BAREKENG: Journal of Mathematics and Its Applications
June 2026 Volume 20 Issue 2 Page 1643–1658
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss2pp1643-1658

STATISTICAL MODELING FOR DOWNSCALING USING PRINCIPAL COMPONENT REGRESSION AND DUMMY VARIABLES: A CASE OF SIAK DISTRICT

Arisman Adnan[1*], Elsa Riesta Alika[2], Divo Dharma Silalahi[3], Felia Rizki Aulia[4], Gustriza Erda[5]

[1,2,5]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Riau*
Kampus Bina Widya KM. 12,5, Pekanbaru, 28293, Indonesia

[3,4]*The SMART Research Institute (SMARTRI)*
Kandis, Siak Regency, 28686, Indonesia

*Corresponding author's e-mail: * **arisman.adnan@lecturer.unri.ac.id***

Article Info

Article History:

Received: 13th July 2025
Revised: 15th August 2025
Accepted: 16th October 2025
Available Online:26th January 2026

Keywords:

Global Circulation Model (GCM);
Multicollinearity;
Prediction;
Principal Component;
Regression;
Statistical Downscaling.

ABSTRACT

Indonesia, as a tropical country, is characterized by two primary seasons: the rainy season and the dry season. It is evident that meteorological shifts can exert considerable influence on the agricultural sector, a notable example being the cultivation of palm oil. Consequently, the ability to predict rainfall has emerged as a pivotal element in the broader endeavor to mitigate the adverse effects of climate change. This study employs statistical downscaling using the Principal Component Regression (PCR) approach to model rainfall predictions. The issue of multicollinearity, a common occurrence in Global Circulation Model (GCM) data, is addressed through the use of Principal Component Regression (PCR). This method has been demonstrated to stabilize the model structure and reduce variance in the regression coefficients. The data utilized encompass observed rainfall from LIBO Estate, which is owned by PT SMART Tbk (SMART Research Institute), for the period from 2013 to 2022. This data serves as the response variable, while the CMIP6 GCM simulation output data functions as the predictor variable. The findings indicated that the initial PCR model exhibited an RMSE value ranging from 97.06 to 131.69, along with an R^2 value ranging from 14.25% to 20.49%. The incorporation of dummy variables into the model resulted in a substantial enhancement in its performance, as evidenced by a decline in RMSE to 24.46–35.83 and an increase in R^2 to 89.02%–90.24%. The findings indicate that the use of PCR with dummy variables is an effective approach for enhancing the accuracy of rainfall modeling through statistical downscaling.

How to cite this article:

A. Adnan, E. R. Alika, D. D. Silalahi, F. R. Aulia and G. Erda., "STATISTICAL MODELING FOR DOWNSCALING USING PRINCIPAL COMPONENT REGRESSION AND DUMMY VARIABLES: A CASE OF SIAK DISTRICT", *BAREKENG: J. Math. & App.*, vol. 20, no. 2, pp. 1643-1658, Jun, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Indonesia is an agrarian country, with its economy heavily reliant on the agricultural sector, including oil palm plantations. Climate change has increasingly affected agricultural productivity worldwide, with shifts in rainfall patterns posing significant challenges to crop management and yield stability [1]. Among various agricultural commodities, oil palm cultivation serves as a critical case study due to its economic significance and vulnerability to climatic variability [2]. Climate variability has been shown to reduce fresh fruit bunch (FFB) yield, disrupt flowering cycles, and increase susceptibility to pests and diseases [3], making oil palm a particularly vulnerable crop in the face of ongoing climate change [4]. As a pivotal national commodity, oil palm production exhibits heightened sensitivity to meteorological conditions, particularly the accessibility of water from precipitation [5]. Consequently, seasonal fluctuations in rainfall can disrupt productivity [6], especially during periods of drought [7].

Rainfall is a key climatic factor influencing the growth of oil palm. It is imperative to ensure sufficient water availability from rainfall to maintain optimal soil moisture levels in the root zone, as oil palms exhibit substantial water demands [8]. The selection of oil palm as the focus of this study is particularly relevant given Indonesia's position as the world's largest palm oil producer, contributing approximately 58% of global production, making the sector's climate resilience crucial for both national economic stability and global supply chain [6]. Conversely, water deficits resulting from low rainfall can adversely affect yield and disrupt the physiological development of the crop [9].

To anticipate the impacts of climate change and ensure the sustainability of the agricultural sector, it is essential to utilize an accurate weather prediction system. A prevalent approach involves the use of Global Circulation Model (GCM) data for rainfall modeling, a technique that provides large-scale numerical climate forecasts [10]. GCMs are widely recognized as essential tools for projecting future rainfall patterns under various climate scenarios [11], though their coarse resolution often necessitates statistical downscaling to enhance their applicability at the local level [12]. However, due to their limited spatial and temporal resolution, GCM outputs necessitate downscaling to the regional level, which is typically accomplished through statistical downscaling (SD) methods [13].

Statistical downscaling is a process that links global climate data (predictors) with local data (responses) through statistical approaches such as linear regression. This technique addresses a critical limitation of GCMs, which operate at coarse spatial resolutions (typically 100-300 km) that are insufficient to capture local climate variability and site-specific precipitation patterns. The downscaling process involves three key steps: (1) identifying relevant large-scale atmospheric predictors from GCM outputs, (2) establishing statistical relationships between these predictors and observed local climate variables using historical data, and (3) applying these relationships to generate local-scale climate predictions. Furthermore, this approach facilitates the investigation of relationships between global-scale data and local-scale data over a designated time period [14]. By transforming coarse-resolution GCM data into high-resolution local estimates, downscaling enables more accurate and actionable climate predictions for agricultural planning and water resource management at regional scales.

