# Application of Statistical Downscaling with Principal Component Regression for Local Rainfall Forecasting in Jember Regency

***Alfian Futuhul Hadi*[1]*, Abduh Riski*[2]*, Okit Tazkiyah*[3]*, Dian Anggraeni*[4]*, Izdihar Salsabila*[5]*, Dimas BC Wicaksono*[6]**

[1]Data Science Research Group, Department of Mathematic, University of Jember, Jember, Indonesia
Email: afhadi@unej.ac.id
[2]Data Science Research Group, Department of Mathematic, University of Jember, Jember, Indonesia
Email: riski.fmipa@unej.ac.id
[3]Data Science Research Group, Department of Mathematic, University of Jember, Jember, Indonesia
Email: tazkiyah@students.unej.ac.id
[4]Data Science Research Group, Department of Mathematic, University of Jember, Jember, Indonesia
Email: dian_a.fmipa@unej.ac.id
[5]Data Science Research Group, Department of Mathematic, University of Jember, Jember, Indonesia
Email: izdiharsalsabila.is@gmail.com
[6]Department of Biostatistics and Epidemiology, University of Jember, Jember, Indonesia
Email: wicaksono@unej.ac.id

**Abstract-** Global climate change causes various changes and extreme fluctuations in weather circumstances, including extreme changes in rainfall. An accurate rainfall forecasting was indeed needed in various agricultural activities. The statistical downscaling (SD) was developed to model the global climate circumstance data from the satellite, called the General Circulation Model (GCM). Combine with data on the earth from the weather station; the GCM predict the future local weather. The functional relationship in the SD was modeling the GCM output data as the predictors and the local-scale rainfall data as the response. The GCM's ability to display predictive data for decades to come was a technological leap in forecasting the rainfall to study long term on weather/climate change. Statistically, this modeling requires the twos below: (1) a dimensional reduction in GCM data and (2) accurate predictive models on the functional relationship. In this study, rainfall forecasting was conducted in Jember Regency using Principal Component Analysis (PCA) for dimensional reduction and a predictive model of Principal Component Regression (PCR). The accuracy was measured in each cluster in the 8×8, and 10×10 domains with the RMSE statistic was around 80.41-101.35.

***Keywords***: General Circulation Model, Statistical Downscaling, Principal Component Regression.

## 1. Introduction

Most of Jember Regency consists of lowlands with an average height of 83 meters and is a relatively fertile area and very suitable for the development of agricultural and plantation [1]. Global climate change causes various changes and extreme fluctuations in weather circumstances, including extreme changes in rainfall. it will influence the cropping patterns, planting time, production, and quality of yields, so that information on rainfall forecasting was indeed needed in various agricultural activities.

One of time series data-based rainfall forecasting using the Kalman Filter method has been carried out a weakness in the long period forecasting time intervals [2]. Therefore, we need a better forecasting model by considering the information on climate parameters globally [3]. The data was obtained from the General Circulation Models (GCM) output, which is a computer-based model by simulating global climate variables in each grid and an atmospheric layer which is then used to predict long-term climate patterns. However, the GCM information is still on a global scale resulting in a too low resolution so that it is dificult to use to obtain information on smaller local scale phenomena [3]. The Statistical downscaling (SD) can be connector the global scale of GCM with a smaller local scale based on the functional relationship.

As a downscaling process, the SD technique is static. By using data on large-scale grids within a specified period as a basis for determining data on a smaller scale of grid, the SD was used to predict monthly rainfall.

The GCM output data usually had a multicollinearity problem since its highly correlated each other. This multicollinearity was a violation in term of regression analysis assumption. Principal component analysis (PCA) will solve this problem [4]. In this study, principal component analysis (PCA) method was used to reduce dimensions. The results of the reduction of the grid domain dimension were regressed with the response variable; for example, the monthly rainfall of 2005-2016s period to obtain the SD model. The SD model obtained is then evaluated using the RMSE value criteria and the correlation value of R.

*Corresponding Author: *Afilan Futuhul Hadi*

Article History: Received:, Revised: , Accepted:

## 2. Materials and Methods

The data of this study are the GCM out-put monthly precipitation data (from the web of https://climexp.knmi.nl/start.cgi) as global scale data (predictor variables) and the monthly rainfall data of Jember Regency as local data (response variable) in each 2005-2016s period.

