Journal of Sciences, Islamic Republic of Iran 35(2): 159 - 166 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

The Principal Component Linear Spline Quantile Regression Model in Statistical Downscaling for Rainfall Data

A. S. Yulianti, A. Islamiyati*, E. T. Herdiani

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, Indonesia

Received: 19 April 2024 / Revised: 6 June 2024 / Accepted: 27 July 2024

Abstract

Information regarding rainfall can be obtained from global data, namely the global climate model that can be accessed through the statistical downscaling approach. Linear spline quantile regression with principal component is a statistical method that can be employed in statistical downscaling to address multicollinearity and outliers in data by using nonparametric estimators. This method is applied to rainfall data in Pangkep Regency from January 2008 to December 2022 as the response variable and global climate model data as the predictor variable. The aim of this research is to obtain the best regression model used for predicting rainfall data. The results obtained indicate that statistical downscaling with two principal components at the 0.50 quantile with respective knot points of -10.20 and -0.30 is the best model with the lowest generalized cross-validation value. The forecasted rainfall data using this model shows a high level of accuracy with a correlation of 89%.

Keywords: Principal Component; Quantile Regression; Rainfall; Spline; Statistical Downscaling.

Introduction

Rainfall is one of the climate elements that significantly influences the natural ecosystem framework. The intensity of rainfall determines the availability of water on Earth and plays a key role in sustaining human life (1). The intensity and timing of rainfall events that deviate from normal or extreme conditions are caused by global climate change (2). Extreme rainfall events have often been highlighted as they can have a significant impact on various aspects of life (1). Changes in the intensity of low or high rainfall can lead to droughts or floods. In certain conditions, it is crucial to have information about rainfall to mitigate the adverse effects (3).

Information about rainfall can be obtained from global climate models that contain information about atmospheric circulation to depict various climate subsystems on Earth (4). Statistical downscaling is a statistical approach used to understand the functional relationship between globally impacting climate variables, such as those from global climate models, and locally impacting climate variables, such as rainfall (5). One of the important aspects of statistical downscaling is the presence of a strong correlation between the two variables (6). Therefore, a time lag is applied in statistical downscaling (7). The time lag is applied to the global circulation model data so that the resulting patterns match those of the rainfall (8). The use of rainfall data in statistical downscaling has been applied in various conditions, including seasonal rainfall data, monthly rainfall, daily rainfall, and hourly rainfall (9–

* Corresponding Author: Tel: +628124261584; Email: annaislamiyati701@gmail.com

12).

The functional relationship between rainfall and global climate models in statistical downscaling can be established by applying a regression model, which has the potential to simulate past climate, current climate, and predict future climate (4). Quantile regression is a statistical method used to assess extreme changes in rainfall, providing flexibility across various quantile values (13). Quantile regression has been employed in extreme rainfall conditions in countries such as India, Bangladesh, Rwanda, and Taiwan (3,14–16). The extreme changes in rainfall result in irregular rainfall patterns (17), leading to the development of estimators for rainfall data in statistical downscaling. Among these estimators, nonparametric regression methods such as artificial neural networks (18), splines (19), kernels (5), and wavelets have been utilized (20). Spline is one of the estimators with better flexibility in estimating regression functions than other nonparametric estimators (21). Therefore, this research utilizes quantile regression with the nonparametric spline estimator to assess extreme changes in rainfall.

A high multicollinearity issue among predictor variables was identified in rainfall research using a global climate model. This issue was addressed by employing a penalized spline estimator, which incorporates a penalty function to tackle multicollinearity problems (22, 23). The penalized spline has estimation criteria consisting of a goodness of fit function and a penalty function (24). In addition to the penalized spline estimator, another estimator, such as the truncated spline, was utilized. The capability of the truncated spline lies in its ability to identify knot points that indicate changes in the data behavior patterns within different intervals (21). The application of truncated spline has been extended by incorporating principal component analysis to address multicollinearity issues in the data (23). Principal component analysis has been employed in statistical downscaling for daily rainfall and monthly rainfall conditions (25,26). Therefore, this study adopts a linear truncated spline approach to highlight changes in rainfall patterns in statistical downscaling, and principal component analysis is employed to address multicollinearity issues.