However, a notable challenge in implementing SD is the issue of multicollinearity, which arises due to the high correlation among numerous predictor variables [15]. This can result in high variance in estimated parameters and reduce model accuracy. The decline in precision inherent to statistical downscaling methodologies can adversely affect the reliability of climate estimations at particular locations, thereby necessitating the mitigation of multicollinearity to ensure optimal outcomes.

A prevalent methodology for addressing multicollinearity is Principal Component Regression (PCR). This approach integrates Principal Component Analysis (PCA) with linear regression, a technique that has been employed to reduce dimensionality and eliminate correlations among predictor variables [16]. Principal component analysis (PCA) is a statistical technique that transforms a set of correlated predictors into a new set of uncorrelated principal components (PCs). These PCs are then incorporated into a regression model to assess their influence on the response variable [17].

Research on statistical downscaling using Principal Component Regression (PCR) for rainfall forecasting has previously been conducted by [7]. The study incorporated dummy variables to enhance the accuracy of rainfall prediction models in Indramayu. The utilization of GCM data, arranged in a 6×4 grid spanning the period 1979–2007, The utilization of GCM data, arranged in a 6×4 grid spanning the period 1979–2007, revealed multicollinearity issues among the predictors. The findings demonstrated that the

combination of PCR with dummy variables yielded more accurate rainfall estimates compared to models devoid of dummy variables. The incorporation of dummy variables addresses a key limitation of standard PCR models, which assume a linear relationship between predictors and response variables across all observations. However, rainfall patterns often exhibit distinct regimes or clusters that may respond differently to the same atmospheric predictors. Dummy variables, generated through K-means clustering based on rainfall intensity groupings, enable the model to capture these non-linear regime-dependent relationships by allowing different intercepts for each rainfall cluster while maintaining the same slope coefficients from principal components. This approach effectively partitions the data into homogeneous groups with similar rainfall characteristics, thereby improving model fit and prediction accuracy without violating the multicollinearity assumptions already addressed by PCR.

The optimal model employed a single principal component, achieving a high correlation value (0.99) and a Root Mean Square Error of Prediction (RMSEP) of 28.84 millimeters. Furthermore, [18] employed dummy variables generated through K-means clustering in combination with Liu-Ridge Regression (LLR) to address multicollinearity. The study's findings indicated that incorporating dummy variables led to a substantial enhancement in the model's accuracy, with an observed improvement of up to 15%.

Another study was conducted by [14]. The present study applied the PCR in the Statistical Downscaling (PPSD) method to predict daily rainfall in Kupang City, including the handling of missing values. The most optimal outcomes indicated that the model comprising 11 primary components from the 6x6 grid domain (with cumulative variance of 94.01%) yielded high precision, exhibiting a mean absolute percentage error (MAPE) of 2.81% and a root mean square error (RMSE) of 10.81 millimeters. [19] also employed the use of SD and PCR to assess climate change in the Cauvery River basin, India. The present study sought to compare the projected results of various CMIP5 GCMs in terms of rainfall and temperature with observational data that has undergone validation at local stations. The PCR model demonstrated satisfactory performance, with determination coefficients (R^2) ranging from 70% to 83% across various GCM satellite scenarios.

The present study aims to develop a local-scale rainfall prediction model using statistical downscaling and principal component regression (PCR) methods. The model is based on GCM data from CMIP6 (MPI-ESM1-2-HR) with a spatial resolution of 100 km. The model is based on GCM data from CMIP6 (MPI-ESM1-2-HR) with a spatial resolution of 100 km. This research makes a significant methodological contribution to the field of climate modeling by integrating K-means clustering-derived dummy variables into the PCR framework, effectively addressing both multicollinearity and non-linear regime-dependent rainfall patterns simultaneously, a combination that has received limited attention in existing downscaling literature. Model performance is evaluated using multiple metrics: the coefficient of determination (R^2) and Root Mean Squared Error (RMSE). It is anticipated that the resulting model, particularly the PCR-dummy variable approach, will contribute to mitigating risks associated with climate change in oil palm agriculture and support more effective water resource management in the future.

The novelty of this research lies in the incorporation of dummy variables derived from k-means clustering of rainfall data to enhance model accuracy, demonstrating a hybrid approach that combines dimensionality reduction with categorical partitioning to improve predictive performance. While this study focuses on oil palm cultivation in Siak Regency, the proposed methodology is generalizable and can be adapted to other agricultural systems, hydrological applications, and climate-sensitive sectors across various geographical regions where accurate local-scale rainfall prediction is crucial for informed decision-making. It is anticipated that the resulting model will contribute to mitigating risks associated with climate change in oil palm agriculture and support more effective water resource management in the future.

2. RESEARCH METHODS

2.1 Data Sources

The observational data utilized in this study encompasses daily rainfall data, which is subsequently aggregated into monthly rainfall data (in units of millimeters per month). The rainfall variable was employed as the response variable in the modeling. The data was obtained from the Automated Weather Station (AWS) owned by PT SMART Tbk, SMART Research Institute Division, located at LIBO Estate, for the period from 2013 to 2022. LIBO Estate operates three distinct AWS locations: The geographical locations in question

correspond to the following divisions: Division 2 (latitude 0.9543°; longitude 101.2167°), Division 3 (latitude 0.9471°; longitude 101.1832°), and Division 7 (latitude 0.9264°; longitude 101.2061°).

