Applied Mathematical Sciences, Vol. 9, 2015, no. 126, 6263 - 6272
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.57502

Quantile Spline Regression on Statistical Downscaling Model to Predict Extreme Rainfall in Indramayu

Noor Ell Goldameir

Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Anik Djuraidah and Aji Hamim Wigena

Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Abstract

Statiscal Downscaling (SD) was a method to model the relationship between local scale data as response variable and global scale data as predictor variable. The response variable was the rainfall and the predictor variable was the global circulation model (GCM) ouput. In general, GCM output had a large dimension and multicollinearity so the principal component analysis was used to solve this problem. This study modeled the rainfall with the selected principal component using quantile regression. This functional relationship could be parametric or nonparametric relationship. In this case, nonparametric functional relationship used spline to accommodate the extreme value with the quantile regression. The result showed that the quantile spline model was better than the quantile polynomial regression model especially in predicting the extreme values in the 90^{th} and 95^{th} quantile with correlation values of 0.95 and 0.93.

Keywords: Global circulation model, statistical downscaling, quantile regression and spline

1. Introduction

Rainfall was one of the climate element with the highest fluctuation and was the most dominant element that characterized the climate in Indonesia. Rainfall had a great effect in agriculture especially the extreme rainfall. Extreme rainfall resulted in flood and low rainfall resulted in dryness causing lower production of rice. Therefore, the analysis that examine the extereme event was needed to acquire the predicted information of rainfall in order to lessen the effect of the extreme climate.

Rainfall prediction needed a model which was statistical downscaling (SD). The model determine the functional relationship between GCM output and rainfall data [5]. GCM output potentially stimulated the climate in the past, in the present, and could predict the climate that might occur in the future [15].

The main problem of the SD model was finding the statistical method that could describe the relationship between the response variable and the predictor variable [11]. Statistical method was developed from the parametric, nonparametric, and semiparametric approaches. The study of the SD modeling with parametric approach that was used was the principal component regression [10]. The study of the SD modeling with nonparametric approach that was used was spline multivariate additive regression [11]. The study of the SD modeling with semiparametric that was used was penalty spline regression (P-sline) with mixed linear model approach [14]. However, the SD modeling predicted the average rainfall.

Some studies of the SD modeling that discussed about the extreme were quantile regression [8] [9] [13] and block maxima [6]. The type of the functional relationship in the quantile regression could be developed into nonparametric type. The best type of the functional relationship between rainfall and GCM was the spline [14]. Quantile spline regression had been used to predict the air pollution in Surabaya [1]. This study examined the SD modeling using the quantile spline regression to predict the extreme rainfall in Indramayu.

2. Literature Review

2.1 Statistical Downscaling

Statistical downscaling (SD) used statistical model in describing the relationship between GCM output (global scale) and rainfall data (local scale) to translate the global scale anomalies to anomalies of some local climate variable [16]. The basic idea of SD were finding the relationship between global scale climate parameter and local scale climate parameter and using this relationship for projection the GCM output simulation result in the past, in the present, or in the future globally. GCM output was a computer based model that consisted of some numeric and deterministic equation that was integrated and followed physics principles. GCM output simulated global climate variables on every grid (the grid size is $2{,}5^0 \times 2{,}5^0$) or every atmosphere layer, henceforth was used to predict the climate pattern in long period of time (annually) [12].

In general SD model could be seen in the following equation (1).

$$\boldsymbol{y}_{t\times 1} = f(\boldsymbol{X}_{t\times g}) \tag{1}$$

with $\boldsymbol{y}_{t\times 1}$ was the local climate variable (e.g. rainfall), $\boldsymbol{X}_{t\times g}$ was GCM output variable (e.g. precipitation), t was the amount of the time (e.g. monthly), and g was the amount of the GCM output grid domain. SD model would give the best result on condition that the relationship between response and predictor variables had to have high correlation in order to explain the variety of the local climate well, predictor variable had to be simulated well by GCM output, and the relationship between response and predictor variables were consistent for a long time period and the climate change in the future [3].

