Eastern Journal of Agricultural and Biological Sciences (EJABS)
Website: https://qabasjournals.com/index.php/ejabs

Prediction of Daily Rainfall in Dodokan Watershed Based on Statistical Downscaling Model: An Effort to Manage Watershed Ecosystems

Mustika Hadijati[1]*, Irwansyah [1] , Nurul Fitriyani [1]
[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram – Indonesia*
Corresponding Author: Mustika Hadijati, E-mail: mustika.hadijati@unram.ac.id

ARTICLE INFO

Received: May 1, 2021
Accepted: June 1, 2021
Volume: *1*
Issue: 2

KEYWORDS

CART, Ecosystems, Kernel Nonparametric Regression, Statistical Downscaling, Rainfall.

ABSTRACT

Climatic conditions in watersheds, especially rainfall, can affect ecosystems in the watershed. Changes in ecosystem conditions can affect the condition of the watershed. It also affects the condition of water resources in the watershed. Therefore, developing a rainfall model that provides an accurate rainfall prediction is necessary to manage the watershed ecosystems. The rainfall model was developed using the statistical downscaling method. This method utilizing global-scale GCM output information, so data pre-processing is needed. The CART algorithm is one method that can be used for data pre-processing to lessen the GCM data dimension. Moreover, the functional form of the rainfall relationship with GCM data was determined using nonparametric kernel regression. The final result obtained a rainfall model that provides fairly good prediction accuracy with a relatively small RMSE value.

1. Introduction

Management of watershed ecosystems is essential. It will affect watershed conditions, so that also affects water resources in the watershed. Climatic conditions, especially rainfall, influence the conditions of watershed ecosystems. Therefore, it is necessary to develop a rainfall model to provide accurate rainfall prediction to manage the watershed ecosystem.

There are many methods to develop a rainfall model. One of those methods is the statistical downscaling model using the General Circulation Model (GCM) information. However, GCM as the main predictor is the global scale variables, so local forecasting is less suitable. Therefore, a downscaling technique is necessarily needed for localization. One of them is Statistical Downscaling (SD) which arranges a relationship function among the GCM results and the local climate details [1].

Data regarding the GCM output is a dimensionality curse, spatial, and temporary data. They cause multicollinearity, a condition where there is a correlation among several grids in a single domain. Hence, degrading the data dimension is required as a pre-processing in conducting SD modeling. GCM data is ordinarily non-normal, so that the pre-processing method applied is the one that does not assume the data normality; that is, the algorithm based on Classification and Regression Tree (CART) ([2], [3], [4]). GCM data is projected on some dominant predictor variables by the CART algorithm in the SD model.

GCM results and regional climate data are usually temporal, spatial, not linear, not stationary, and not follow the normal distribution. Nonparametric regression is a suitable SD method for these features, as it explains tentative varieties and tremendous utilities, especially for climate data ([3], [4], [5]). This research aimed to utilize a nonparametric kernel SD model to simulate the rainfall data of the Dodokan watershed. The rainfall simulation data will be used to create the river water discharge model.

2. Literature Review

2.1 Previous Research

Standard SD techniques sometimes fail to respond to the spatial dependence of total rainfall and extreme hydrometeorological temperatures. The downscaling algorithm was proposed in research [3] to overcome this problem by simulating rainfall throughout the studied river area using CART and projecting it using a nonparametric regression estimator. This approach is proven to produce models with good performance.

In addition, the kernel regression-based SD methodology introduced by [3] also provided a fairly good resolution daily rainfall projection for the studied watershed. This model was applied in reducing the scale of daily rainfall in and near the watershed. Several variables were used as GCM independent variables, i.e., air heat, determined humidity, meridional and wind zones, pressure at sea level, and the top geopotential height. The resulting model showed the likelihood of an extreme event occurring with a rise in the watershed rainfall incidence in the time ahead [5].

Literature studies show that previous research predicting the daily rainfall in the Dodokan Watershed based on a statistical downscaling model has not been done before. Therefore, by referring to the results of previous research which are used as a reference, the researchers are interested in applying the explained method in this study.