The predictor variables are derived from precipitation data from the Global Circulation Model (GCM) based on CMIP6 simulations, obtained from the website https://aims2.llnl.gov/search/cmip6 **[20]**. The GCM data utilized in this study possesses a nominal spatial resolution of 100 km (0.9375° × 0.9349°). These data are derived from the MPI-ESM1-2-HR satellite model. Subsequently, the global data underwent a series of processing stages, including interpolation and clipping, to obtain data relevant to the study area, specifically Siak Regency. In this process, an 8 × 8 grid was formed with a resolution of 0.1° × 0.1° per grid, resulting in 64 grid points as predictor variables. Each grid encompasses 120 monthly observation data points, which will be utilized in the statistical modeling process.

2.2 Analysis Method

The present study employs statistical downscaling, a statistical method that describes the relationship between global-scale data and local-scale data within a specified time period. This relationship can be expressed as follows [21]:

$$y_{(t\times 1)} = f\left(X_{t\times g}\right), \tag{1}$$

with $y_{(t\times 1)}$ is the local climate variable (response variable), $X_{t\times g}$ is the GCM output variable (predictor variable), t is the time period (monthly), and g is the number of GCM output grid domains. [21] shows that the 8×8 domain is better than the other domains of sizes 10×10, 12×12, 14×14, and 16×16. An illustration of the statistical downscaling process is presented as follows [22]:

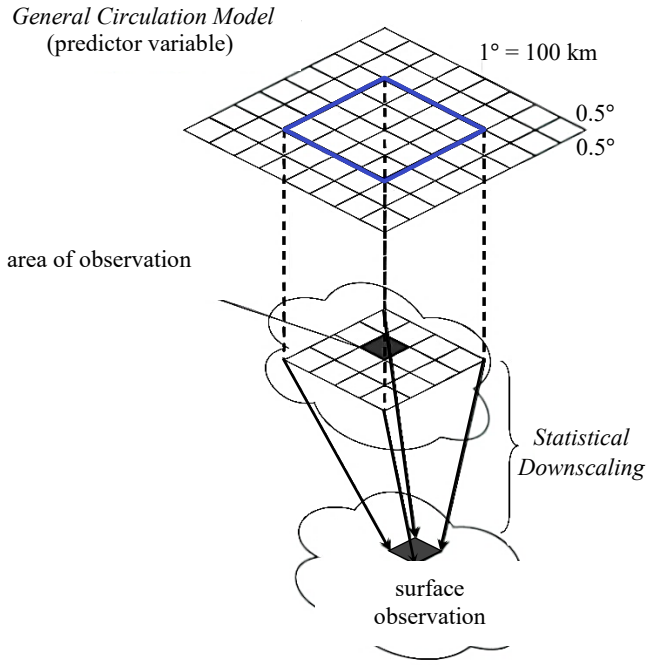


Figure 1. Illustration of Statistical Downscaling
Source: [22]

The primary objective of the present study was to employ the Principal Component Regression (PCR) method to reduce the dimensions of monthly rainfall observation data from 2013 to 2022 through statistical downscaling. Prior to PCR modeling, multicollinearity among the 64 continuous GCM predictor variables was assessed using the Variance Inflation Factor (VIF), a widely used diagnostic tool for identifying collinearity in regression analysis [23]. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, resulting in inflated coefficients and unstable parameter estimates. The application of VIF in this study serves to detect and quantify such multicollinearity, which justifies the use of PCR as a dimension reduction technique capable of transforming correlated predictors into orthogonal principal components. Dummy variables representing rainfall categories were added to the final model; these do not introduce multicollinearity concerns since the principal components are orthogonal by construction,

and the categorical variables serve to capture nonlinear rainfall effects without overlapping with the continuous predictors [24]. The Variance Inflation Factor (VIF) method is a statistical technique used to identify multicollinearity.

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}, \tag{2}$$

with R_j^2 is the coefficient of determination$-j$.

Then, Principal Component Analysis (PCA) was performed to reduce the dimension of predictor variables by selecting principal components based on eigenvalue values greater than one, as indicated by the elbow plot analysis. Principal Component Analysis (PCA) is a statistical method used to reduce variables in a case into important features of the principal components. The formation of principal components using a correlation matrix begins with transforming the original variable X into a standard form Z or standardizing the variable [25]:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sqrt{\sigma_j^2}}, \tag{3}$$

with Z_{ij} a is the standardized j variable in i $-$row, X_{ij} the original value of variable j in row i, μ_j is the mean of variable-j, σ_j^2 is the standard deviation of variable -j. The eigenvalue (λ) of the correlation matrix ρ is calculated with the condition:

$$|\rho - \lambda \mathrm{I}| = 0, \tag{4}$$

while the value of *eigenvector* $\boldsymbol{e_j'} = [e_{1j}, e_{2j}, \dots, e_{jp}]$ is calculated using the following formula:

$$(\rho - \lambda \mathrm{I})\mathbf{e_j} = 0. \tag{5}$$

The-j principal component is formed based on the variable $\mathbf{Z}' = [z_1, z_2, \dots, z_p]$ determined using the eigenvector:

$$\mathbf{w_j} = \mathbf{e_j'Z} = e_{j1}z_1 + e_{j2}z_2 + \cdots + e_{jp}z_p, \tag{6}$$

with w_j is the j-th principal component, e_j is the j-th *eigenvector* to j, and z_j is the j-th standardized variable value.