2.2 Spline

Spline was continuing piecewise polynomial so it could describe the characteristic of local data. The joint of broken line in which the characteristics change occurred in different interval was also known as knot. The number of knots used needed to be defined in advance by trying all knot combination that might be determined manually. In general the spline function of p degree could be presented in equation (2) [4].

$$s(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} u_{pk}(x-\kappa_k)_+^p \tag{2}$$

with $\boldsymbol{\beta} = \left(\beta_0, \cdots, \beta_p, u_{p1}, \cdots, u_{pK}\right)'$ was the spline coefficient vector with $p \geq 1$ are positive integer, $(w)_+^p = w^p \boldsymbol{I}\ (w \geq 0)$ was truncated power function (TPF), and $\kappa_1, \cdots, \kappa_K$ were the knots. The degree form of spline function consisted of degree 1 (linear), 2 (quadratic) and 3 (cubic).

2.3 Quantile Spline Regression

Quantile regression was a statistical method used for estimating the relationship between the response variable and predictor variable on quantile function in specific condition. Quantile regression could measure the effect of predictor variables not only in the center of the data distribution, but also in the tails of distribution. This method is very useful in the application, especially when extreme values were important issue [2].

In general the quantile regression model could be presented in equation (3).

$$\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

with $\boldsymbol{y} = (y_1, \ldots, y_n)'$ was the sized response vector $(n \times 1)$, $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})'$ was the sized predictor matrix $(n \times p)$, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)'$ was the sized parameter vector $(p \times 1)$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ was the sized error vector $(n \times 1)$. Parameter estimation β was estimated by minimizing the sum of squared errors, τ weighted for positive errors and $(1-\tau)$ for negative errors. Functional relationship in quantile regression was functional relationship between the conditional quantile forming linear function that could be shown in the equation (4).

$$Q(\tau|X = x) = \boldsymbol{x}'\boldsymbol{\beta}(\tau) \quad (4)$$

In general, according to [7] estimator of τ-order quantile regression for $\tau \in (0,1)$ was the problem solution of minimizing function of the equation (5).

$$\min_{\boldsymbol{\beta}\in\mathbb{R}} \left[\sum_{i\in\{i:y_i\geq \boldsymbol{x}_i'\boldsymbol{\beta}\}}^{n} \tau|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| + \sum_{i\in\{i:y_i<\boldsymbol{x}_i'\boldsymbol{\beta}\}}^{n}(1-\tau)|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|\right] \quad (5)$$

The solution of equation (5) was denoted $\hat{\beta}(\tau)$. Testing of parameter $\hat{\beta}$ for each quantile is using t-test with hypotheses.

$$H_0: \beta_k(\tau) = 0$$
$$H_1: \beta_k(\tau) \neq 0$$

with $k = 1,2,\cdots,r$. t-test statistics could be prensented in the equation (6).

$$t_{hit}(\tau) = \frac{\hat{\beta}_k(\tau)}{s\left(\hat{\beta}_k(\tau)\right)} \quad (6)$$

with $\hat{\beta}_k(\tau)$ was parameter $\hat{\beta}$ of-k order on quantile of τ-order and $s\left(\hat{\beta}_k(\tau)\right)$ was the standard deviation of parameter $\hat{\beta}$ on quantile of τ-order. H_0 was rejected if $|t_{hit}(\tau)| > t_{\left(\frac{\alpha}{2};n-k-1\right)}$. Quantile spline regression was the quantile regression in the form of a functional relationship spline. The general model of the quantile spline regression model with $0 < \tau < 1$ can be presented in the following equation (7).

$$Q(\tau|X = x) = s(x,\tau) = y_\tau = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} u_{pk}(x-\kappa_k)_+^p \quad (7)$$

3. Methodology

3.1 Data

The data of this study were the rainfall data in Indramayu used as the response variable and GCM-lag (precipitation) of the Climate Model Inter comparison Project (CMIP5) used as the predictor variables in KNMI Netherland. GCM-lag gave a better rainfall estimation result [10]. The domain size was 8×8 (64) grids located at the position of 16.25^0north latitude to 1.25^0 south latitude and 98.75^0west longtitude to 116.25^0 east longtitude above the area of Indramayu. The 8×8 grids domain size above the area of Indramayu gave more consistent result and was not sensitive to outliers [12]. The Climate data (rainfall and GCM-lag) was a monthly climate data from 1979-2008. In this study, data was divided into two parts, the data in 1979-2007 for modeling and the data in 2008 for model validation.