2.2 Theoretical Framework

a. Statistical Downscaling

Downscaling is a data conversion process, transforming the higher-scale into lower-scale data units. One of the approaches to downscaling is Statistical Downscaling (SD). This approach is a steady procedure where specific data periods with higher-scale grids form lower-scale grids. SD utilizes universal data to get connection function between GCM global and local scale, formularized by [1]:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \quad (1)$$

$\mathbf{Y}$ indicates the response, $\mathbf{X}$ reveals the predictors unify the spatial decrement output of GCM variables, and $\boldsymbol{\varepsilon}$ shows the model error. SD methods have been evolved and can be categorized into transfer function, weather recording, and generator [6].

b. Algorithm for Classification and Regression Tree

Classification and Regression Tree, commonly known as CART, categorize historical data called studied samples to construct a decision tree. This tree is expressed as a question collection that divides the learning sample into slighter parts, and it is used for new data division [7].

CART concept appears when T examination sample computes on Y variable with the value of $1,2,\cdots,T$ and k independent variables with many related features. The relationship is not linear, complex, and leads to hardship in making the functional models between the variables. A nonparametric method that splits data into slighter spaces must be developed to merely the relations. Dividing up is needed to be done recursively up to getting the new predictor variable [8]. Three parts of the CART algorithm are given with the following outputs: a maximum tree, the appropriate tree size, and constructed tree for new data classification [7].

c. Nonparametric Regression with Multivariate Kernel

A multivariate nonparametric model can be described as the following:

$$Y_t = m(\mathbf{X}) + u_t, \quad t = 1,2,\ldots,T \quad (2)$$

given error u_t and $E(u_t) = 0$. The function $m(.)$ is supposed to be unrevealed but smooth, $\mathbf{X} = (X_1, X_{2,} \ldots, X_k)$ are some random predictor variables, and Y_t is the response variable. The curve $m(\mathbf{X})$ is evaluated using a multivariate kernel approach. One of the kernel estimators that can be implemented is the estimator of Nadaraya-Watson. This estimator is frequently applied for random independent variables illustrations [9]. The estimator of Nadaraya-Watson can be formulated as [10]:

$$\hat{m}_{\mathbf{H}}(x_j) = \frac{\sum_{t=1}^{T} \left\{ \prod_j K_{\mathbf{H}}\left(\frac{x_j - X_{ij}}{h_j} \right) \right\} Y_i}{\sum_{t=1}^{T} \left\{ \prod_j K_{\mathbf{H}}\left(\frac{x_j - X_{ij}}{h_j} \right) \right\}} \quad (3)$$

$K(\cdot)$ indicates the function of the kernel and $\mathbf{H}$ reveals a vector of bandwidth. One of the kernel functions that can be used is the quartic kernel function, formulated as:

$$K(u) = \frac{15}{16}\left(1 - u^2\right)^2 \ , \quad I(|u| \le 1) \quad (4)$$

The kernel estimator curve is affected by the utility of bandwidth. The curve estimator will be under-smoothed if the bandwidth is too small and over-smoothed when the bandwidth is too big. Therefore, it is required to decide optimum bandwidth to obtain the finest kernel estimator with a smooth regression curve. The bandwidth selection can be made by considering the Generalized Cross-Validation (GCV) indicator, with the following formula [11]:

$$GCV = \frac{T^{-1} \sum_{t=1}^{T} \left(y_t - \hat{m}_{\mathbf{H}}(x_t) \right)^2}{\left(T^{-1} trace(\mathbf{I} - \mathbf{H}(h_0)) \right)^2} \quad (5)$$

3. Methodology

This study uses two types of data: daily rainfall data at observed stations (Kuripan and Kabul) in the Dodokan watershed and daily precipitation GCM data (total) obtained from CPC Global Precipitation data. The five years GCM data zone gathered is 8×8 domain on the Dodokan watershed, located in Lombok Island.

The rainfall data fabrication of the Dodokan watershed is derived from a nonparametric regression SD model oncoming. The dependent variable originates from daily rainfall data and the independent variables founts from GCM data daily precipitation at the 8×8 grid spots. As a preparatory for the modeling process, GCM data pre-processing is conducted to lessen the data dimension. This step occupies the CART algorithm with the following steps [7].

a) *Constructing a maximum tree according to a specific separated arrangement that forms the studied sample partition becomes smaller.*

Let the parent node t_p, the child node (left) t_l, the child node (right) t_r, the j-th variable x_j, and the best splitting measure of the j-th variable x_j^R. Then, with maximum homogeneity, the research data are separated into two components. The impurity function $i(t)$ explains the maximum homogeneousness of child nodes, and the impurity node of t_p is persistent for every possible partition. Therefore, the maximum homogeneity of the splitting child nodes will be analogous with the impurity function difference $\Delta i(t)$ maximization, formulated as:

$$\Delta i(t) = i(t_p) - E(i(t_c)) = i(t_p) - (P_l i(t_l) + P_r i(t_r))$$

$P_l i(t_l)$ and $P_r i(t_r)$ are the separated nodes possibility, respectively. Hence, the CART nodes accomplish the maximization matter:

$$\underset{x_j \le x_j^R,\, j=1,2,\ldots,k}{\arg\max} \left(i(t_p) - P_l i(t_l) - P_r i(t_r)\right) \tag{6}$$