Two approaches are employed in the modeling process: The first approach involves the utilization of a select array of principal components in the PCR model. The second approach incorporates dummy variables, which function as supplementary predictors within the PCR framework. The PCR equation, when reduced to m components, can be expressed as follows:

$$Y = \alpha_0 + \alpha_1 w_1 + \cdots + \alpha_m w_m + \varepsilon, \tag{7}$$

where α_0 is the principal component regression constant, $\alpha_1, \alpha_2, \dots, \alpha_m$ are the principal component regression parameters, and $w_1, w_2, \dots w_m$ are the principal components used.

To enhance the accuracy of the model and address issues of multicollinearity, dummy variables were constructed using k-means clustering based on rainfall groupings. K-means clustering was selected for several reasons: (1) it effectively captures natural groupings in rainfall patterns; (2) it provides an objective, data-driven approach to categorize observations rather than arbitrary threshold selection, and (3) it has been successfully applied in previous statistical downscaling studies to improve model performance by incorporating seasonal or magnitude-based variations. By converting continuous rainfall data into categorical dummy variables, the model can better account for non-linear relationships and regime-specific behaviors that may not be adequately captured by principal components alone.

Dummy variables serve as analytical tools that facilitate the categorization of data into distinct groups based on specific characteristics or attributes [26]. K-means clustering is a non-hierarchical method that groups data according to similarity by predefining the number of clusters and initial centroid values. The algorithmic process involves the iterative updating of centroids until optimal clustering is achieved [27]. This approach has proven effective for rainfall classification. The assessment of cluster quality is facilitated by the utilization of the silhouette coefficient, a metric that quantifies the extent to which each data point aligns with

its designated cluster. The optimal number of clusters (k) is determined by identifying the peak value on the silhouette plot [28], [29]. Subsequent to the establishment of the clusters, dummy variables are generated to represent each identified rainfall group. The model's performance is then evaluated using the coefficient of determination (R^2), the RMSE, and the correlation between predicted and observed rainfall values.

The statistical downscaling methodology employed in this study involves multiple sequential stages, each with distinct objectives and computational procedures. To provide a concise overview of the complete analytical framework, Fig. 2 presents a comprehensive flowchart illustrating the workflow from data collection through model validation.

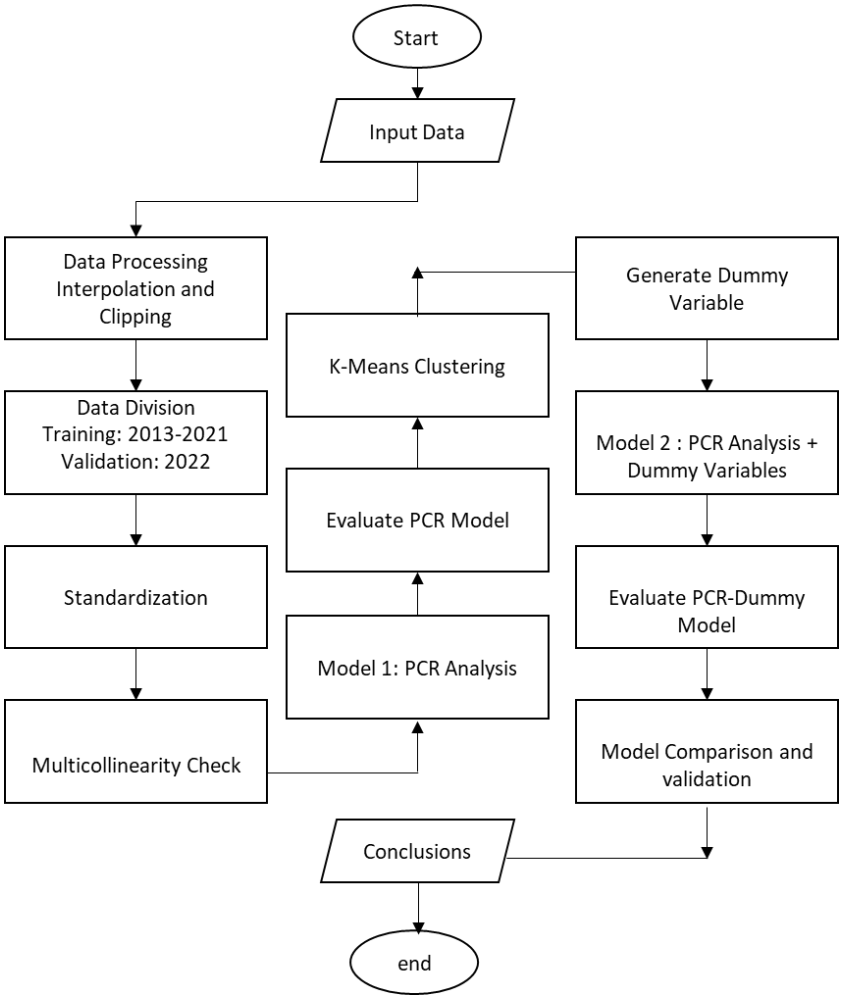


Figure 2. Statistical Downscaling Analysis Workflow

3. RESULTS AND DISCUSSION

3.1 Data Exploration

The rainfall patterns presented in Fig. 3 demonstrate analogous trends across all three weather stations, with the highest average monthly rainfall consistently recorded in November. This consistent seasonal pattern indicates that November typically marks the peak of the rainy season in the study area. The highest average monthly rainfall during the observation period ranged from 153.35 to 172.43 mm, with the maximum monthly rainfall was 587 mm in November 2015. This consistency in rainfall distribution can be attributed to various factors such as wind patterns, topography, and the geographical location of each weather station. The high rainfall in November may also reflect the influence of the La Niña phenomenon, which is often associated with increased rainfall intensity in tropical regions, including Indonesia.