3.2 Methods

The analysis methods of this study were as follows:

1. Identifying multicollinearity based on the variance inflation factor (VIF) in GCM-lag data.
2. Reducing the dimension of GCM-lag data using principal component analysis (PCA).

3. Determining the functional relationship pattern between the rainfall and the selected principal components (PCs) using the minimum criteria of generalized cross validation (GCV) in the equation (8).

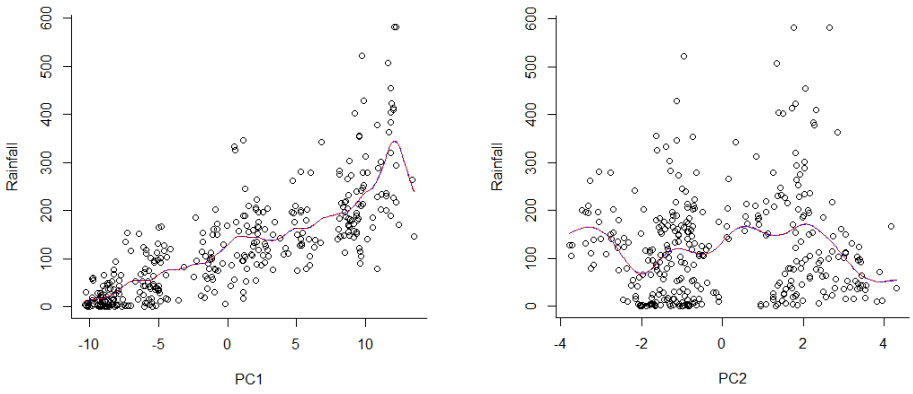
$$GCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2}{\frac{1}{n}\left(tr(I-S)\right)^2} \quad (8)$$

4. Determining TPF basis using cubical spline degree according to step 3.
5. Modeling using quantile regression with the spline functional relationships in various quantile values ($\tau = 50, 75, 90$ and 95).
6. Choosing the best model based on the criteria of determination coefficient (R^2) and root mean square error ($RMSE$) value.
7. Predicting the best model based on the criteria of correlation and root mean square error of prediction ($RMSEP$) value.

4. Results and Discussion

4.1 Data Exploration

This study was begun with identifying multicollinearity. There were 62 of 64 grids with VIF values more than 10. It indicated the multicollinearity problem in GCM-lag data. PCA was performed to reduce the data dimension. The amount of the principal component (PC) which had root characteristics greater than one were four PCs. The total variance proportion of PC1 to PC4 was 95%. It showed that PC1 to PC4 was able to explain 95% proportion of total variance of 64 predictor variables. The relationship patterns between rainfall and PC1 to PC4 were conducted with various possible number of degree of freedom. The optimum number of degree of freedom was determined by using the criteria of minimum GCV. The results showed that the optimum degrees of freedom based on the value of the minimum GCV for the selected PC1 to PC4 were 18, 11, 9 and 7. The plot between the rainfall and PC1 was closed to linear pattern. And then, the plot between the rainfall and PC3 was closed to quadratic pattern. And then, the plots between the rainfall and PC2 and PC4 did not tend to make a specific pattern (Figure 1). The number of spline knot in a model depended on the number of parameter and degree the basic model. The cubical model with the spline combination knot of 14, 8, 7, 5 were the best spline combination knot [14].