Regarding the index of Gini, the impurity function is defined as follows.

$$i(t) = \sum_{k \ne l} p(k|t) p(l|t) \tag{7}$$

$k, l = 1, 2, \cdots, K$ indicates the class concordance and $p(k|t)$ indicates the possibility of class k when in node t. Therefore,

$$\Delta i(t) = -\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2(k|t_r) \tag{8}$$

b) *Choosing the suitable size of the tree*

The topmost tree possibly has extremely substantial complexity and contains a hundred levels. Therefore, before the new data are characterized, it is required to determine the optimal measurement of the tree. One of the optimization methods is the Cross-Validation method. This method considers tree complicatedness and misclassified error. The optimal portion between the two mentioned considerations will be acquired by minimizing the following cost complexity function:

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \tag{9}$$

$R(T)$ indicates the misclassified error of the tree T and $\alpha(\tilde{T})$ indicates the measure of complexity that relies on $\tilde{T}$ (tree's terminal node number total sum).

c) *Classifying new data applying the constructed tree*

CART algorithm assigns the influential GCM variables as a projection. These variables are then utilized in SD modeling, and the models for rainfall are established using a multivariate kernel approach in nonparametric regression.

4. Results and Discussion

The following figure reveals CART analysis for Kuripan station rainfall data regarding GCM precipitation data. The outcomes of the classification tree are drawn in Figure 4.1.