Studies have shown that La Niña tends to increase rainfall over the Maritime Continent, particularly in Southeast Asia and Indonesia, due to enhanced moisture transport and strengthened Walker circulation. For instance, central Pacific La Niña events can trigger severe flooding across Southeast Asia, including parts of

Indonesia, during the September–November period [30]. Another study showed that La Niña events contribute to positive precipitation anomalies across the Maritime Continent, driven by increased humidity and local sea surface temperature gradients [31]. This regional linkage confirms that the intense rainfall observed in November may indeed be influenced by La Niña dynamics.

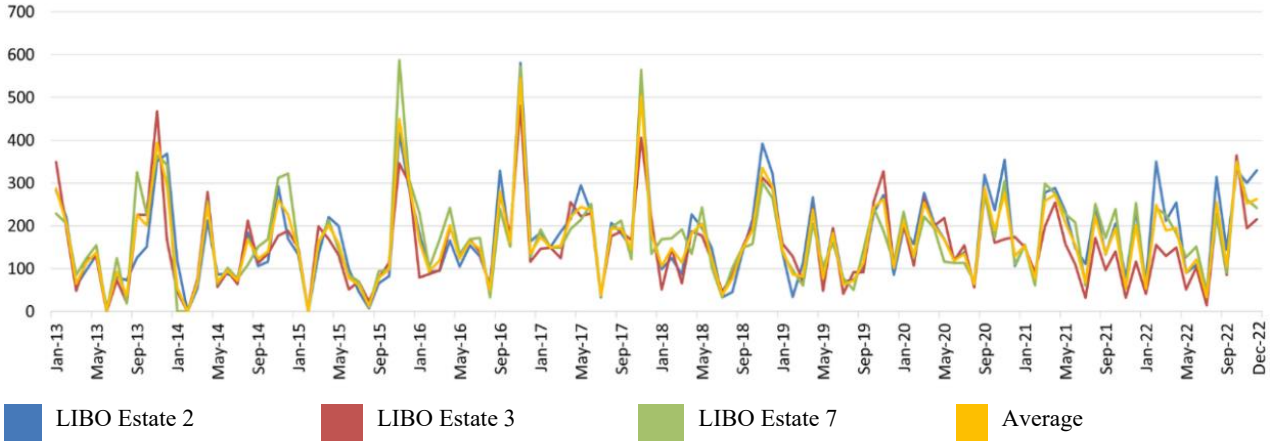


Figure 3. Rainfall patterns at Each Weather Station in LIBO Estate

Based on the data shown in Table 1, rainfall in LIBO Estate 7 exhibits relatively similar observation values to those of weather stations in other divisions. This similarity is due to the location of the weather stations, which are still within the same plantation area and have relatively close coordinates. During the observation period, the highest monthly rainfall recorded was 587 mm in November 2015. Meanwhile, the average monthly rainfall across the entire division ranged from 153.35 to 172.43 mm.

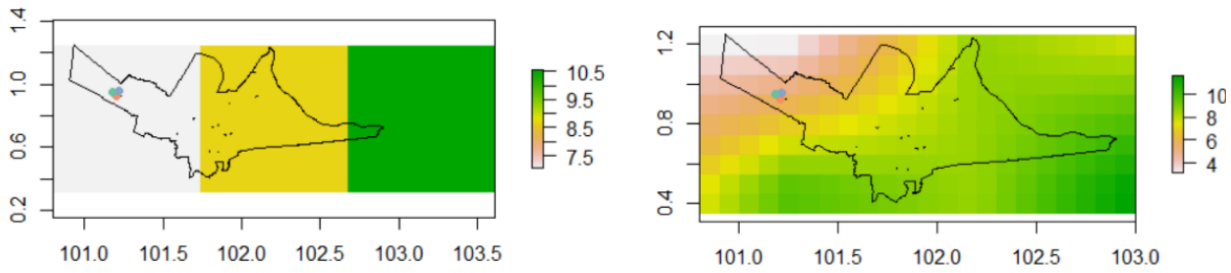
Table 1. Summary of Rainfall Statistics for Each Weather Station

Statistics	LIBO Estate 1	LIBO Estate 3	LIBO Estate 7
Minimum	0	0	0
Median	151.5	147.5	156
Average	172.43	153.35	171.9
Maximum	581	481	587

3.2 Precipitation Data from GCM Data

The global data used in this study comes from the output of the Global Climate Model (GCM) based on the CMIP6 (Coupled Model Intercomparison Project Phase 6) projections. Several institutions provide GCM data, including MPI-M (Max Planck Institute for Meteorology), which provides global data projections for each sub-experiment conducted over a decadal range. In this study, GCM data with a nominal spatial resolution of 100 km were used, selected because the observation location is in a fairly specific local area. The GCM data was downloaded in NetCDF (Network Common Data Form) format, which contains several data dimensions, such as longitude, latitude, time, and pressure level.

GCM data with a resolution of 100 km has a grid size of 0.937° × 0.934°. On a global scale, this grid size is quite fine and capable of providing detailed data representation. However, problems arise when this global data is cropped for use in smaller or local areas. At this coarse resolution, the study area is represented by only three grid cells (Figure 3a), where large portions are assigned identical precipitation values, making it inadequate for capturing local climate variations. In such cases, spatial interpolation, specifically bilinear interpolation, is required to generate a more varied and appropriate grid distribution at the local scale. This interpolation transforms the original three-grid configuration into a finer 0.1° × 0.1° resolution grid, dramatically increasing the number of data points available for analysis. For example, GCM data for precipitation variables for the Siak Regency, as of January 1, 2013, are shown in Fig. 4.