Materials and Methods

This study used global climate model data as predictor variables, totaling 64 variables. The response variable used is the rainfall data for Pangkep Regency from January 2008 to December 2022, consisting of 180 data points. This rainfall data represents the average rainfall from three rainfall stations: Bungoro, Ma'rang, and Labakkang, obtained from the Meteorology, Climatology, and Geophysics Agency Station Region IV Makassar.

Suppose given data $(\mathrm{y_i}, \mathrm{x_{ij}}), \mathrm{i} = 1,2, \dots, \mathrm{n};\ \mathrm{j} = 1,2, \dots, \mathrm{p}$ where y_i is the response variable and x_{ij} is the predictor variable. Global circulation model data, which serve as predictor variables, were time-lagged to obtain the highest correlation values with rainfall data using the cross-correlation function (8). The regression model formed from the response and predictor variables is written as follows (26):

$$y_i = f(x_{ij}) + \varepsilon_i \tag{1}$$

Multicollinearity testing is conducted based on the variance inflation factor (VIF) values. To address this issue, principal component analysis is employed. The data obtained from the principal component analysis is represented as $(y_i, w_{ik}), i = 1, 2, \dots, n$ and $k = 1, 2, \dots, q$ for $k < j$ with w_{ik} representing the principal component scores. Furthermore, outlier testing is conducted based on the Mahalanobis distance test. In addressing outliers, quantile regression is employed with a quantile value $\tau \in (0,1)$. When Eq. (1) is expressed in the form of a quantile regression model with principal component analysis, it is obtained as follows:

$$y_i(\tau) = f(w_{i1}) + f(w_{i2}) + \cdots f(w_{iq}) + \varepsilon_i(\tau) \tag{2}$$

The relationship pattern between rainfall and the main irregularly formed components in the regression function is modeled using a linear truncated nonparametric spline. The spline has an order of 1 and knot points, denoted as k_h, as follows:

$$\begin{aligned} f(w_{i1}) &= \beta_{01}(\tau) + \beta_{11}(\tau) w_{i1} + \sum_{h=1}^{r} \beta_{(1+h)1}(\tau)(w_{i1} - k_{h1})_+ \\ f(w_{i2}) &= \beta_{02}(\tau) + \beta_{12}(\tau) w_{i2} + \sum_{h=1}^{r} \beta_{(1+h)2}(\tau)(w_{i2} - k_{h2})_+ \\ f(w_{iq}) &= \beta_{0q}(\tau) + \beta_{1q}(\tau) w_{iq} + \sum_{h=1}^{r} \beta_{(1+h)q}(\tau)(w_{iq} - k_{hq})_+ \end{aligned} \tag{3}$$

The function $(w_i - k_h)_+$ in Eq. (3) is a truncated linear function described as follows:

$$(w_i - k_h)_+ = \begin{cases} (w_i - k_h), & w_i \geq k_h \\ 0, & w_i < k_h \end{cases}$$

If Eq. (3) is expanded, the result is as follows:

$$f(w_{i1}) = \beta_{01}(\tau) + \beta_{11}(\tau)w_{i1} + \beta_{(1+1)1}(\tau)(w_{i1} - k_{11})_+ + \cdots + \beta_{(1+r)1}(\tau)(w_{i1} - k_{r1})_+ \quad (4)$$

$$f(w_{i2}) = \beta_{02}(\tau) + \beta_{12}(\tau)w_{i2} + \beta_{(1+1)2}(\tau)(w_{i2} - k_{12})_+ + \cdots + \beta_{(1+r)2}(\tau)(w_{i2} - k_{r2})_+$$

$$f(w_{iq}) = \beta_{0q}(\tau) + \beta_{1q}(\tau)w_{iq} + \beta_{(1+1)q}(\tau)(w_{iq} - k_{1q})_+ + \cdots + \beta_{(1+r)q}(\tau)(w_{iq} - k_{rq})_+$$