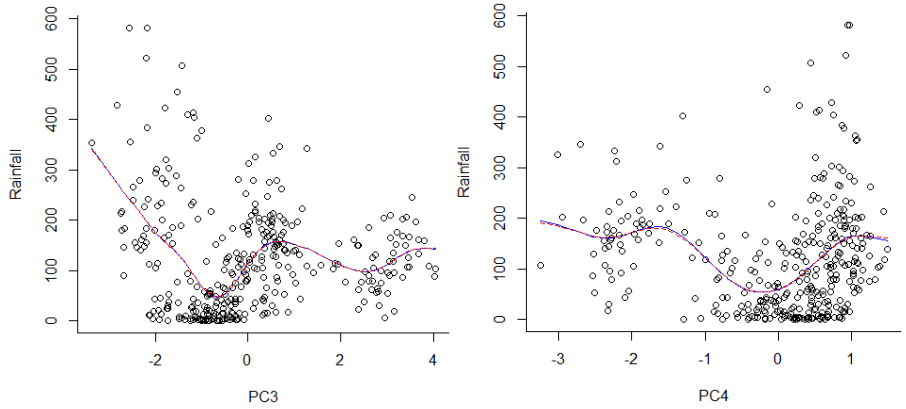


Figure 1 The functional relationship plot between the rainfall and PC1to PC4

4.2 Statistical Downscaling Modeling

4.2.1 Quantile Polynomial Regression Model

R^2, r, $RMSE$ and $RMSEP$ values on quantile polynomial regression (QPR) model tended to increase with the increasing of quantile values (Tabel 1 and Figure 4). R^2, r values especially for the extreme values did not show significant differences or in other words the values were almost equal. RMSE and RMSEP values for the extreme value were $RMSE_{90} = 110.832$, $RMSE_{95} = 145.928, RMSEP_{90} = 105.430$ and $RMSEP_{95} = 140.738$.

In general, QPR model could predict the intensity of rainfall well (Figure 2). The monthly prediction of rainfall in Indramayu from January to December could follow the pattern of actual data well. In the dry season (April to September) the estimate values for the 50th, 75th, 90th and 95th quantiles were higher than the actual values and could follow the actual pattern well. In the rainy season (October to March), especially February had the highest rainfall intensity. The actual value could be estimated well by the prediction of the 95th quantile. However, the estimate values of October to January and March had higher prediction value than the actual value. The estimate value in February was closer to the actual value.

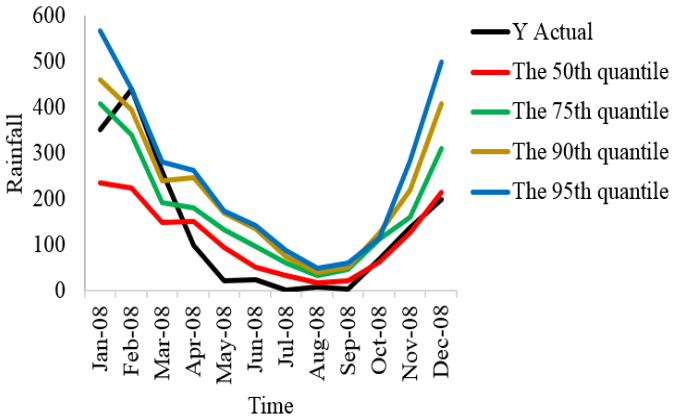


Figure 2 The plot between the actual rainfall values and QPR rainfall model of the 50th, 75th, 90th and 95th quantiles in 2008

4.2.2 Quantile Spline Regression Model

Quantile spline regression (QSR) model in Table 4 and Figure 10 showed that R^2, r, $RMSE$ and $RMSEP$ values on the quantile spline regression (QSR) model tended to increase following the increasing of quantile value. R^2 and r values between QPR and QSR models for the 50th, 75th, 90th and 95th quantiles did not showed significant differences or in other words the values were almost equal. However, $RMSE$ and $RMSEP$ values on QSR model were lower than on the QPR model especially for the extreme values ($RMSE_{90} = 110.679$, $RMSE_{95} = 136.433, RMSEP_{90} = 93.000$ and $RMSEP_{95} = 128.791$). $RMSE$ value showed that there was a little difference between the estimate and the actual values, which meant that the established model became more accurate in generating the estimated value of the extreme rainfall and the pattern became more similar to the actual data.