(a) (b)

Figure 4. GCM Data (a) Before Interpolated and (b) After Interpolation
Source: Processed using RStudio 4.3.2

Fig. 4 shows that without interpolation, GCM data tends to be less diverse, with only three data grids available for use. This is because areas within a single grid are considered to have the same precipitation value. To overcome this limitation, spatial interpolation is performed to increase data diversity at the local scale. By applying a grid resolution of 0.1° × 0.1° to the Siak Regency area, a grid domain of 9 × 22 is obtained. After the global data is scaled down to the local level according to the research location, the next step is to determine the main grid domain to be used in the analysis. The main domain is set at 8 × 8 grids, with the observation weather station located at the center of the grid area used. The selected grid will function as a predictor variable; therefore, the 8 × 8 grid configuration yields a total of 64 predictor variables. The visualization of the grid distribution in the Siak Regency area is presented in Fig. 5.

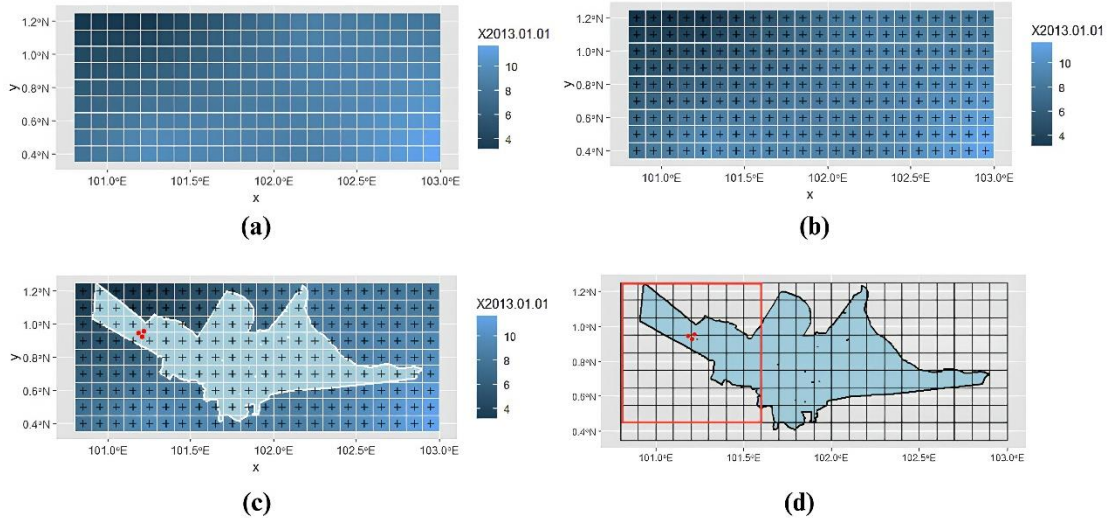


Figure 5. (a) GCM Grid Data, (b) Point Data at Each Grid, (c) Addition of Siak Map and Weather Station Locations, (d) Main Grid to be Used
Source: Processed using RStudio 4.3.2

Fig. 5 presents a visualization of the stages of GCM output data usage. In part (a), a global grid of GCM data covering the research location is displayed. Each grid has different precipitation values, depending on the time of observation. The data on each grid is then represented as data points, as shown in part (b). Next, in part (c), the administrative map of Siak Regency is added along with the location of the weather observation station. From the entire grid available, a main domain measuring 8 × 8 grids is selected, as shown in part (d). This grid will be used as the basis for analysis in the next stage.

3.3 Detection of Multicollinearity

Multicollinearity assessment using Variance Inflation Factor (VIF) analysis as shown in Table 2 revealed that all 64 GCM predictor variables exhibited excessively high VIF values, ranging from 4.17×10^{13} to 4.17×10^{13}, substantially exceeding the standard threshold of VIF > 10. This severe multicollinearity, typical of gridded climate model data due to high spatial-temporal correlation among neighboring grid points, would produce inflated coefficients and unstable estimates in conventional OLS regression. Principal Component Regression (PCR) was therefore employed to eliminate multicollinearity by

transforming correlated predictors into uncorrelated principal components while preserving predictor information.