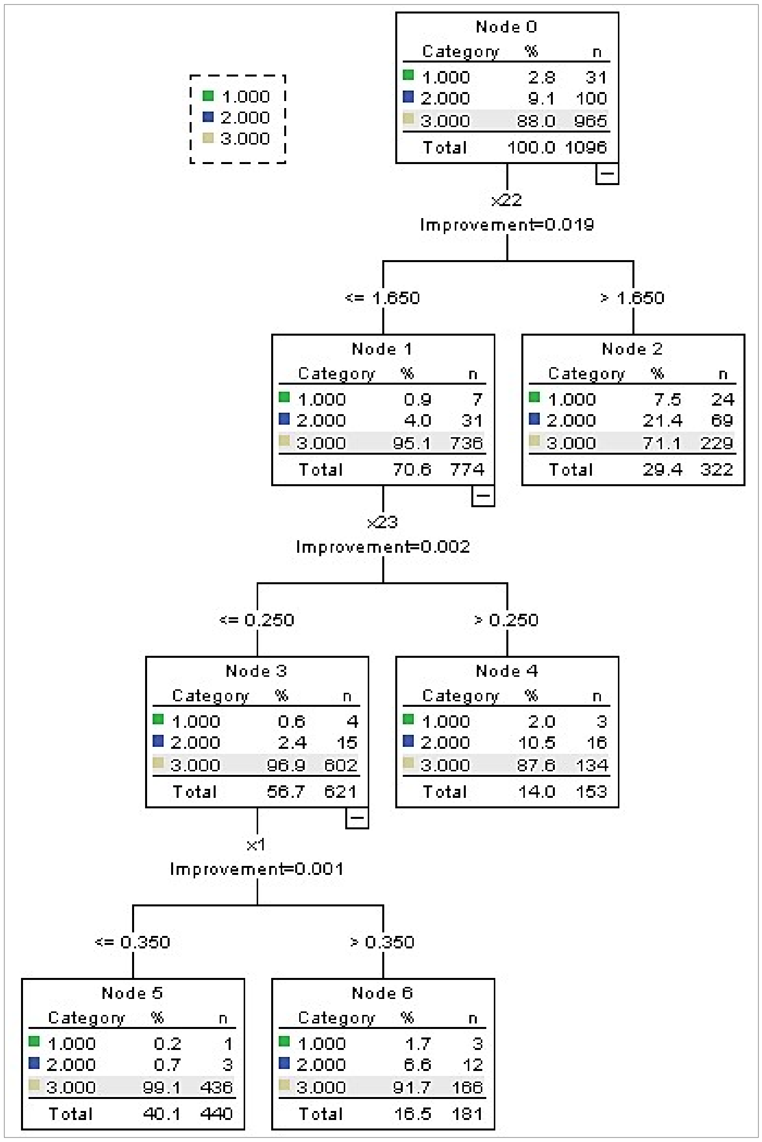

Figure 4.1
Rainfall Data Classification Tree of Kuripan Station

Figure 4.1 describes the seven nodes with four terminal nodes of the classified tree conducted. Three model variables arise for the GCM precipitation data on different grids, that are X_{22} (–8.75° N and 116.25° E grid), X_{23} (–8.75° N and 116.75° E grid), and X_1 (–7.75° N and 114.25° E grid). These three variables denote the variables that affect the rainfall classification at Kuripan station. Moreover, the following drawing illustrates the CART analysis of Kabul station rainfall data. The constructed tree is expressed in Figure 4.2.

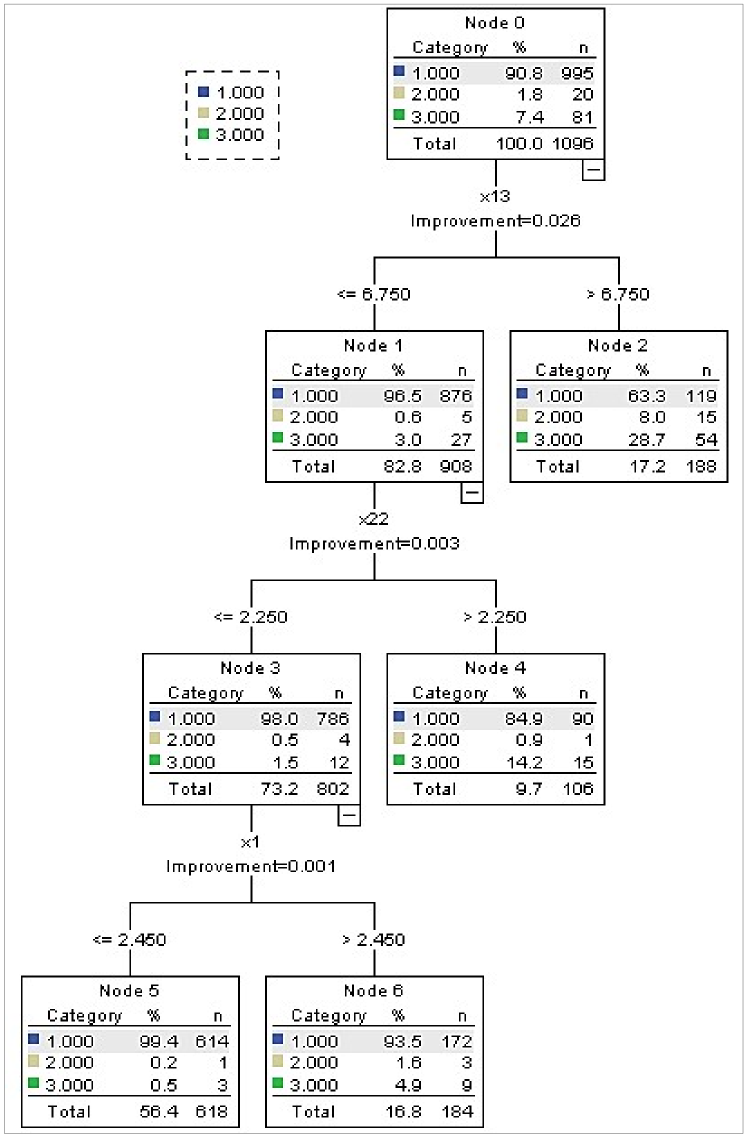

Figure 4.2
Rainfall Data Classification Tree of Kabul Station

Figure 4.2 represents the classification tree with seven nodes with four final nodes. Three model variables emerge for the GCM precipitation data on various grids, namely X_{13} (–8.25° N and 115.75° E grid), X_{22} (–8.75° N and 116.25° E grid), and X_1 (–7.75° N and 114.25° E grid). The recapitulation of the CART analysis outcomes is pointed out in Table 4.1.
Table 4.1 reveals that CART analysis results in high accuracy classification, with a more than 75% percentage. The accuracy value for the Kuripan station is 88%, it means that 88% of the CART classification outcome are suitable with the proper classification. Moreover, a 90.8% classification resulted by CART appropriates with the actual classification at Kabul station. This Table also indicates that the model variables mostly have an upper percentage of normalized importance with more than 85%, except variable X_1. Those three variables (X_{22}, X_{23}, X_{13}) substantially influenced the classified outcome for the rainfall at Kuripan and Kabul stations. Therefore, the GCM data are projected as independent variables in constructing the rainfall model.