The GCM data is very accurate satellite image data with global scale or large size area. Hence it has a low resolution in small areas or local scale of rainfall phenomena screening. The SD technique was connecting both of the global scale variables from the GCM and the local scale variables. The local scale variables in this study are monthly rainfall data in Jember, as many as 144 months or 12 years; it called 2005-2016s period. Meanwhile, the global scale data used was GCM data with the center coordinates of the Jember located at coordinates -8.18448 latitudes and 113.668076 longitudes. Data domains are $30_{o}\times30_{o}$ square. This data was in the form of 144 grids, where each grid has a size of $2.5_{o}\times2.5_{o}$ (See Figure 1). Each grid contains the value of monthly precipitation. The coordinates of the observation point are used to determine the size of the domain from GCM output data. Domain size uses 3×3, 4×4, 5×5, up to 12×12.

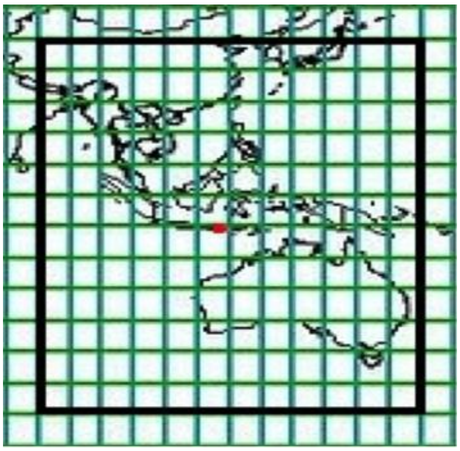

**Figure 1.** Center of the domain and the domain size of the GCM output data.

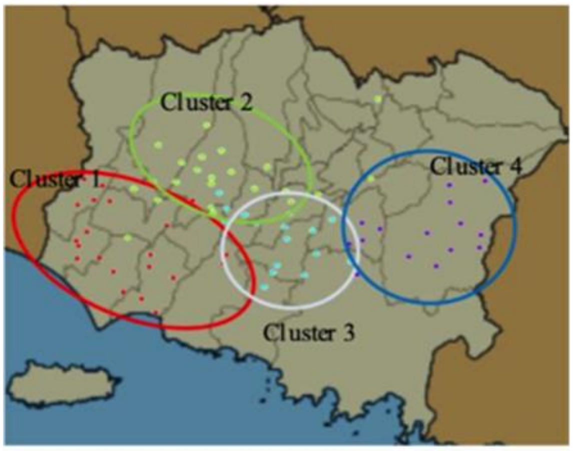

**Figure 2.** The 77 rainfall stations at Jember Regency and the 4 clusters of its.

The correlation value between the grid and the response variables of each cluster was calculated - the optimal size of the domain defined with the highest correlation value. The local data of monthly rainfall data of Jember Regency in the 2005-2016s period was come up from about 77 weather stations around the whole area. Then we clustered the 77 stations into 4 clusters (See Figure 2). We then divided both data into two parts: the training set (2005-2015s period) and testing set (2016 period). Then we forecast the rainfall for 2017-2020.

The GCM output data generally have multicollinearity or highly correlated, and this violates the terms of regression analysis, and the PCA will solve this problem. The PCA was used to reduce the predictor dimensions of GCM outcomes. The latest dimension is used in regression analysis, called principal component regression (PCR). The best model has a high correlation value and a small value of Root Mean Square Error (RMSE). The optimum model used for predicting the 2016 period. The use of 2016 period data is a forecasting correction with real data comparison.