Next, Eq. (4) is substituted into Eq. (2), then it is obtained

$$y_i(\tau) = \beta_{01}(\tau) + \beta_{11}(\tau)w_{i1} + \beta_{(1+1)1}(\tau)(w_{i1} - k_{11})_+ + \cdots + \beta_{(1+r)1}(\tau)(w_{i1} - k_{r1})_+ + \beta_{02}(\tau) + \beta_{12}(\tau)w_{i2} + \beta_{(1+1)2}(\tau)(w_{i2} - k_{12})_+ + \cdots + \beta_{(1+r)2}(\tau)(w_{i2} - k_{r2})_+ + \cdots + \beta_{0q}(\tau) + \beta_{1q}(\tau)w_{iq} + \beta_{(1+1)q}(\tau)(w_{iq} - k_{1q})_+ + \cdots + \beta_{(1+r)q}(\tau)(w_{iq} - k_{rq})_+ + \varepsilon_i(\tau) \quad (5)$$

So, Eq. (5), which is the principal component linear spline quantile regression model, when expressed in matrix form, is as follows:

$$\boldsymbol{y}(\tau) = \boldsymbol{W}[k]\boldsymbol{\beta}(\tau) + \boldsymbol{\varepsilon}(\tau)$$

The parameter estimation in the principal component linear spline quantile regression is obtained by minimizing the sum of absolute errors. In quantile regression, errors are assigned different weights, where a weight of τ is given to non-negative errors, while a weight of $(1-\tau)$ is given to negative errors, ensuring that the quantile obtained matches the specified τ. Therefore, the parameter estimation β is obtained by minimizing the following estimation criteria:

$$\hat{\beta}(\tau) = \min_{\beta}\left\{\tau\sum_{i=1,\varepsilon_i\geq 0}^{n}|y_i - \boldsymbol{w}_i'[k]\boldsymbol{\beta}(\tau)| + (1-\tau)\sum_{i=1,\varepsilon_i<0}^{n}|y_i - \boldsymbol{w}_i'[k]\boldsymbol{\beta}(\tau)|\right\}$$

Finding a solution for parameter estimation β can be done both analytically and numerically. One commonly used numerical method to solve it is the simplex method. The best linear spline quantile regression model depends on the optimal knot point. The method often used to find the optimal knot point is the Generalized Cross-Validation (GCV). The GCV value that provides the optimal knot point is the minimum GCV value. The formula for GCV used is as follows:

$$GCV(k) = \frac{MSE(k)}{(n^{-1}trace[\boldsymbol{I} - \boldsymbol{A}(k)])^2}$$

with $MSE(k) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i(\tau))^2$, $\boldsymbol{A}[k] = \boldsymbol{W}[k](\boldsymbol{W}[k]'\boldsymbol{W}[k])^{-1}\boldsymbol{W}[k]'$ and $\boldsymbol{I}$ is the identity matrix.

Results and Discussion

The average rainfall in Pangkep Regency from 2008 to 2022 is 311.22 mm/month. The highest rainfall occurred in January at 1540.50 mm/month, and the lowest rainfall was 0.00 mm/month in July, August, September, and October. Based on the standard deviation, the rainy season with the highest value is in January, and the lowest standard deviation occurs at the peak of the dry season in August. A high standard deviation in January indicates that rainfall in January from 2008 to 2022 is very variable. Based on the average rainfall values, January and December have the highest rainfall compared to other months.