Table 2. Variance Inflation Factor (VIF) Values

No	Variable	VIF	No	Variable	VIF
1	V1	1.801×10^{15}	33	V33	4.5×10^{14}
2	V2	4.504×10^{15}	34	V34	7.51×10^{14}
3	V3	2.252×10^{15}	35	V35	5.63×10^{14}
4	V4	9.007×10^{14}	36	V36	3×10^{14}
5	V5	6.005×10^{14}	37	V37	1.67×10^{14}
6	V6	9.007×10^{15}	38	V38	3.46×10^{14}
7	V7	5.004×10^{14}	39	V39	9.01×10^{14}
8	V8	1.365×10^{14}	40	V40	7.51×10^{14}
9	V9	3.002×10^{15}	41	V41	2.81×10^{14}
10	V10	4.504×10^{15}	42	V42	5.3×10^{14}
11	V11	3.002×10^{15}	43	V43	3.6×10^{14}
12	V12	1.501×10^{15}	44	V44	1.67×10^{14}
13	V13	1.287×10^{15}	45	V45	9.19×10^{13}
14	V14	9.007×10^{15}	46	V46	2.2×10^{14}
15	V15	9.007×10^{14}	47	V47	4.29×10^{14}
16	V16	3.217×10^{14}	48	V48	2.31×10^{14}
17	V17	1.801×10^{15}	49	V49	1.84×10^{14}
18	V18	3.002×10^{15}	50	V50	3.34×10^{14}
19	V19	1.801×10^{15}	51	V51	2×10^{14}
20	V20	1.287×10^{15}	52	V52	1.11×10^{14}
21	V21	7.506×10^{14}	53	V53	5.93×10^{13}
22	V22	2.252×10^{15}	54	V54	1.5×10^{14}
23	V23	2.252×10^{15}	55	V55	2.57×10^{14}
24	V24	1.001×10^{15}	56	V56	1.13×10^{14}
25	V25	6.929×10^{14}	57	V57	1.27×10^{14}
26	V26	1.801×10^{15}	58	V58	2.31×10^{14}
27	V27	1.001×10^{15}	59	V59	1.58×10^{14}
28	V28	5.629×10^{14}	60	V60	7.32×10^{13}
29	V29	3.106×10^{14}	61	V61	4.17×10^{13}
30	V30	6.929×10^{14}	62	V62	1.02×10^{14}
31	V31	2.252×10^{15}	63	V63	1.61×10^{14}
32	V32	1.501×10^{15}	64	V64	6.53×10^{13}

3.4 Data Model Division and Validation

The data was partitioned for the purpose of evaluating the performance of the regression model in predicting rainfall. This evaluation was conducted by comparing the predicted results with actual observations. Data quality inspection confirmed that both the observed rainfall data from the three LIBO Estate weather stations and the GCM precipitation data were complete, with no missing values, for the study period (2013-2022), allowing for the direct application of the statistical downscaling procedure without requiring imputation methods. The dataset was segmented into two subsets: a training set comprising monthly data from 2013 to 2021 (108 observations) and a validation set comprising monthly data from 2022 (12 observations). The objective of this division was to evaluate the model's capacity for out-of-sample prediction, thereby ensuring its ability to generalize effectively to data not utilized during the training phase.

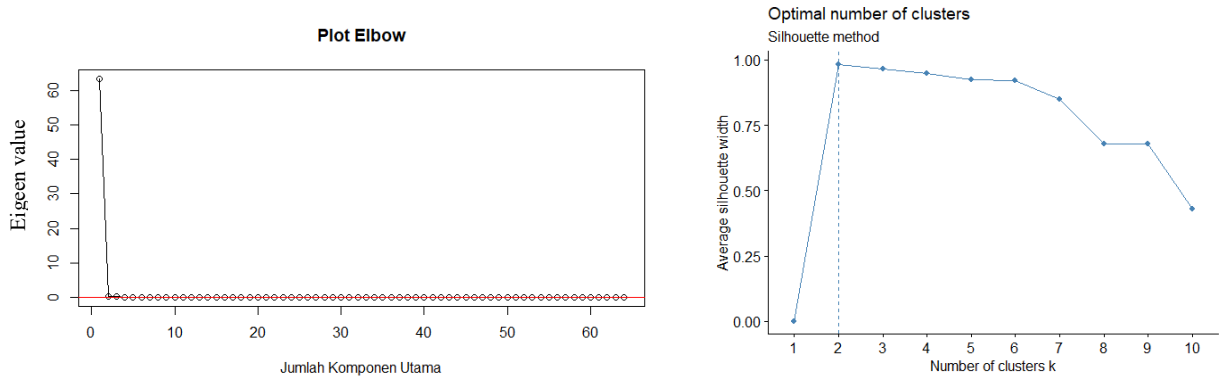
3.5 Statistical Downscaling Modeling with PCR

Prior to applying the model, the predictor variables underwent standardization. The objective of this standardization is to ensure that all variables are measured on a uniform scale, thereby facilitating more consistent analysis and more accurate interpretation. Following standardization, Principal Component Analysis (PCA) is used to derive the principal components that will be incorporated into the regression model. These principal components are subsequently employed as input variables in regression modeling to perform statistical downscaling of rainfall. The PCA commences with the calculation of the eigenvalues based on the correlation matrix between the predictor variables. The value of the eigenvector is employed in the estimation of the number of principal components utilized in the modeling process.

Table 3. Eigenvalue Analysis

	w_1	w_2	w_3	w_4	w_5	w_6
Eigenvalue	63.4116	0.3187	0.213	0.0491	0.0054	0.00175
Cumulative proportion	0.9908	0.99579	0.99912	0.99989	0.99997	1.00000

As illustrated in Table 3, the eigenvalues undergo a precipitous decline following the first and second principal components. The first principal component (w1) accounts for 99.08% of the total variance, while the combination of the first and second components (w1 and w2) explains 99.57% of the total variance in the predictor data. The selection of these two principal components is further substantiated by visual analysis employing the elbow plot and silhouette plot depicted in Fig. 6. As illustrated in Fig. 6 (a), the elbow plot manifests a discernible "elbow" following the second component, thereby suggesting that two components (k = 2) are optimal. Conversely, Fig. 6 (b) demonstrates that clustering based on two principal components yields the most stable and well-separated results, as evidenced by the Silhouette Plot.



(a) **(b)**

Figure 6. (a) Elbow Plot and (b) Silhouette Plot
Source: Processed using RStudio 4.3.2

The results of the selected principal components will subsequently be utilized to calculate the principal component scores w_j ,thereby forming the PCR equation. The component scores for each selected principal component are enumerated in Table 4. These values indicate the contribution of each variable to the formation of the principal components.