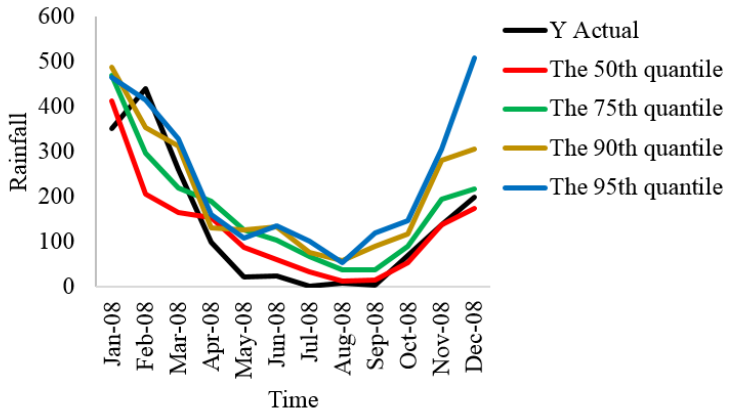


Figure 3 The plot between the actual rainfall value and the QSR rainfall model of the 50th, 75th, 90th and 95th quantiles in 2008

Table 1 The value of determination coefficient (R^2) and correlation (r) of the SD models

Model		R^2	r
Quantile polynomial regression (QPR)	The 50th quantile	63.63%	0.894
	The 75th quantile	94.62%	0.974
	The 90th quantile	98.62%	0.992
	The 95th quantile	98.09%	0.993
Quantile spline regression (QSR)	The 50th quantile	66.21%	0.827
	The 75th quantile	96.51%	0.993
	The 90th quantile	93.15%	0.948
	The 95th quantile	93.85%	0.931

The QSR model could predict rainfall intensity well (Figure 3). In general, the prediction pattern of the rainfall on the 50th, 75th, 90th and 95th quantiles was able to follow the pattern of the actual data well. The actual value on February could be predicted well above the 95th quantile which was the highest rainfall occurrence. In the rainy season, October to January and March had higher estimate values than the actual values. Estimate value in February was also close to the actual value. Furthermore, in the dry season the estimate values of each quantiles were higher than the actual values, but were able to follow the pattern well.

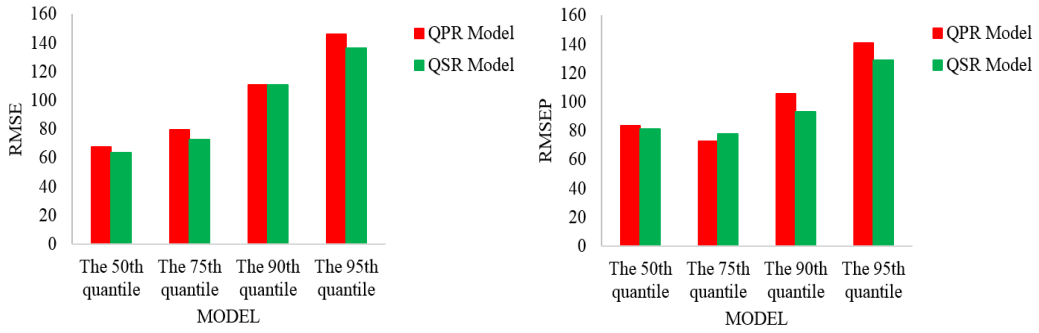


Figure 4 The plot of the RMSE value and the RMSEP value on QPR and QSR models of the 50th, 75th, 90th and 95th quantiles in 2008

5. Conclusions

Based on the analysis result of the extreme rainfall in Indramayu, statistical downscaling technique on the GCM data could be used to predict the extreme rainfall well using the quantile spline regression model. The estimate value of rainfall on Februari could be predicted at extreme value which described the highest actual rainfall value. The rainfall prediction that was made for one year ahead delivered good and consistent result.