*Corresponding Author: *Alfian Futuhul Hadi*
Article History: Received:, Revised: , Accepted:

## 3. Result and Discussions

### 3.1. Domain Size Selection

The domain size selection is a critical factor in Statistical Downscaling modeling [5]. The Observations with a domain of grid which too small will reduce information on global influences, the other side the size of the grid domain that too large will reduced local information. Commonly, the determination of the domain size was determined priory. There was no specific method to determine the size of the research domain grid [3]. However, there was an essential goal in the SD modeling; that is the close relationship between the response variable (local data) and the predictor variable (GCM output). We search an optimal domain in this study refer to the criteria of high correlation value between the predicting response from PCR in the selected grid and the response variables. The correlation value of the domain size and the response variables in each cluster of local data rainfall was shown in Table 1. According to the correlation value, the optimum grid of the domain size was 8×8 for all cluster, except for cluster 2 the optimum ones was 10×10.

### 3.2. Principal Component Analysis (PCA)

Regression-based Statistical Downscaling modeling from the GCM output data tends to have a multicollinearity problem. Furthermore, it was too large of predictor dimension that allows the number of predictor variables to exceed the number of available samples. That matter is that regression analysis was no longer eligible since it required the number of samples greater than the number of predictor variables ($n > p$).

In this study, there are 144 rows of local data or response variables. The 132 rows were used as training data, and 12 rows are used as testing data. The 12×12 grid size is certainly can be done in regression analysis because the number of predictor variables is greater than the number of samples. This problem overcame by reducing the predictor variables using the PCA. After getting the optimal domain size, then the predictor variable size is reduced to optimal variables or new components. The number of predictors in cluster 1, cluster 3 and cluster 4 are 64 variables, because the size of the domain in the cluster is 8×8. Meanwhile, in cluster 2 has 100 predictor variables. Figure 3 shows the cumulative variance in each principal component. In the Cluster 1, Cluster 3, and Cluster 4, the number of principal components selected was 30 principal components. These were reduced from 64 predictor variables in those clusters. Meanwhile, in cluster 2, 41 principal components were reduced from 100 variables of predictor [6]. Suggested choosing the principal components until the main components had a cumulative variance of 75%. This selection is based on the cumulative variance that can be seen in Table 2.

**Table 1.** Correlation value between cluster and domain size.

| Domain | Correlation Value | | | |
|---|---|---|---|---|
| Size | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| 3×3 | 0.844 | 0.857 | 0.851 | 0.829 |
| 4×4 | 0.806 | 0.815 | 0.827 | 0.791 |
| 5×5 | 0.802 | 0.815 | 0.830 | 0.787 |
| 6×6 | 0.868 | 0.872 | 0.852 | 0.859 |
| 7×7 | 0.813 | 0.813 | 0.837 | 0.798 |
| 8×8 | 0.855 | 0.860 | 0.908 | 0.864 |
| 9×9 | 0.808 | 0.808 | 0.828 | 0.792 |
| 10×10 | 0.809 | 0.917 | 0.825 | 0.794 |

**Table 2.** The PCA's cumulative variance of the 8×8 domain size and 10×10.

| Domain size 8×8 | | Domain size 10×10 | |
|---|---|---|---|
| Component | Cumulative Variance | Component | Cumulative Variance |
| PC28 | 0.72 | PC39 | 0.73 |
| PC29 | 0.73 | PC40 | 0.74 |
| PC30 | 0.75 | PC41 | 0.75 |
| PC31 | 0.76 | PC42 | 0.76 |
| PC32 | 0.77 | PC43 | 0.77 |

*Corresponding Author: *Alfian Futuhul Hadi*

Article History: Received:, Revised: , Accepted:

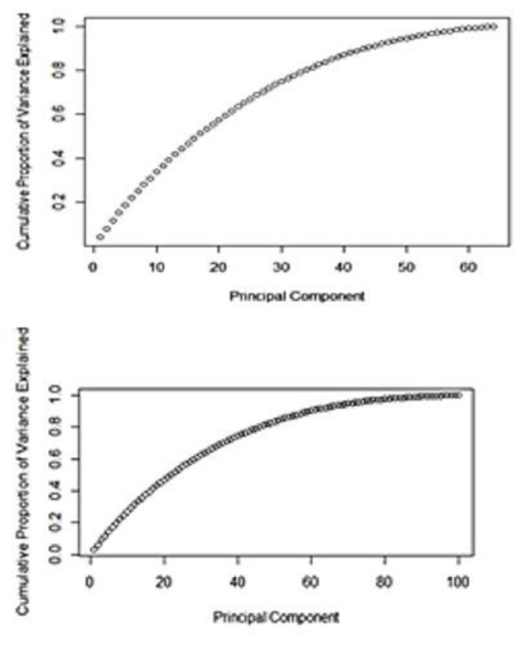

**Figure 3.** Principal component cumulative graphic of the 8×8 domain size (above) and the 10×10 domain size (bottom).

### 3.3. Statistical Downscaling with Principal Component Regression (PCR)

The equation linear least square regression is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

