between dependent variables and independent variables in conditional quantile functions and can analyze a number of data in the form of asymmetric and not homogen [6]. Quantile regression can measure the effects of independent variables not only at the center of the distribution of data but also at the top and bottom of the distribution tail. This is very useful if extreme values are an important problem [5]. The parameter estimates from quantile regression are determined by the simplex method in linear programming [4]. The simplex method is not appropriate for obtaining parameter estimators from Laplace's asymmetric distribution, so that the parameter estimates using the Bayesian method. The Bayesian method uses sample information and prior distribution to get posterior distribution. When estimation based on posterior distribution is difficult to obtain analytically, a numerical Markov Chain Monte Carlo (MCMC) method is needed.

Many kinds of research on Bayesian quantile regression and SD have been conducted. Yu and Moyeed examined Bayes quantile regression with a likelihood function based on Laplace asymmetric distribution and the MCMC method specifically the Metropolis-hasting algorithm [16]. Kozumi and Kobayashi developed Bayes quantile regression with the Gibbs sampling algorithm [7]. Djuraidah and Wigena research on quantile regression to explore rainfall in Indramayu [5]. Mondiana discusses SD modeling with quantile regression using principal component analysis [9]. Santri researches quantile regression modeling at SD using LASSO [12]. Zakarina discusses about gulud quantile regression in SD [17]. Cahyani discusses about SD modeling with Elastic-net quantile regression [3].

Statistical downscaling modeling has a broad scope to be discussed. The problem limitation in this study for problem-solving is more focused. Therefore, this study only discusses statistical downscaling modeling with Bayesian quantile regression. The purpose of this study is statistical downscaling modeling with Bayesian quantile regression using LASSO to overcome multicollinearity and the MCMC method to estimate extreme rainfall in West Java.

2. Material and Method

2.1. Material

The research data is in the form of 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, as the dependent variable. 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 GCM output data is the monthly precipitation data Climate Forecast System Reanalysis (CFSR) with a grid size of $2.5^0 \times 2.5^0$ of domain 5×8 grids as independent variables. The CFSR is the model that describes a global

interaction between lands, oceans, and air. The CFSR issued by the National Centers for Environmental Prediction, NCEP, (https://rda.ucar.edu).

2.2. Method

1. West Java is grouped into three parts, lowland (0-200 masl), medium (201-500 masl), and high (more than 500 masl) [10].
2. Make a boxplot to detect extreme rainfall data based on point 1.
3. Rainfall data is divided into two parts. That are rainfall data from 1981 - 2008 as training data and rainfall data in 2009 as testing data.
4. The Bayesian quantile regression model [7] as follows:

$$y_i = x_i'\beta + pv_i + k\sqrt{\sigma v_i}u_i$$

Where $p = \frac{1-2\tau}{\tau(1-\tau)}$, $k^2 = \frac{2}{\tau(1-\tau)}$, $u \sim N(0,1)$, $v_i = [v_1 \dots v_n]'$ and $v \sim \exp(\sigma)$. LASSO Penalty in quantile 0.75, 0.90, and 0.95 are used to obtain variables that are not multicollinearity with the following formula:

$$\beta_\tau^{LASSO} = \min_{\beta \in R} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

with λ is the penalty parameter and $\sum_{j=1}^{p} |\beta_j| \leq \lambda, \lambda \geq 0$. MCMC, Gibbs sampling, is used to estimate parameters. The Gibbs sampling algorithm is as follows [8]:

i. Suppose the initiation value for β, v_i, σ are $\beta^{(0)}, v_i^{(0)}, \sigma^{(0)}$.

ii. For the first iteration, do

- Generate $\beta^{(1)}$, $\beta^{(1)} \sim \pi(\beta | v_i^{(0)}, \sigma^{(0)}, y)$
- Generate $v_i^{(1)}$, $v_i^{(0)} \sim \pi(v_i | \beta^{(0)}, \sigma^{(0)}, y)$
- Generate $\sigma^{(1)}$, $\sigma^{(0)} \sim \pi(\sigma | \beta^{(0)}, v_i^{(0)}, y)$

iii. Repeat steps 4.ii as many as m iterations.

iv. obtained examples that have a joint posterior distribution $\pi(\beta, v_i, \sigma | y)$

5. The model is evaluated based on root means square error of prediction (RMSEP)

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y)^2}$$

and the correlation between actual rainfall and estimated rainfall

$$r_{y\hat{y}} = \frac{n\sum_{i=1}^{n} y_i\,\hat{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \hat{y}_i)}{\sqrt{[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2][n\sum_{i=1}^{n} \hat{y}_i^2 - (\sum_{i=1}^{n} \hat{y}_i)^2]}}.$$