References

[1] A. Djuraidah, L. Rahman, Quantile Spline Regression for Modeling the Extreme Value on PM_{10} Pollutant in Surabaya (Proceedings of the IX National Statistics Seminar), Sepuluh November Technology Institute (in Indonesian), Indonesia, 2009.

[2] A. Djuraidah, A. H. Wigena, Quantile Regression for Rainfall Exploration in Indramayu, *Journal of Basic Science*, **12** (2011), no. 1, 50-56.

[3] A. Busuioc, D. Chen, C. Hellstrom, Performance of statistical downscaling models in GCM validation and regional climate change estimates:

Application for Swedish precipitation, Int, J. Climatology, **21** (2001), 557-578. http://dx.doi.org/10.1002/joc.624

[4] R. Eubank, *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, 1988.

[5] E. Fernandez, On The Influence of Predictors Area in Statistical Downscaling of Daily Parameters, Onslo: Norwegian Meteorogical Institute, Report 9 (2005), 1-21.

[6] L. Handayani, A. H. Wigena, A. Djuraidah, Statistical Downscaling with Generalized Additive Model For Extreme Rainfall Estimation, *IOSR Journal of Mathematics*, **10** (2014), 21-25. http://dx.doi.org/10.9790/5728-10352125

[7] R. Koenker, *Quantile Regression*, Cambridge: Cambridge Iniversity Press, 2005. http://dx.doi.org/10.1017/cbo9780511754098

[8] YQ. Mondiana, Statistical Downscaling Modeling with Quantile Regression to Estimation Extreme Rainfall (case study of Bankir Station in Indramayu District), [Thesis], Bogor Agricultural University (in Indonesian), Indonesia, 2012.

[9] WJ. Sari, 2015, Statistical Downscaling Modeling with Quantile Regression of Principal Component Fungsional to Predict Rainfall, [Thesis], Bogor Agricultural University (in Indonesian), Indonesia, 2015.

[10] S. Sahriman, A. Djuraidah, A. H. Wigena, Application of principal component regression with dummy variable in statistical downscaling to forecast rainfall, *Open Journal of Statistics*, **4** (2014), 678-686. http://dx.doi.org/10.4236/ojs.2014.49063

[11] Sutikno, Statistical Downscaling of GCM Output and Its Application for Estimating Rice Prediction, [Dissertation], Bogor Agricultural University (in Indonesian), Indonesia, 2008.

[12] A. H. Wigena, Modeling of Statistical Downscaling using Projection Pursuit Regression for Forecasting Monthly Rainfall, [Dissertation], Bogor Agricultural University (in Indonesian), Indonesia, 2006.

[13] A. H. Wigena, A. Djuraidah, Quantile regression in statistical downscaling to estimate extreme monthly rainfall, *Science Journal of Applied Mathematics and Statistics*, **2** (2014), no. 3, 66-70. http://dx.doi.org/10.11648/j.sjams.20140203.12

[14] A. H. Wigena, A. Djuraidah, A. Rizki, Semiparametric Modeling in Statistical Downscaling to Predict Rainfall, *Applied Mathematical Sciences*, **9** (2015), no. 88, 4371-4382. http://dx.doi.org/10.12988/ams.2015.54362

[15] RL. Wilby et al., A review of climate risk information for adaptation and development planning, *International Journal of Climatology*, **29** (2009), 1193-1215. http://dx.doi.org/10.1002/joc.1839

[16] E. Zorita, H.V. Storch, The analog method as a simple statistical downscaling technique: comparison with more complicated methods, *Journal Climate*, **12** (1999), 2474-2489. http://dx.doi.org/10.1175/1520-0442(1999)012<2474:tamaas>2.0.co;2

Received: August 10, 2015; Published: October 14, 2015