Figure 1 shows that the rainfall in Pangkep Regency is concave, resembling a U-pattern. This pattern is one

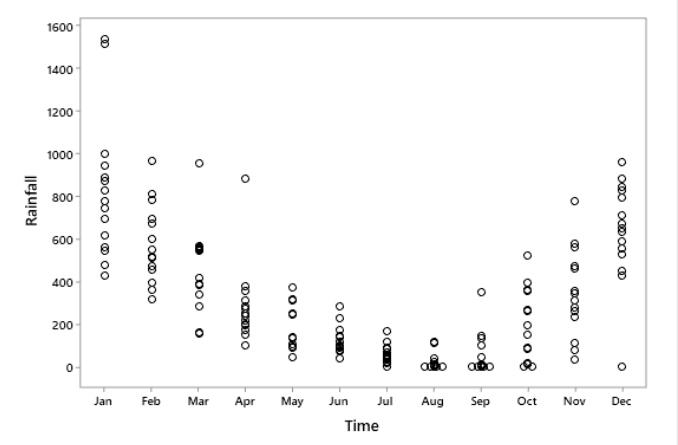

Figure 1. Plot of rainfall in Pangkep Regency 2008 to 2022

of Indonesia's three types of rainfall patterns. The monsoon rainfall type is the pattern observed in Pangkep Regency, characterized by a single peak rainy season typically occurring from October to March and a dry season from April to September. The peak of the rainy season in January has an average of 829.60 mm/month, which decreases in the following months, reaching the lowest average rainfall in August at 26.46 mm/month. Then, in the subsequent months, it increases again, peaking in December with an average of 636.00 mm/month.

The predictor variables exhibit a strong correlation among themselves, as indicated by the VIF values ranging from 28.02 to 4612.10. Additionally, the response variable contains outliers, constituting 25% of the utilized data. Therefore, quantile regression with principal components will be employed to analyze the data due to the presence of multicollinearity and outliers. To determine the main components to be used in the analysis stage is done by selecting the main components with a cumulative diversity proportion of at least 85%. Table 1 shows that the first eigenvalue has explained 83.24% of the sample variance, and the second eigenvalue is 5.15%. Thus, utilizing two main components has explained more than 85% of the total diversity.

A scatter plot visualizes the relationship between the variables under investigation. Through a scatter plot, we can initially discern the nature of the relationship between rainfall data and the two main components. The visualization of this plot serves as the foundation for regression model development. If the relationship between the data does not follow common patterns like linear, quadratic, or cubic, nonparametric methods can be employed. Estimating the linear spline quantile model with principal components involves determining the appropriate number of knot points. The approach begins with one-knot point, then two-knot points, and three-knot points. The addition of knot points is done to enhance the quality of estimation. If the optimal knot point can be identified, it will result in an optimal model. The most suitable knot point is selected by considering the GCV value for each knot point on the selected component.

The regression curve for the linear spline quantile regression model with the principal component is shown in Figure 2. Each component used is approximated with several knot points, one-knot point, two-knot points, and three-knot points. Each component also uses three quantile values: 0.25, 0.50, and 0.75. The blue line represents the 0.25 quantile value, the green line represents the 0.50 quantile value, and the red line represents the 0.75 quantile value. The first and second components with a single knot point indicate two data change patterns for each quantile value. The first pattern of change occurs before the first component reaches -10.20 points, experiencing a decrease in rainfall. After the first component reaches -10.20 points, a second pattern change occurs, which is also a decrease in rainfall. As for the second component, the first pattern of change occurs before the second component reaches -0.30 points, experiencing an increase in rainfall. After the second component reaches -0.30 points, a second pattern change occurs a decrease in rainfall.

The first component and the second component with two knot points indicate three data change patterns for each quantile value. The first pattern of change occurs before the first component reaches -6.90 points, experiencing a decrease in rainfall. After the first component reaches -6.90 points, a second pattern change occurs, which is also a decrease in rainfall. After the first component reaches 0.10 points, a third pattern change occurs, which is also a decrease in rainfall. As for the second component, the first pattern of change occurs before the second component reaches -1.60 points, experiencing a decline in rainfall. After the second component reaches -1.60 points, a second pattern change occurs, an increase in rainfall. After the second component reaches -0.10 points, a third pattern change occurs a decrease in rainfall.