which is $x$ as predictor variables on the model and k is the number of predictor variables. One assumption that must be fulfilled in the regression analysis is that there is no multicollinearity in the predictor variables. Regression analysis with a large number of predictor variables tends to experience multicollinearity. Therefore, one way to overcome this problem is using Principal Component Regression (PCR). Rather than in the form of least square regression model in equation (1), the PCR has the equation as

$$y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_k C_k \quad (2)$$

which $C$ is the principal component as a linear combination from initial predictor variables. This equation has a more concise than the least-squares regression model where $m < k$. In this study, there were four models of PCR-based rainfall forecasting for Jember Regency, where each equation represented each cluster. Based on the results of PCA, for cluster 1, cluster 3 and cluster 4 used 30 components (m = 30), while cluster 2 used 41 components.

Table 3 shows the coefficient values for each variable in the form of rainfall model of equation (2) using SD techniques. One model in this section represents one cluster. To see the goodness and accuracy of this model, we cross-validate the rainfall prediction model using the 2016 period. The predicted value was compared with the observed value of the rainfall real in that period. The level of similarity of patterns between forecasting results and real values becomes a benchmark in this section. This similarity is measured by testing the correlation between both of them. In addition, to evaluate the similarity pattern, and the accuracy, we calculate the RMSE.

### 3.4 Evaluating Predicted Model

The Figure 4 provides the predicted value in the testing set of the period 2016 for all clusters. It can be seen that the comparison of real value and predicted value in cluster 2 is the most similarity pattern. It the smallest deviation compared to the other 3 clusters.

Statistically, the predictive power relations for each cluster can be seen from the correlation value, while the accuracy be observed from the RMSE value generated. The correlation values and the RMSE for each cluster are shown in Table 4. The correlation and the RMSE values for each cluster are not much different, namely in the range of 0.615 - 0.79 and RMSE in the range of 77.73 - 123.40. Cluster 3 has the highest correlation coefficient and the smallest RMSE compared to the other 3 clusters. In accordance with this, the similarity between the predicted value and the real observed value was also shown in Figure 4.