6. Validation and consistency of the model.

3. Result and Discussion

3.1. Data Description

3.1.1. Rainfall

Descriptions rainfall in low, medium and high lands are presented in Figure 1. Patterns of rainfall each land is about U. The rainy season has an average intensity of monthly rainfall greater than 150mm/month and the dry season has an average rainfall intensity of less than 150mm/month [14]. In the lowlands of West Java, the rainy season occurs in October-May and the dry season occurs in June-September. In the medium and high lands in West Java, the rainy season occurs every month.

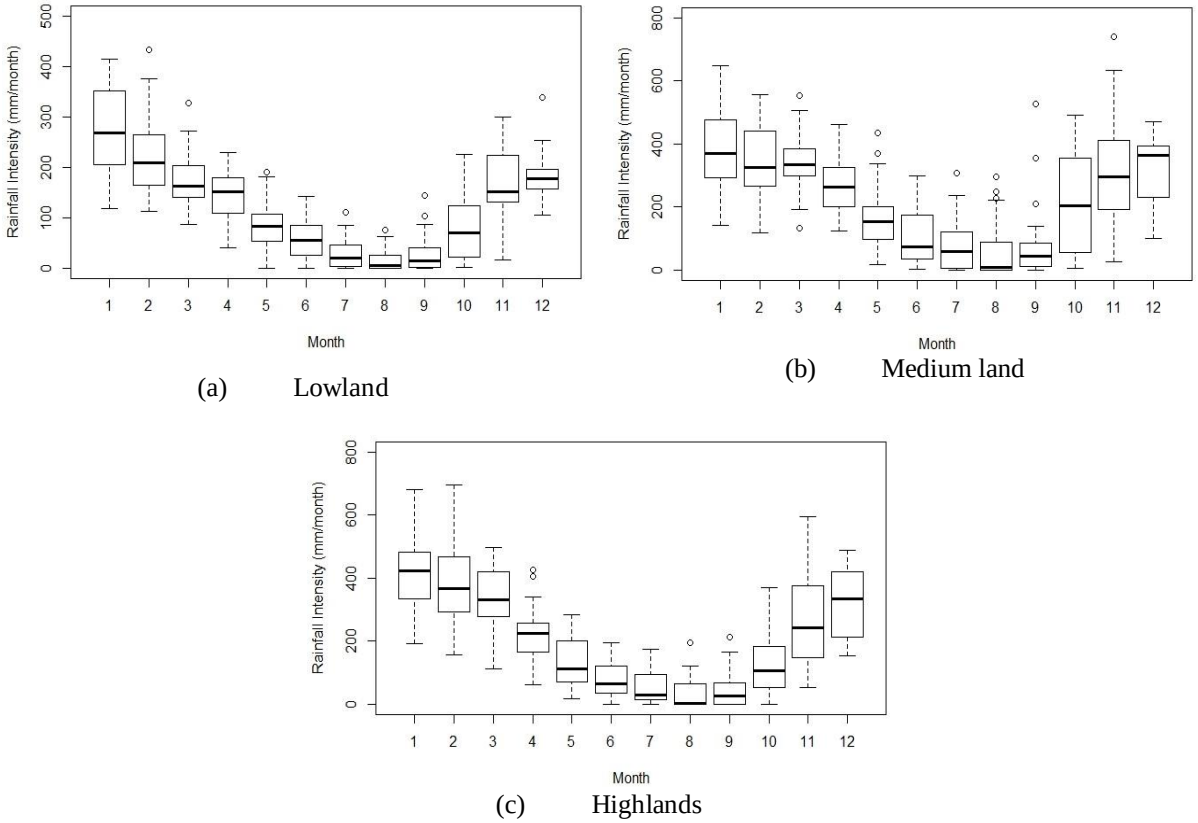


Figure 1: boxplot of the rainfall in West Java from 1981 to 2009

Extreme rainfall has a rainfall intensity greater than 400mm/month [1]. In lowlands, extreme rainfall occurs in January and February. In the medium land, extreme rainfall does not occurs in June - August. In the highlands, extreme rainfall occurs in November - April.