The first component and the second component with three knot points indicate that there are four patterns of data change for each quantile value. The first pattern change occurs before the first component reaches -8.30

Table 1. Eigenvalues and cumulative variance proportions

Principal component	Eigenvalues	Variance proportions	Cumulative variance proportions
1	53.27	83.24%	83.24%
2	3.30	5.15%	88.39%
3	2.09	3.27%	91.66%
4	1.48	2.31%	93.97%
5	0.87	1.36%	95.33%
⋮	⋮	⋮	⋮
64	0.00	0.00%	100.00%

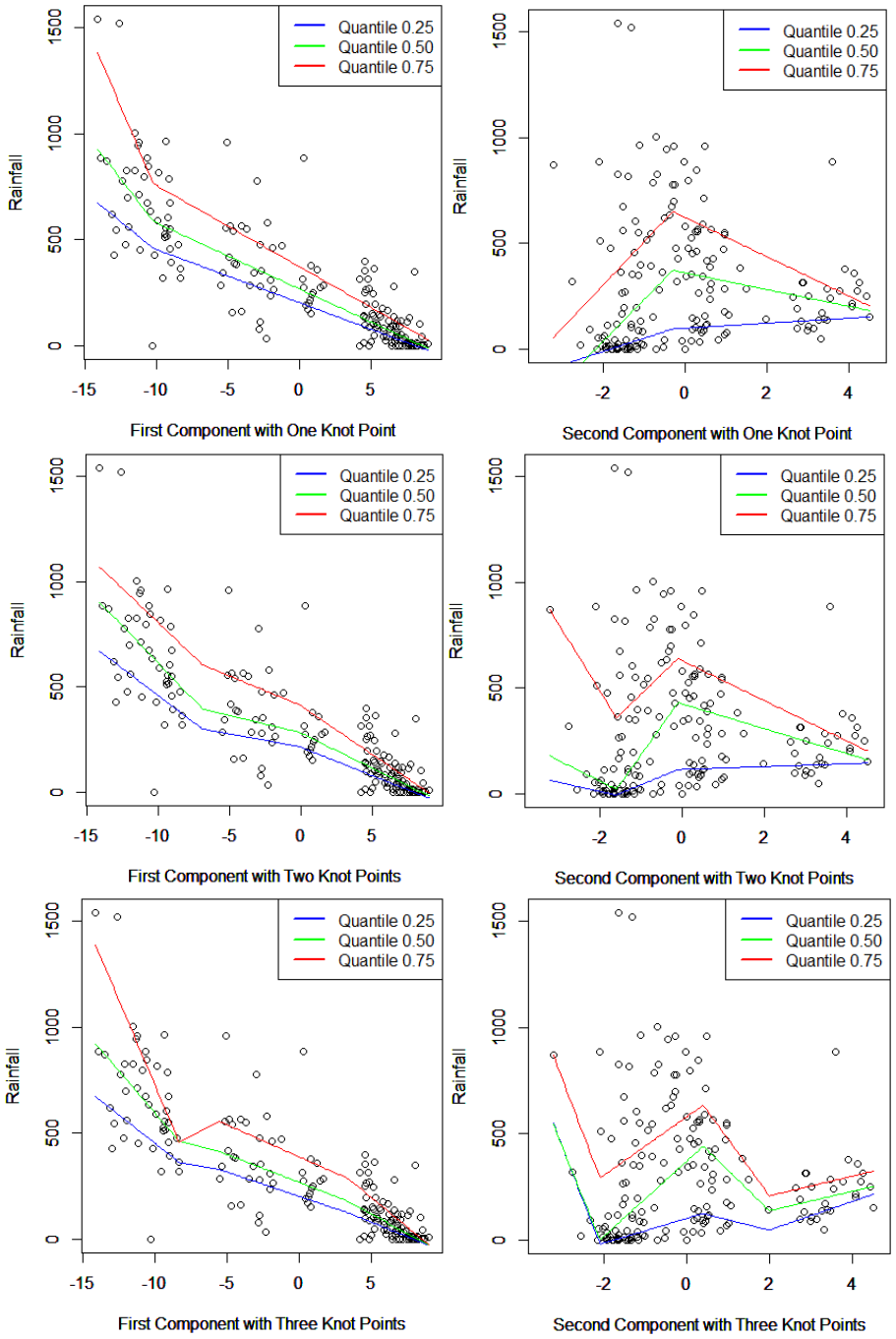

Figure 2. The principal component linear spline quantile regression curve

points, experiencing a decrease in rainfall. After the first component reaches -8.30 points, the second pattern change occurs an increase in rainfall. After the first component reaches -5.30 points, the third pattern change occurs a decrease in rain again. After the first component reaches 3.40 points, the fourth pattern change occurs, which is also a decrease in rainfall. As for the second component, the first pattern change occurs before the second component reaches -2.10 points, experiencing a decline in rainfall. After the second component reaches -2.10 points, the second pattern change occurs an increase in rainfall. After the second component reaches 0.40 points, the third pattern change occurs a decrease in the rain again. After the

Table 2. The GCV value in principal component linear spline quantile regression

Quantile	GCV		
	1 Knot point	2 Knot points	3 Knot points
0.25	47.71	48.74	49.86
0.50	32.33	33.78	33.51
0.75	50.04	46.12	49.58

second component reaches 2.00 points, the fourth pattern change occurs an increase in rainfall again.

Table 2 shows the GCV values for each principal component's linear spline quantile regression model. The regression model with a one-knot point with the lowest GCV value is at the 0.50 quantile, 32.33. In the regression model with two-knot points, the lowest GCV value is at the 0.50 quantile of 33.78. In the regression model with three-knot points, the lowest GCV value