**Table 3.** The PCR coefficient value of each clusters.

| No. | P | Cluster | | | | No | P | Cluster | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $C1$ | $C2$ | $C3$ | $C4$ | | | $C1$ | $C2$ | $C3$ | $C4$ |
| 1 | $\beta_0$ | 114 | 169.9 | 155.5 | 128.5 | 22 | $\beta_{21}$ | 77.05 | 105.2 | 89.67 | 81.27 |
| 2 | $\beta_1$ | 89.88 | 139.5 | 119.8 | 102.4 | 23 | $\beta_{22}$ | 77.89 | 105.9 | 89.10 | 81.77 |
| 3 | $\beta_2$ | 75.23 | 110.4 | 91.50 | 84.34 | 24 | $\beta_{23}$ | 78.40 | 107.2 | 89.66 | 81.88 |
| 4 | $\beta_3$ | 75.85 | 109.5 | 91.63 | 84.80 | 25 | $\beta_{24}$ | 78.56 | 106.5 | 89.84 | 83.34 |
| 5 | $\beta_4$ | 76.49 | 108.9 | 92.07 | 84.88 | 26 | $\beta_{25}$ | 79.46 | 106.1 | 91.11 | 84.24 |
| 6 | $\beta_5$ | 75.05 | 107.3 | 89.41 | 83.09 | 27 | $\beta_{26}$ | 78.71 | 105.3 | 92.75 | 84.39 |
| 7 | $\beta_6$ | 74.9 | 107.7 | 88.73 | 83.18 | 28 | $\beta_{27}$ | 80.33 | 105.4 | 93.06 | 84.94 |
| 8 | $\beta_7$ | 73.12 | 106.3 | 86.50 | 80.21 | 29 | $\beta_{28}$ | 80.43 | 105.6 | 94.69 | 85.52 |
| 9 | $\beta_8$ | 73.39 | 107.1 | 86.46 | 80.65 | 30 | $\beta_{29}$ | 82.16 | 107.1 | 94.86 | 85.63 |
| 10 | $\beta_9$ | 73.76 | 108.3 | 86.83 | 81.21 | 31 | $\beta_{30}$ | 82.9 | 108.8 | 96.67 | 86.80 |
| 11 | $\beta_{10}$ | 74.43 | 108.6 | 87.55 | 81.08 | 32 | $\beta_{31}$ | 0 | 108.7 | 0 | 0 |
| 12 | $\beta_{11}$ | 74.88 | 107.7 | 88.39 | 79.63 | 33 | $\beta_{32}$ | 0 | 109 | 0 | 0 |
| 13 | $\beta_{12}$ | 73.89 | 108.7 | 86.57 | 79.12 | 34 | $\beta_{33}$ | 0 | 108.1 | 0 | 0 |
| 14 | $\beta_{13}$ | 74.32 | 106.5 | 86.91 | 79.15 | 35 | $\beta_{34}$ | 0 | 109.6 | 0 | 0 |
| 15 | $\beta_{14}$ | 74.86 | 108.3 | 86.48 | 79.34 | 36 | $\beta_{35}$ | 0 | 109 | 0 | 0 |
| 16 | $\beta_{15}$ | 73.99 | 107.1 | 85.98 | 80.28 | 37 | $\beta_{36}$ | 0 | 110.1 | 0 | 0 |
| 17 | $\beta_{16}$ | 73.93 | 104.7 | 86.29 | 79.73 | 38 | $\beta_{37}$ | 0 | 110.6 | 0 | 0 |
| 18 | $\beta_{17}$ | 74.20 | 103.2 | 87.03 | 79.56 | 39 | $\beta_{38}$ | 0 | 111.3 | 0 | 0 |
| 19 | $\beta_{18}$ | 75.14 | 103.4 | 87.19 | 80.04 | 40 | $\beta_{39}$ | 0 | 111.5 | 0 | 0 |
| 20 | $\beta_{19}$ | 76.10 | 104.4 | 86.70 | 81.20 | 41 | $\beta_{40}$ | 0 | 111.4 | 0 | 0 |
| 21 | $\beta_{20}$ | 75.28 | 104.8 | 88.74 | 81.24 | 42 | $\beta_{41}$ | 0 | 113.8 | 0 | 0 |

**Table 4.** Validation results on correlation values and RMSE.

| Test | Cluster | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Correlation | 0.62 | 0.73 | 0.79 | 0.60 |
| RMSE | 95.27 | 134.70 | 77.72 | 123.40 |

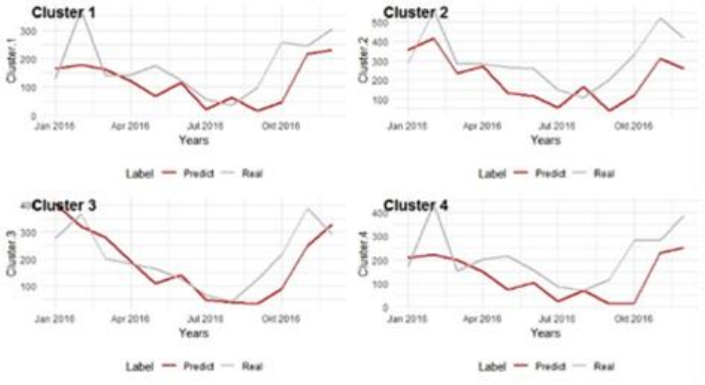

**Figure 4.** Predicted value of the test set for all cluster.

### 3.5 The Forecasting Result

Finally, we provide a little bit long-run forecasting result for 2017-2020. Figure 5 was the graphically representative of the real rainfall data, the predicted value in training and test set, and also the forecasting valued for next periods. As we knew the strong-powerful mathematical relationship of the GCM output data and the rainfall observed data in the SD schemes would be resulting in the accurate forecasting value. Some important things we would stripe for was that find the strong predicted model in the test set.