3.1.2. GCM output

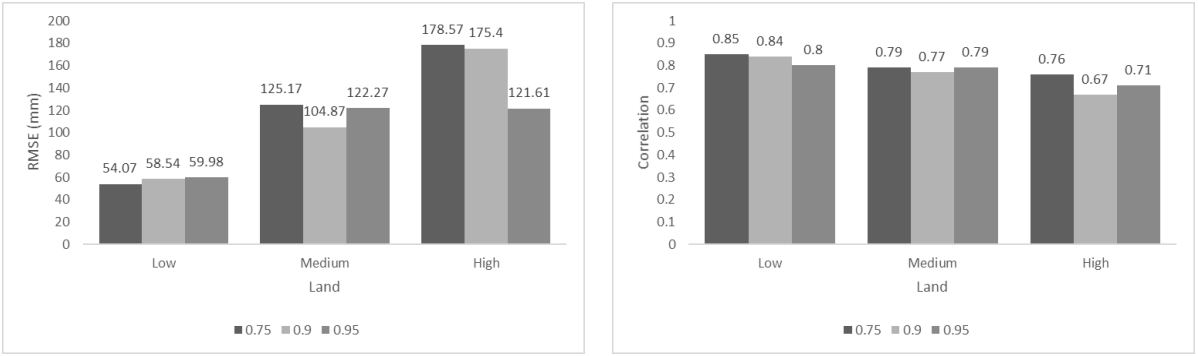
GCM output data are a high dimension so it is checked multicollinearity using the method of variance inflation factors (VIF). Based on Table 1, there are grids VIF value greater than 10. This shows that there is a multicollinearity problem.

Table 1: The VIF value of the GCM output data

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

3.2. Bayesian Quantile Regression

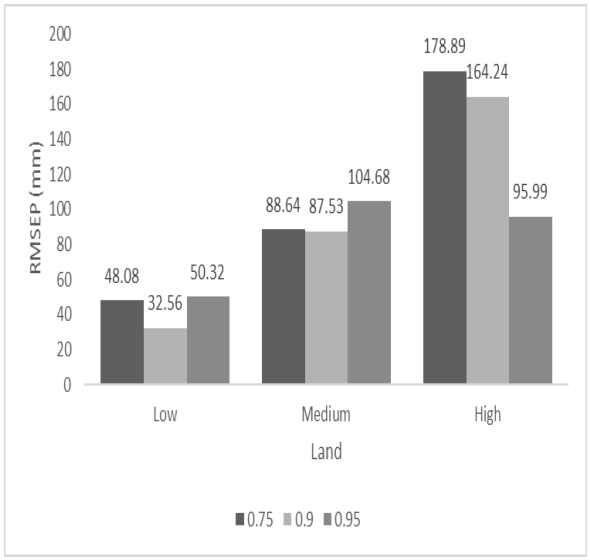
The Model of Bayesian quantile regression based on Kozumi and Kobayashi [7]. Each land uses the domain 5×8 grids. The best model is based on the smallest RMSE value and the biggest correlation. Figure 2 shows the comparison of the RMSE value and the correlation between the Bayesian quantile regression model on each land. The Bayesian quantile regression model in the lowlands is better than the model in the medium and high lands. This is because the Bayesian quantile regression model in the lowlands has the smallest RMSEP value and the largest correlation.



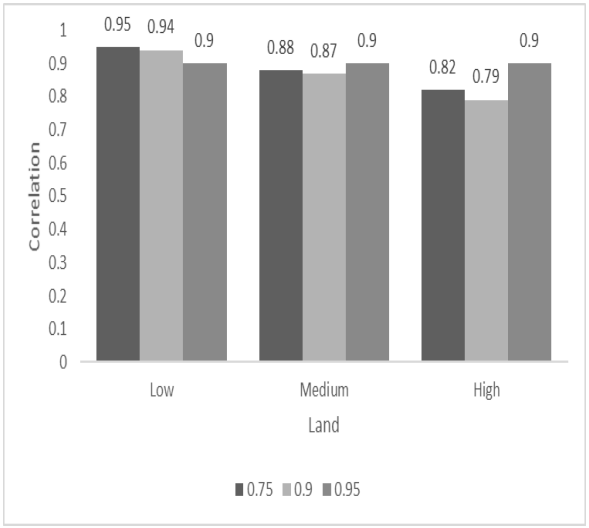
(a) RMSE (b) Correlation

Figure 2: Plots value of RMSE and correlation

Figure 3 shows a comparison of the values of RMSEP and correlation. The Bayesian quantile regression model in the lowlands is better than the Bayesian quantile regression model in the middle and high lands for predictions. This is because the Bayesian quantile regression model in the lowlands has the smallest RMSEP value and the largest correlation.