is at the 0.50 quantile, which is 33.51. This indicates that the 0.50 quantile is better for each different use of knot points. Thus, it is found that the best model is at the 0.50 quantile and one-knot point with the lowest GCV value compared to other quantiles and knot points. This result indicates that this model is more effective in explaining the variation of data and the impact of the two components used on rainfall in Pangkep Regency. The following is the best model produced from the principal component linear spline quantile regression in statistical downscaling:

$$\hat{y}_i(0.50) = -518.89 - 105.53w_{i1} + 76.37(w_{i1} - (-10.20))_+ + 10.91w_{i2} - 6.79(w_{i2} - (-0.30))_+$$

The equation will be used for model validation to assess how closely the obtained model matches the actual measurements. The root mean square error of prediction (RMSEP) and the correlation between the measured rainfall data and predicted rainfall are used as evaluation parameters in this validation process. Rainfall data from 2021 to 2022 is utilized as test data for validation. The model produced a correlation value of 0.89, indicating that the actual and estimated rainfall have a relatively strong linear relationship with a strength of 0.89. The RMSEP value obtained from the model is 153.16. This RMSEP value can be seen in Figure 3, which shows that there are several estimated rainfalls with significant differences in capturing the actual rain, such as in March 2021, October 2021, February 2022, May 2022, June 2022, October 2022, and November 2022.

Figure 3 shows that in January 2021, January 2022, March 2022, and April 2022, the principal component linear spline quantile model in statistical downscaling estimated the rainfall value to be higher than its actual value. Whereas in other months, namely February to December 2021, February 2022, and May to December 2022, the principal component linear spline quantile model in statistical downscaling estimated the rainfall value to be lower than its actual value. The model was