Tabel 4.1
Recapitulation of CART Analysis

Stations	Number of Nodes	Number of Terminal Nodes	Accuracy	Model Variable	Normalized Importance
Kuripan	7	4	88.0 %	X_{22}	100.0 %
				X_{23}	99.4 %
				X_{1}	22.1 %
Kabul	7	4	90.8 %	X_{13}	94.8 %
				X_{22}	100.0 %
				X_{1}	4.0 %

The models for rainfall data are encouraged using nonparametric regression with a multivariate kernel approach. At Kuripan station, local rainfall is affected by the activity of GCM precipitation on the two different grids: (–8.75)° N – 116.25° E and (–8.75)° N – 116.75° E. Moreover, the motion of GCM precipitation on the two different grids, (–8.25)° N – 115.75° E and (–8.75)° N – 116.25° E, are influencing the neighborhood rainfall at Kabul station. These models are then applied as the simulation base of rainfall data for Kabul and Kediri stations.
The simulated results of rainfall data at Kuripan and Kabul stations are slightly close to the original rainfall data. These can be seen from the accuracy of data simulation measured by the Root Mean Square Error of Prediction (RMSEP) values. Table 4.2 reveals the RMSEP values of simulated data conducted for Kuripan and Kabul stations.

Tabel 4.2
Root Mean Square Error of Simulated Data Prediction

Stations	Root Mean Square Error of Prediction
Kuripan	6.65
Kabul	4.28

Based on Table 4.2, simulated data RMSEP values at Kuripan and Kabul stations are reasonably slight. These results denote that the simulated outcomes for rainfall data at Kuripan and Kabul stations in the Dodokan watershed are highly accurate. Therefore, research outcomes can be utilized to model the water discharge of the river at the Dodokan watershed.

5. Conclusion

The activity of GCM precipitation on two different grids has affected the rainfall at Kuripan station, in line with the activity of GCM precipitation at Kabul station. The precipitation of GCM on the (–8.75)° N – 116.25° E grid (variable X_{22}) significantly influenced the rainfall classification at both stations. Rainfall models at the Kuripan and Kabul stations provide relatively good prediction accuracy with a relatively small RMSE value, i.e., 6.65 and 4.28, respectively.

Funding: This research was funded by the Directorate General of Higher Education (DIKTI) Indonesia.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Wigena, A. H., Djuraidah, A., and Rizki, A. (2015). Semiparametric Modeling in Statistical Downscaling to Predict Rainfall. *Applied Mathematical Sciences*, 9(88), 4371-4382.
[2] Zorita, E. and Storch, H. V. (1999). The Analog Method as A Simple Statistical Downscaling Technique: Comparison with Complicated Methods. *J Clim,* 12, 2474-2489.
[3] Kannan, S. and Ghosh, S. (2013). A Nonparametric Kernel Regression Model for Downscaling Multisite Daily Precipitation in the Mahanadi Basin. *Water Resources Research,* 49, 1360–1385.

[4] Hadijati, M., Komalasari, D., and Fitriyani, N. (2016). Statistical Downscaling Model Using Nonparametric Regression to Predict Temperature in Selaparang Lombok. *Proceeding of the 1st International Conference on Science and Technology*, University of Mataram, Mataram, 129-132.
[5] Singh, S., Kannan, S., and Timbadiya, P. V. (2016). Statistical Downscaling of Multisite Daily Precipitation for Tapi Basin Using Kernel Regression Model. *Current Science,* 10, 1468-1484.
[6] Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P., and Mearns, L.O. (2004). Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods. *IPCC Reports*.
[7] Timofeev, R. (2004). Classification and Regression Trees (CART): Theory and Application. *Master Thesis,* Humboldt University, Berlin.
[8] Loh, W. Y., (2011). Classification and Regression Trees. *WIREs Data Mining Knowledge Discovery,* 1, 14–23.
[9] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, Inc.
[10] Hardle W., Muller M., Sperlich S., and Werwatz A. (2004). *Nonparametric and Semiparametric Models: An Introduction*. Germany: Springer.
[11] Sarda, P. and Vieu, P. (2012). Kernel Regression. *Wiley Series in Probability and Statistics,* 43–70.