(a) RMSEP



(b) Correlation

Figure 3: Plots value of RMSEP and correlation

Figure 4 shows the value of rainfall predictions for each quantile with actual data in the lowlands. Extreme rainfall in January and February is in the rainfall predictions of Bayesian quantile regression. Extreme rainfall in January is in Q(0.75). Extreme rainfall in February is in Q(0.90).

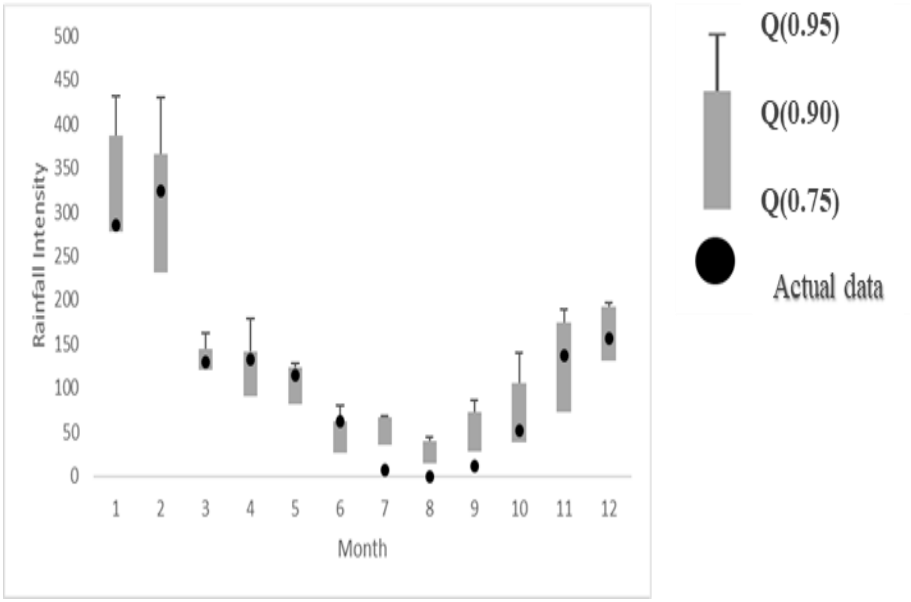


Figure 4: Plot of actual and predicted rainfall in the lowlands

Figure 5 shows the value of rainfall predictions for each quantile with actual data in the medium land. Extreme rainfall in January, February, March, April, May, and November is in the rainfall predictions of Bayesian quantile regression. Extreme rainfall in March and May are in Q(0.75). Extreme rainfall in January and November is in Q(0.90). Extreme rainfall in February and April is in Q(0.95). Figure 6 shows the value of rainfall predictions for each quantile with actual data in the highlands. Extreme rainfall in October is in the rainfall predictions of Bayesian quantile regression. Extreme rainfall in October is in Q(0.95).