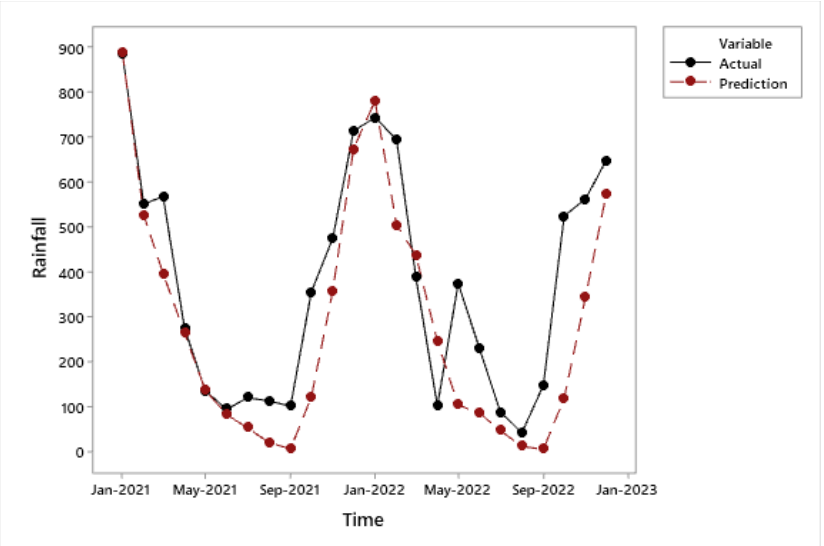

Figure 3. The actual and estimated rainfall plot for 2021 to 2022

able to accurately assess actual rainfall in January 2021, February 2021, April 2021, May 2021, June 2021, December 2021, January 2022, March 2022, July 2022, and August 2022 with a smaller difference to the actual rainfall compared to other months.

Conclusion

The relationship between variables that have a global impact from global climate model data with variables that have a local effect from rainfall data in statistical downscaling cannot be directly used in regression modeling. This is due to the presence of multicollinearity problems and outliers in the data. Quantile regression with principal component analysis is used to address these issues. The relationship pattern between response variables and predictors is not parametric; therefore, a nonparametric approach, namely truncated linear spline, is utilized. The results of this modeling provide fairly accurate estimates of rainfall. Using a single knot point at the 0.50 quantile produces more effective results in data modeling than using two or three-knot points at several other quantiles. This is seen based on the lowest GCV value. Additionally, this model can accurately forecast actual rainfall in the years 2021 to 2022 with a correlation level of 0.89.