Especially for the next 2020 rainfall forecasting, we can see in Table 5 that mostly the highest rainfall in wet seasons will occur in early of 2020, January - February and at the end of 2020 from October to December. The

*Corresponding Author: *Alfian Futuhul Hadi*

Article History: Received:, Revised: , Accepted:

dry seasons will occur in June to September of 2020. There are slight differences between clusters; cluster 1 will be the driest one then followed by cluster 4, will have the driest in June and August 2020. While cluster 2 and cluster 3 will be the wettest area in January to February 2020 with around 400mm of rainfall forecasting value.

**Table 5.** Rainfall forecasting value (mm) for all clusters in next 2020.

| Month | Cluster | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Jan | 264.41 | 464.29 | 409.29 | 284.63 |
| Feb | 231.03 | 432.95 | 352.44 | 249.25 |
| Mar | 158.14 | 256.21 | 154.46 | 200.07 |
| Apr | 190.24 | 267.69 | 279.45 | 185.58 |
| May | 120.65 | 241.34 | 148.98 | 108.85 |
| Jun | 18.26 | 96.49 | 78.67 | 36.64 |
| Jul | 69.52 | 97.25 | 129.32 | 62.93 |
| Aug | 31.52 | 109.38 | 84.64 | 53 |
| Sep | 75.97 | 96.35 | 149.79 | 51.49 |
| Oct | 72.49 | 125.58 | 133.14 | 62.46 |
| Nov | 157.03 | 210.48 | 257.22 | 161.17 |
| Dec | 219.59 | 388.48 | 349.98 | 255.88 |

## 4. Conclusion

For the next 2020 rainfall forecasting, the highest rainfall in wet seasons will occur in early of 2020, January - February and at the end of 2020 from October to December. The cluster 1 seem will be the driest one then followed by cluster 4, will have the driest in June and August 2020. While the cluster 2 and cluster 3 will be the wettest area in January to February 2020 with around 400 mm of rainfall forecasting value. Cluster 3 has the highest correlation coefficient and the smallest RMSE compared to the other 3 clusters.

For further research, we thought that it is required to develop the SD forecasting for the extreme value representing the potential of drought resister and also the floods. Statistically, the important things we would strive for is to find the strong predicted model in the SD scheme.

## Acknowledgments

This research was supported by the Ministry of Research, Technology and Higher Education of Indonesia. Research funding by Grant No.061/SP2H/LT/DRPM/2019. For the preparation of this paper, we thanks to all member of Data Science Research Group, and all member of the Statistical Laboratory, Department of Mathematics, University of Jember, Indonesia.

## References

[1] The official portal of the agriculture and food government of East Java, online aviable from http://pertanian.jatimprov.go.id/index.php/komoditas/sentra-hortikultura/14-kab-jember.

[2] R. Tresnawati and K.E. Komalasari. The Scenario of Nino 3.4's SST grace period for rainfall to improve the accuracy of the Kalman Filter prediction, Pusat Penelitian dan Pengembangan. BMKG, Jurnal Meteorologi dan Geosika, Vol 12, No 3, 241-249, 2011.

[3] A. J. Wigena. Statistical Downscaling Dengan Pergeseran Waktu Berdasarkan Korelasi Silang, Pusat Penelitian dan Pengembangan. BMKG, Jurnal Meteorologi dan Geosika, Vol 16, No 1, 2015.

[4] D. F., Morrison. Multivariate Statistical Methods Series in Probability and Statistics, Mc Graw Hill, Singapore, 1978.

[5] E. Fernandez. On the influence of predictors area in statistical downscaling of daily parameters, Report no. 09/2005, Oslo: Norwegian meteorological institute, 2005.

[6] R. L. Wilby and T.M. Wigley. Precipitation Predictors for Downscaling: Observed and General Circulation Model Relationships, International Journal of Climatology, Vol 20, Issue 6, 641-661, 2000.

*Corresponding Author: *Alfian Futuhul Hadi*
Article History: Received:, Revised: , Accepted: