

## **Statistical Downscaling with Generalized Additive Model For Extreme Rainfall Estimation**

Lilies Handayani<sup>1</sup>, Aji Hamim Wigena<sup>1</sup>, Anik Djuraidah<sup>1</sup>

<sup>1</sup>*Department of Statistics, Bogor Agricultural University, Indonesia*

---

**Abstract:** *The analysis of the phenomenon of extreme values of climate, especially rainfall is very important for government to reduce the negative impacts. Global circulation model (GCM) is an important data in the climate system because it can provide information about the climate in the future on a large scale. Techniques to reduce the size of the spatial scale using statistical downscaling (SD). SD modeling method requires a more flexible alternative to the assumption that the resulting models can be used to describe the climate events. Generalized additive model (GAM) is a method that accommodates the influence of linear and nonlinear in extreme rainfall events. The methodology is applied to forecast monthly extreme rainfall in Indramayu District.*

**Keywords:** *extreme rainfall, generalized additive model, global circulation model, statistical downscaling*

---

### **I. Introduction**

Rainfall is an element that has a diversity of climate and the high fluctuation in Indonesia, so rainfall is the most dominant element characterizes the climate of Indonesia. Extreme rainfall is the rainfall above or below normal conditions. The analysis examines the extreme events are needed to minimize the adverse impacts due to the extreme rainfall [1].

In anticipation of the adverse effects of extreme rainfall, forecast information is required about the maximum extreme rainfall events. To get information on climate change is to use climate models that can be used to estimate climate conditions. One of the climate models used to estimate the climatic conditions are global circulation model (GCM). However, the GCM does not provide information on the local climate, but only the global climate. On climate forecasting, information is needed not only on a global scale but on a local scale, the local scale to obtain information from the GCM is very difficult because it has a low resolution GCM. Therefore the use of Statistical downscaling (SD) which is a technique that uses statistical models to examine the relation between a global data with local scale data [2].

Data GCM has the characteristic that are nonlinear, high-dimensional, and there is multicollinearity, thus modeling SD requiring precise analytical techniques so that the resulting model can be used to describe the phenomena. This problem can be overcome by a more flexible alternative methods of assumptions called nonparametric methods. Nonparametric methods that can properly accommodate the nonlinear effect is generalized additive model (GAM) [3].

This research examines the SD modeling approach the generalized additive model to describe extreme rainfall events that occur within a specified period. The data used is the rainfall data in Indramayu district which is the largest supplier of rice in West Java, Indonesia. Rice fields in Indramayu now diminishing because the extreme impacts of climate change and global warming.

### **II. Literature Review**

#### **2.1. Extreme Rainfall**

Rainfall is rain water level be accommodated in a flat, non volatile, not pervasive, and not flowing. Unit rainfall measured in millimeters (mm) or inches. Rainfall of one millimeter means in area of one square meter in a flat, there is water as high as one millimeter. Rainfall that falls in the territory of Indonesia is influenced by several factors such as topography, direction of slope terrain, wind direction parallel to the shoreline, and wind trip distance above the terrain. Extreme weather is the conditions above normal weather conditions that occur in a particular area of short term scale [4].

To identify extreme rainfall can use the block maxima method, that is to create a block period of time that annual period. Blocks that formed were analyzed to determine the maximum value and that observations are considered as extreme observations [5].

#### **2.2. Global Circulation Model (GCM)**

GCM is the data important in understanding the climate system because it can provide information about the climate shift in the future. GCM data in the form of grids indicate that the GCM is a form of spatial data. GCM is a mathematical depiction of the interaction of physics, chemistry, and dynamics of the Earth's atmosphere. This model assumed changes in weather elements in the form of an output grids according to

latitude and longitude. This model can be run to estimate the climate sensitivity to different conditions. Research using the GCM is often constrained in the broadest selection, methods, and domain, it can certainly make it difficult to continue the analysis. The approach can be adopted to solve the problem is to determine the domain using statistical downscaling (SD) [6].

### 2.3. Statistical Downscaling (SD)

Statistical downscaling (SD) is an empirical approach to the statistical relation between the atmospheric circulation and rainfall. SD gives good results with the condition 1) the close relation between the predictors and response that explains exactly local climate variability, 2) the predictors well simulated by GCM, and 3) the relation between the predictors and response does not change with time changes and remains the same even though there is climate change [7].

SD method is based on the assumption that the local climate is controlled by two factors, i.e large scale climatic characteristics (low resolution) and local physiography. Regional or local climate information can be obtained by determining a statistical model that relates large scale climate variable ( $\mathbf{X}$ ) with local variable ( $\mathbf{y}$ ). Outcome data from the large scale GCM simulations used as input for the statistical model to estimate the characteristics of the local climate. A common form of SD models is :

$$\mathbf{y}_{(t)} = f(\mathbf{X}_{(t \times g)})$$

- where,  $\mathbf{y}_{(t)}$  : local climate variable (rainfall)  
 $\mathbf{X}_{(t \times g)}$  : GCM outputs variable (presipitation)  
 $t$  : amount of time (monthly)  
 $g$  : amount of domains GCM grid

### 2.4. Generalized Additive Model

Expansion of additive model for exponential family distribution be GAM. The general form of GAM are as follows :

$$\eta = g[E(y|x_1, x_2, \dots, x_k)] = s_0 + \sum_{j=1}^p s_j(x_{ij}), \text{ for } i = 1, 2, \dots, n$$

where  $g$  is link function and  $s_j$  modeled with a nonparametric smoothing function.

Smoothing is basically a process that can systematically eliminate rough data pattern (fluctuating). Smoothing spline function is a function that has high flexibility and is capable of handling data change behavior in certain subintervals. Spline is a piecewise polynomial that are segmented continuous, can model the data pattern that shows up/down with sharp points by using knots, and the resulting curve is smooth [8].

GAM estimation method using a backfitting algorithm. Estimation of smooth functions  $s_0, s_1(\cdot), s_2(\cdot), \dots, s_p(\cdot)$  in

the additive model, where  $E(s_j(\mathbf{X}_j)) = 0$  is using partial error  $\mathbf{R}_j = \mathbf{Y} - s_0 - \sum_{k \neq j}^p s_k(\mathbf{X}_k)$  so that

$E(\mathbf{R}_j | \mathbf{X}_j) = s_j(\mathbf{X}_j)$  by minimizing  $E\left[\mathbf{Y} - s_0 - \sum_{k=1}^p s_k(\mathbf{X}_k)\right]^2$ , then the estimator  $s_j$  be obtained iteratively if

given estimator  $s_i$  for  $i \neq j$  [9].

## III. Methodology

### 3.1. Data

The GCM data used are monthly rainfall data from Climate Model Intercomparison Project (CMIP5) issued by KNMI, Netherlands in 1979 to 2008. The area of a square area measuring 8 x 8 grid around Indramayu district, so there are 64 variables  $X$ . Local rainfall data used is the data from climatological station Indramayu district. Rainfall data from this station are in the period 1979 to 2008 (the data length of 30 years) so there are 360 observation data.

### 3.2. Methods

1. Exploration of Indramayu district rainfall data and GCM data
2. Modeling SD based on principal component (PC), where the parameter estimation using the ordinary least square
3. Modeling SD based on GAM :
  - a. Creating plot of extreme rainfall with each PC selected to indicate the nonlinear effect on the data
  - b. Determine the degree of smoothing in non-linear predictor

- c. Estimate the  $s_j$  in model  $\eta = s_0 + \sum_{j=1}^p s_j(x_{ij})$  backfitting algorithm
  - d. Creating a plot of the functional form of the model to evaluate the dependence of extreme rainfall with each predictor
4. Comparing the results of SD based on PC with GAM

#### IV. Results And Discussion

##### 4.1. Data Exploration

Figure 1 below shows the pattern of rainfall in Indramayu district from 1979 to 2008 monsoon pattern. The pattern of monsoon rainfall pattern is U-shaped or have a rainy season peak (unimodal). Monsoon pattern in Indramayu district has a peak in January because the average rainfall is the highest among the other months. Figure 1 also shows the rainy season generally occurs between the months of October to March and the dry season occurs between April and September.

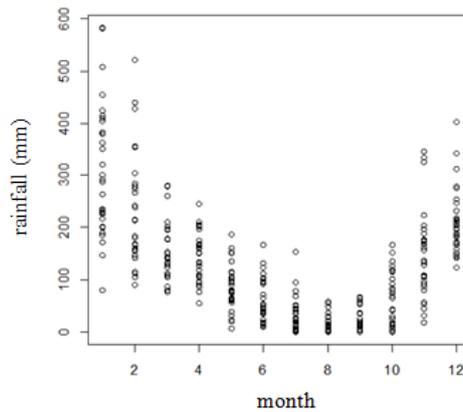


Figure 1 Rainfall Indramayu District in 1979-2008

Identification of extreme rainfall in the district of Indramayu using block maxima method, by creating a block period of years so that there are 30 blocks (from 1979 to 2008). Then determine the value of the highest rainfall in each block and the value of the rainfall is as extreme rainfall.

GCM data used are in accordance with the observational data of extreme rainfall data. GCM data are monthly data with a square area measuring 8 x 8 grid, so there are 64 variables  $X$ . This dimension is considered large enough to be analyzed, so that the required reduction of the dimension from the component origin variable into principal component (PC) dimension using principal component analysis (PCA). PC obtained a summary of the original variables as possible to explain the origin variables. The number of principal components that had eigen value greater than one are 6 principal components (PC). Proportion cumulative from PC-1 to PC-6 is 90.3%, it shows that PC is able to explain 90.3% of its original variables. Subsequent analysis of the response ( $y$ ) in this case is extreme rainfall in Indramayu district will be modeled with six predictors (variables of data GCM is reduced using PCA) which is the principal component score ( $PC_1, PC_2, PC_3, PC_4, PC_5, PC_6$ ).

##### 4.2. Modeling SD Based on PC

Extreme rainfall data in Indramayu district then analyzed using principal component regression and estimate its parameters using ordinary least squares.

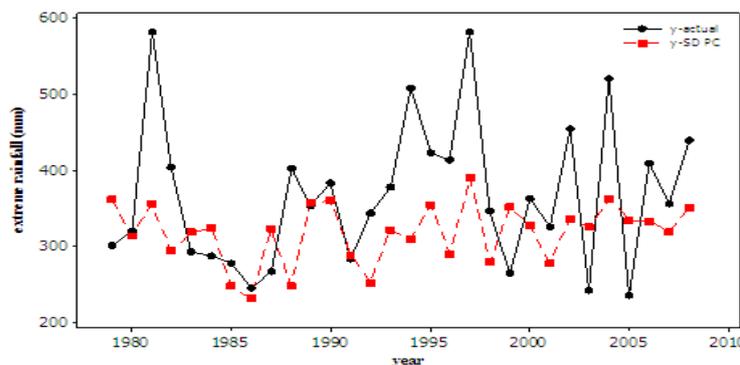
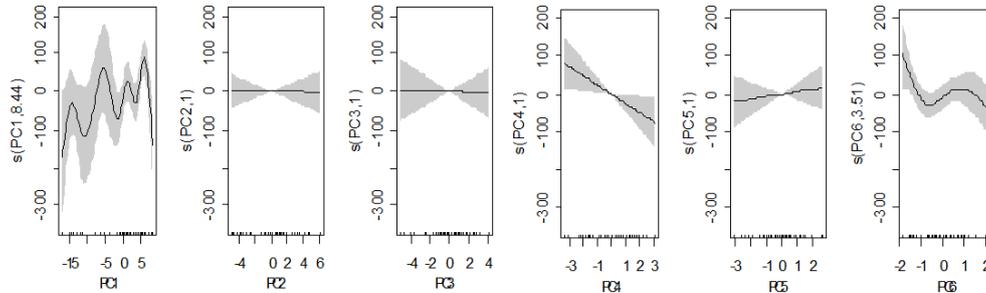


Figure 2 the result of SD based on PC model

Figure 2 shows  $R^2 = 16.3\%$  that the values obtained in model SD based on PC, the data can be explained by the GCM only 16.3% of the variability of extreme rainfall data of Indramayu district. The result gives RMSE value of 101.433, RMSEP of 70.839, and the correlation value of 0.404. These values indicate that SD based on PC has not give a good prediction model.

**4.3. Modeling SD Based on GAM**

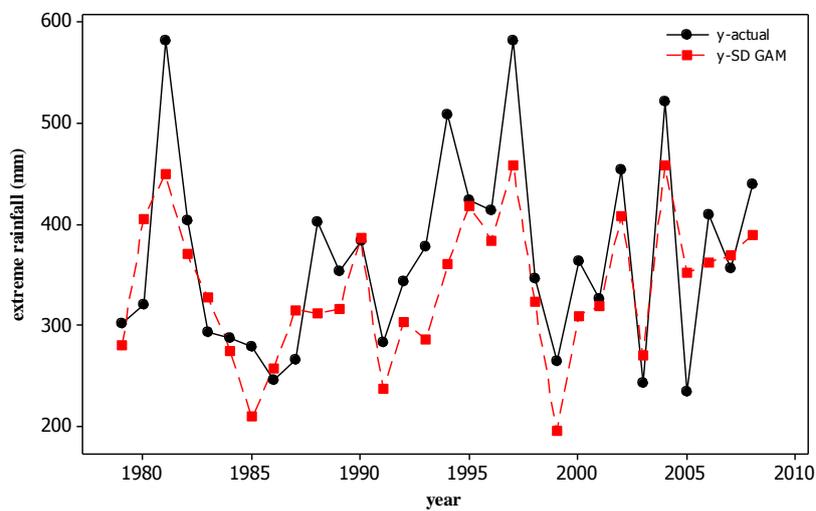
The next step is exploration of the data. Exploration by creating a matrix plot of the data to indicate a nonlinear effect. Figure 3 shows that there the partial relation between the response with each predictors are not linear, it is seen by the distribution of the data that follows a certain nonlinear distribution.