References

1. Chaudhuri RR, Sharma P. An integrated stochastic approach for extreme rainfall analysis in the National Capital Region of India. J Earth Syst Sci. 2021;130(16):1–15.
2. Tunas IG, Azikin H, Oka GM. Impact of Extreme Rainfall on Flood Hydrographs. IOP Conf Ser Earth Environ Sci. 2021;884(1):1–6.
3. Mohsenipour M, Shahid S, Ziarh GF, Yaseen ZM. Changes in monsoon rainfall distribution of Bangladesh using quantile regression model. Theor Appl Climatol. 2020;142:1329–42.
4. Keller AA, Garner KL, Rao N, Knipping E, Thomas J. Downscaling approaches of climate change projections for watershed modeling: Review of theoretical and practical considerations. PLOS Water. 2022;1(9):1–20.
5. Mulyati AE, Wigena AH, Djuraidah A. Statistical Downscaling Using Kernel Quantile Regression to Predict Extreme Rainfall. Int J Sci Basic Appl Res. 2018;42(2):1–9.
6. Singh P, Shamseldin AY, Melville BW, Wotherspoon L. Development of statistical downscaling model based on Volterra series realization, principal components and ridge regression. Model Earth Syst Environ. 2023;9(3):3361–80.
7. Kalu I, Ndehedehe CE, Ferreira VG, Janardhanan S, Currell M, Kennard MJ. Statistical downscaling of GRACE terrestrial water storage changes based on the Australian Water Outlook model. Sci Rep. 2024;14(1):1–19.
8. Sahriman S, Djuraidah A, Wigena AH. Application of Principal Component Regression with Dummy Variable in Statistical Downscaling to Forecast Rainfall. Open J Stat. 2014;04(09):678–86.
9. Lu Z, Guo Y, Zhu J, Kang N. Seasonal forecast of early summer rainfall at stations in south china using a statistical downscaling model. Weather Forecast. 2020;35(4):1633–43.
10. Pahlavan HA, Zahraie B, Nasseri M, Varnousfaderani AM. Improvement of multiple linear regression method for statistical downscaling of monthly precipitation. Int J Environ Sci Technol. 2018;15(9):1897–912.
11. Salimi AH, Samakosh JM, Sharifi E, Hassanvand MR, Noori A, Rautenkranz H Von. Optimized artificial neural networks-based methods for statistical downscaling of gridded precipitation data. Water (Switzerland). 2019;11(1653):1–20.
12. Sharifi E, Steinacker R, Saghafian B. Multi time-scale evaluation of high-resolution satellite-based precipitation products over northeast of Austria. Atmos Res. 2018;206:1–33.
13. Uranchimeg S, Kwon HH, Kim B, Kim TW. Changes in extreme rainfall and its implications for design rainfall using a Bayesian quantile regression approach. Hydrol Res. 2020;51(4):699–719.
14. Malik N, Bookhagen B, Mucha PJ. Spatiotemporal patterns and trends of Indian monsoonal rainfall extremes. Geophys Res Lett. 2016;43(4):1710–7.
15. Wagesho N, Claire M. Analysis of Rainfall Intensity-Duration-Frequency Relationship for Rwanda. J Water Resour Prot. 2016;8(7):706–23.
16. Shiau JT, Huang WH. Detecting distributional changes of annual rainfall indices in Taiwan using quantile regression. J Hydro-Environment Res. 2014;9(3):1–13.
17. Wigena AH, Djuraidah A, Rizki A. Semiparametric modeling in statistical downscaling to predict rainfall. Appl Math Sci. 2015;9(88):4371–82.
18. Sharifi E, Saghafian B, Steinacker R. Downscaling Satellite Precipitation Estimates With Multiple Linear Regression, Artificial Neural Networks, and Spline Interpolation Techniques. J Geophys Res Atmos. 2019;124(2):789–805.
19. Islamiyati A. Spline Longitudinal Multi-response Model for the Detection of Lifestyle- Based Changes in Blood Glucose of Diabetic Patients. Curr Diabetes Rev. 2021;18(7):98–104.
20. Sachindra DA, Ahmed K, Rashid MM, Sehgal V, Shahid S, Perera BJC. Pros and cons of using wavelets in conjunction with genetic programming and generalised linear models in statistical downscaling of precipitation. Theor Appl Climatol. 2019;138(1):617–38.
21. Lestari B, Fatmawati, Budiantara IN, Chamidah N. Estimation of Regression Function in Multi-Response Nonparametric Regression Model Using Smoothing Spline and Kernel Estimators. J Phys Conf Ser. 2018;1097(1):1–9.
22. Goldameir NE, Djuraidah A, Wigena AH. Quantile Spline Regression on Statistical Downscaling Model to Predict Extreme Rainfall in Indramayu. Appl Math Sci. 2015;9(126):6263–72.
23. Islamiyati A, Kalondeng A, Sunusi N, Zakir M, Amir

AK. Biresponse nonparametric regression model in principal component analysis with truncated spline estimator. J King Saud Univ - Sci. 2022;34(3):1–9.

24. Islamiyati A, Sunusi N, Kalondeng A, Fatmawati F, Chamidah N. Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model. J Sci Islam Repub Iran. 2020;31(2):175–83.
25. Saputra MD, Hadi AF, Riski A, Anggraeni D. Principal Component Regression in Statistical Downscaling with Missing Value for Daily Rainfall Forecasting. Int J Quant Res Model. 2021;2(3):139–46.
26. Sari WJ, Wigena AH, Djuraidah A. Quantile regression with functional principal component in statistical downscaling to predict extreme rainfall. Int J Ecol Econ Stat. 2017;38(1):1–9.