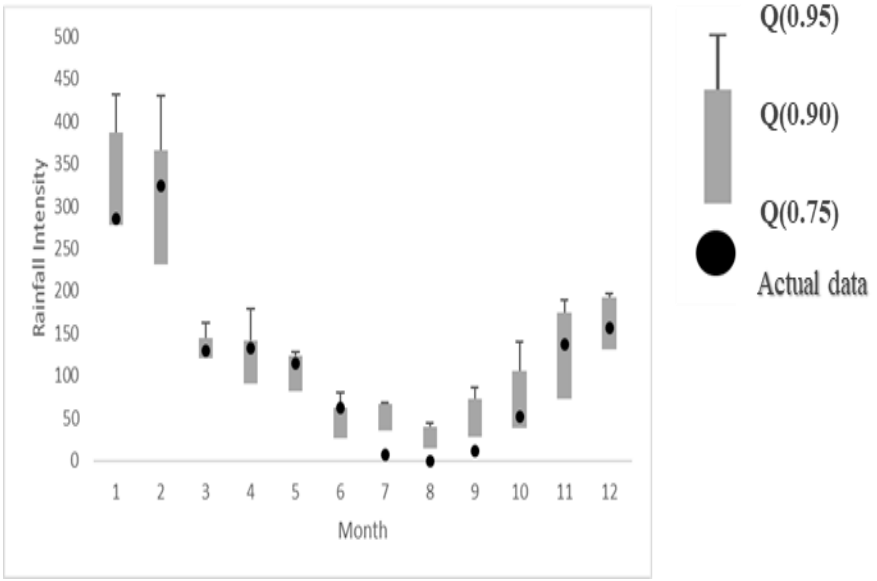


Figure 5: Plot of actual and predicted rainfall in the medium land

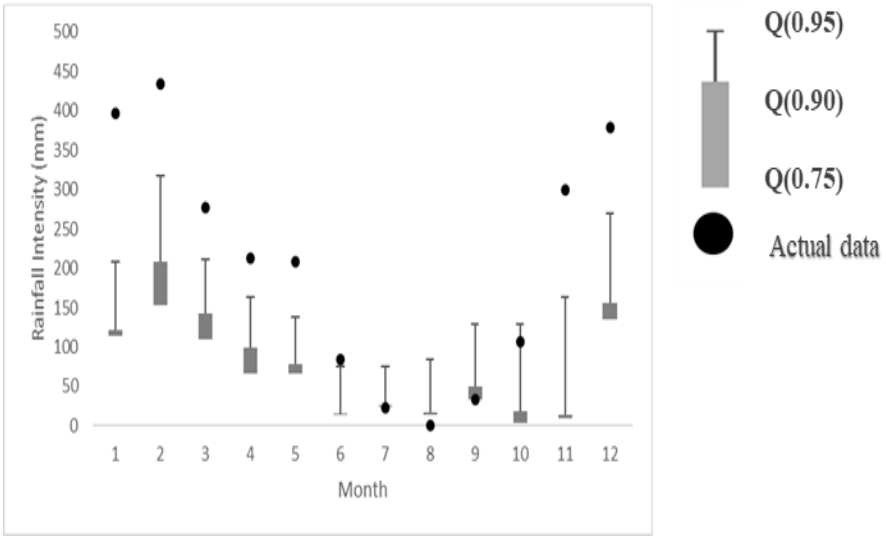


Figure 6: Plot of actual and predicted rainfall in the highlands

3.3. Validation and consistency of the model

Validation is an important step because it reflects the accuracy of the predictions of the model. Table 2 shows that accuracy of the Bayesian quantile regression model in the low, medium and high lands is better used to predict extreme rainfall for the next 1 year because it has the smallest RMSEP value and the largest correlation value.

The consistency of the model seen from the estimation results at different times. The model will give the best results if the relationship between dependent and independent variables is not much different if there is a change in time. The consistency of the model based on the standard deviation value of the correlation value in each estimation year. The smaller the standard deviation, then more consistent the model [15]. Based on Table 3, the standard deviation value of each model is small. The Bayesian quantile regression model of each land is

consistent to predict the rainfall for the next year.

Table 2: value of RMSEP and correlation for each different data length in each land

Training Data	Testing Data	Quantile	Lowlands		Medium Land		Highlands	
			RMSEP	Correlation	RMSEP	Correlation	RMSEP	Correlation
1981-2008	2009	0.75	48.08	0.95	88.64	0.88	178.89	0.82
		0.90	32.56	0.94	87.53	0.87	164.24	0.79
		0.95	50.32	0.90	104.69	0.90	95.99	0.90
1981-2007	2008-2009	0.75	42.88	0.94	115.52	0.86	167.38	0.83
		0.90	34.07	0.96	106.20	0.87	209.14	0.71
		0.95	53.70	0.92	119.78	0.87	119.79	0.88
1981-2006	2007-2009	0.75	58.61	0.94	145.02	0.85	197.73	0.74
		0.90	68.51	0.85	85.61	0.85	177.77	0.62
		0.95	93.75	0.86	105.94	0.80	128.48	0.75
1981-2005	2006-2009	0.75	108.89	0.84	179.33	0.77	166.10	0.79
		0.90	81.02	0.85	107.09	0.80	197.63	0.68
		0.95	58.32	0.89	100.43	0.82	111.08	0.82

Table 3: Value of Correlation and standard deviation

Training Data	Testing Data	Quantile	Correlation		
			Lowlands	Medium Land	Highlands
1981-2008	2009	0.75	0.95	0.88	0.82
		0.90	0.94	0.87	0.79
		0.95	0.90	0.90	0.90
1981-2007	2008	0.75	0.94	0.86	0.81
		0.90	0.96	0.87	0.82
		0.95	0.94	0.86	0.87
1981-2006	2007	0.75	0.93	0.89	0.74
		0.90	0.84	0.90	0.63
		0.95	0.86	0.83	0.71
1981-2005	2006	0.75	0.84	0.78	0.83
		0.90	0.92	0.83	0.79
		0.95	0.94	0.87	0.83
Standard Deviation		0.75	0.05	0.05	0.04
		0.90	0.06	0.03	0.09
		0.95	0.04	0.03	0.09

4. Conclusion

The Bayesian quantile regression model in the low, medium and high lands can predict extreme rainfall more accurate and consistent for the next year. The Bayesian quantile regression model in the lowlands has a better model than the Bayesian quantile regression model in the medium and high lands.