**Figure 3** Matrix plot between response with each predictors

The GAM is an alternative method that is subsequently used to accommodate the nonlinear effect of predictors through a nonparametric regression approach. Nonparametric regression modeling has mainly flexibility in determining the shape of the curve need not be specified a priori, but the curve is formed in accordance with the data (data driven). GAM analysis by modeling extreme rainfall as an additive combination of the functional form from the variable PC selected.

While the estimation of the SD based on GAM model gives RMSE value of 67.676, 40.113 of RMSEP and correlation value of 0.787. RMSE and RMSEP values are small and the correlation value is large enough indicate that the model SD based on GAM is good enough to give an estimates value of extreme rainfall in the Indramayu district.



**Figure 4** The result of SD based on GAM model

**4.4. Comparison**

The results of the extreme rainfall analysis from both methods are presented in Figure 5 which gives the information that the SD based on GAM model has a pattern more similar with actual data of extreme rainfall. It means that the estimation SD based on GAM approach actual data of extreme rainfall in Indramayu district.

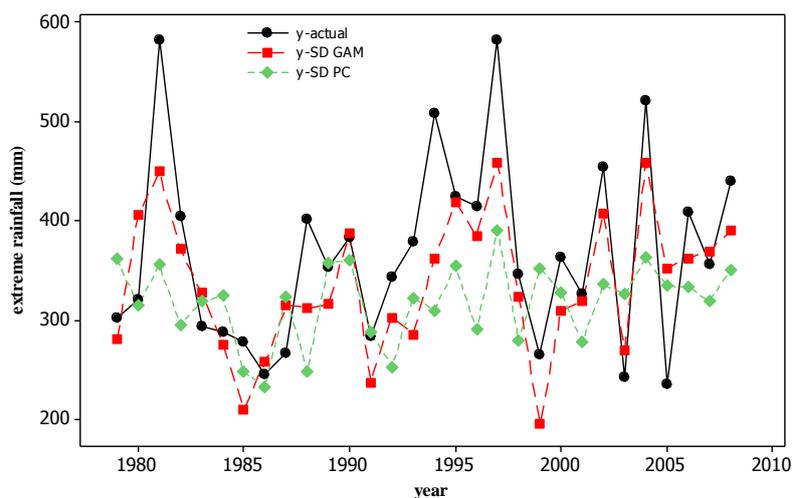


Figure 5 The plot between extreme rainfall with the estimated value

These results are also confirmed by the RMSE, RMSEP, and the correlation value of two methods. SD based on GAM model gives the smallest value of RMSE and RMSEP, and also have the largest correlation value than SD based on PC model. It means that SD based on GAM model can estimate rainfall extreme values better.

## V. Conclusion

Estimation of extreme rainfall in Indramayu district by using SD based on GAM obtained showed a trend similar to the pattern of the actual data, so it can be concluded that the SD based on GAM model is relevant for used to estimate the extreme rainfall in Indramayu district.

## Acknowledgements

The financial support from the Ministry of Higher Education (MOHE) on the statistical downscaling project is gratefully acknowledged.

## References

- [1]. J.D. Prang, Sebaran Nilai Ekstrim Terampat Dalam Fenomena Curah Hujan, Bogor Agricultural University, Indonesia, 2006.
- [2]. A. Busuioc, D. Chen, C. Hellstrom, Performance of statistical downscaling models in GCM validation and regional climate change estimates (Application for Swedish Published date 09June 2014, precipitation), International Journal of Climate, 21, 2001, 557-578.
- [3]. N. Beck, S. Jackman, Getting The Mean Right is a Good Thing: GAMs (San Diego: University of California, 1997)
- [4]. S. Hidayat, Iklim Lokal (Indonesia: Mercubuana University, 2008).
- [5]. M. Gilli M, E. Kellezi, An Application of Extreme Value Theory for Measuring Risk, Department of Econometrics, University of Geneva, Switzerland, 2003.
- [6]. U. Haryoko, Pendekatan Reduksi Dimensi Luanan GCM untuk Penyusunan Model SD, Bogor Agricultural University, Indonesia, 2004.
- [7]. C.B. Uvo, J. Olsson, O. Morita, K. Jinno, A. Kawamura, K. Nishiyama, N. Koreeda, T. Nakashima, Statistical atmospheric downscaling for rainfall estimation in Kyushu Island Japan, Hydrology and Earth System Sciences, 5, 2001, 259-271.
- [8]. W. Hardle, Applied Nonparametric Regression (New York: Cambridge University Press, 1994).
- [9]. T. Hastie, R. Tibshirani, Generalized Additive Models (London: Chapman and Hall, 1990).