5. Suggestion

In this study, statistical downscaling with Bayes quantile regression uses LASSO penalty. For further research, statistical downscaling modeling with Bayes quantile regression uses principal component analysis to compare predictions from both methods.

Acknowledgements

This work is fully supported by Kemenristek DIKTI (Kementrian Riset Teknologi dan Pendidikan Tinggi) of Indonesia.

References

[1]. Badan Metorologi, Klimatologi dan Geofisika (BMKG). Laporan Meteorologi, Klimatologi, dan Geofisika : Jakarta, 2008

[2]. Badan Pusat Statitik. Ringkasan Eksekutif, Luas Panen dan Produksi Padi di Indonesia 2018 : Jakarta, 2018.

[3]. Cahyani, T.B.N. “Statistical Downscaling Modelling with Ridge and Elastic-net Regularized Quantile Regression for Rainfall Prediction in Indramayu.” M.Sc. Thesis, IPB University, Bogor, 2016.

[4]. Chen. C and Wei. Y. “Computation Issues For Quantile Regression.” The Indian Journal Statistics, vol. 67(2), pp. 399-417, 2005.

[5]. Djuraidah. A. and Wigena. A. H. “Regresi kuantil untuk Eksplorasi Pola Curah Hujan di Kabupaten Indramayu”. Jurnal Ilmu Dasar. vol. 12(1), pp. 50-56, 2011.

[6]. Koenker. R. and Bassett. G. “Regression Quantile”. Econometrica. vol. 46(1), pp. 33-50, 1987.

[7]. Kozumi. H. and Kobayashi. G. “Gibbs Sampling Methods for Bayesian Quantile Regression”. Journal of Statistical Computation and Simulation. vol. 81(11), pp. 1565-1578, 2011.

[8]. Geman. S. and Geman. D. “Stochastic relation, gibbs distribution, and the bayesian restoration of image”. IEEE Transaction on Pattern Analisys and Machine Intelligence. vol. 6, pp. 721-741. 1984.

[9]. Mondiana. Y.Q. “Statictical Downscaling Modeling with Quantile Regression to Estimate Extreme Precipitation (A Case Study in Bangkir Station, Indramayu).” M.Sc. Thesis, IPB University, Bogor, 2012.

[10]. Nuryanto. B. et al. “Pengaruh Tinggi Tempat dan Tipe Tanaman Padi Terhadap Keparahan Penyakit Hewan Pelepak”. Jurnal. Penelitian Pertanian Taman Pangan. vol. 33(1), 2014.

[11]. Rosyidie. A. “Banjir: Fakta dan Dampaknya, Serta Pengaruh Guna Lahan. Journal Perencanaan Wilayah dan Kota”. vol. 24(3), pp. 241-249, 2013.

[12]. Santri. D. "Statistical Downscaling Modeling with Quantile Regression using LASSO to Estimate Extreme Rainfall." M.Sc. Thesis, IPB University, Bogor, 2016.

[13]. Suciantini. "Interaksi Iklim (Curah Hujan) Terhadap Produksi Tanaman Pangan di Kabupaten Pacitan," in Proc. Seminar Nasional Masyarakat Biodiversitas Indonesia, 2015, pp. 358-365.

[14]. Pribadi. H.Y. "Variabilitas Curah Hujan dan Pergeseran Musim di Wilayah Banten Sehubungan dengan Variasi suhu Muka Laut Perairan Indonesia, Samudra Pasifik dan Samudra Hindia." M.Sc. Thesis, Universitas Indonesia, Depok, 2012.

[15]. Wigena. A.H. "Statistical Downscaling Modeling using projection Pursuit Regression to Forecash Monthly Rainfall." Dr. dissertation, IPB University, Bogor, 2006.

[16]. Yu. K. and Moyeed. R.A. "Bayesian Quantile Regresion". Statistics & Probability Latter. vol. 54(4), pp. 437-447, 2001.

[17]. Zakarina. H. "Lasso and Ridge Quantile Regression using Cross Validation to Estimate Extreme Rainfall." M.Sc. Thesis, IPB University, Bogor, 2016.