Table 4. Principal Component Score Values

Variable	w_1	w_2
Z_1	0.1245	-0.1460
Z_2	0.1247	-0.1640
Z_3	0.1248	-0.1806
Z_4	0.1247	-0.1959
Z_5	0.1246	-0.2100
Z_6	0.1247	-0.2039
Z_7	0.1246	-0.1973
Z_8	0.1243	-0.1902
Z_9	0.1248	-0.1013
⋮	⋮	⋮
Z_{64}	0.1242	0.1707

The principal component formation equation can then be written as follows:

$$w_1 = 0.1245\, Z_1 + 0.1247\, Z_2 + 0.1248\, Z_3 + \cdots + 0.1242\, Z_{64},$$

$$w_2 = -0.1460\, Z_1 - 0.1640\, Z_2 - 0.1806 Z_3 + \cdots + 0.1707\, Z_{64}.$$

These principal components will be continued as the PCR equation.

The principal component values w_1 and w_2 are subsequently employed as predictor variables in regression analysis with the response variable Y, which represents rainfall at each weather station in Libo Estate. The initial PCR model, excluding dummy variables, is delineated in Table 5.

Table 5. PCR Model

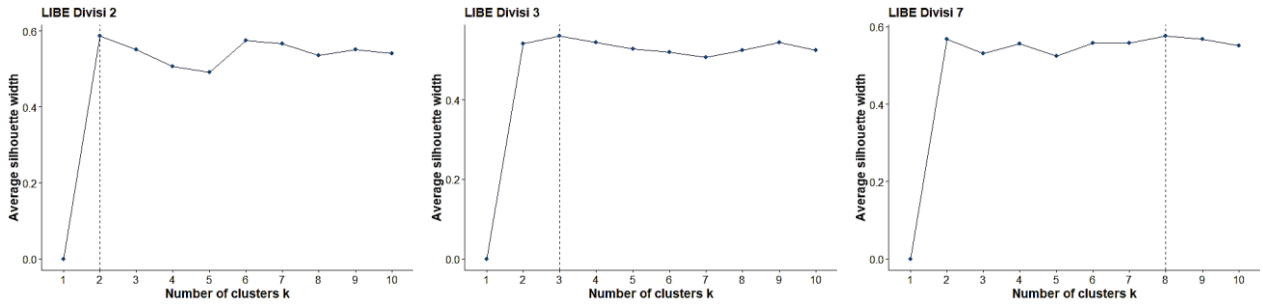
Location	Model
LIBO Estate 2	$Y = 168.102 + 3.862\ w_1 + 27.224\ w_2$
LIBO Estate 3	$Y = 154.329 + 3.020\ w_1 + 25.812\ w_2$
LIBO Estate 7	$Y = 170.583 + 3.569\ w_1 + 4.210\ w_2$

PCR modeling at each weather station uses two main components. The results obtained show that PCR modeling is still inadequate in describing rainfall variability based on the coefficient of determination (R^2) values ranging from 6.97 to 10.62% and RMSE values ranging from 83.136 to 115.427, as shown in Table 5. To improve the accuracy of the model, dummy variables were applied to the PCR modeling at each weather station.

3.6 Statistical Downscaling Modeling with PCR and Dummy Variables

To enhance the accuracy of the principal component regression (PCR) model, dummy variables based on rainfall groupings were introduced. The generation of these dummy variables was achieved through the implementation of the K-means clustering technique, which was applied to the monthly rainfall data from each weather station. The determination of the optimal number of clusters was achieved through the implementation of silhouette analysis, as depicted in Fig. 7 for the LIBE station across Divisions 2, 3, and 7.

The generation of these dummy variables was achieved through the implementation of the K-means clustering technique, which was applied to the monthly rainfall data from each weather station. The determination of the optimal number of clusters was achieved through the implementation of silhouette analysis, as depicted in Fig. 7 for the LIBE station across Divisions 2, 3, and 7.



(a) **(b)** **(c)**

Figure 7. Silhouette plots for LIBE (a) Division 2 (b) Division 3 (c) Division 7

Source: Processed using RStudio 4.3.2

The optimal clustering results vary between stations. LIBO Estate 2 is comprised of two clusters, Division 3 consists of three clusters, and Division 7 is divided into eight clusters (see Fig. 6 and Table 6). Each cluster reflects the range of monthly rainfall (in millimeters per month) based on observation data from the 2013–2022 period. This discrepancy in the number of clusters has a direct impact on the number of dummy variables generated for each station. The variation in cluster numbers across stations indicates differences in rainfall variability: LIBO Estate 2 exhibits homogeneous patterns (low and high rainfall), whereas LIBO Estate 7 demonstrates greater complexity with multiple intermediate rainfall regimes.

Table 6. Summary of K-Means Clustering Results

Group	LIBO Estate 2		LIBO Estate 3		LIBO Estate 7	
	Range	Total	Range	Total	Range	Total
1	0 - 200	79	0 - 120	48	0 - 38	9
2	205 - 581	41	123 - 241	55	47 - 86	15
3	-	-	254 - 481	17	90 - 128	21
4	-	-	-	-	134 - 172	26
5	-	-	-	-	181 - 222	14
6	-	-	-	-	224 - 276	22
7	-	-	-	-	299 - 365	10
8	-	-	-	-	564 - 587	3
Total observation		**120**		**120**		**120**

For instance, in LIBO Estate 2, two distinct rainfall groups are identified: the first cluster (0–200 mm, 79 observations) represents low rainfall and the second cluster (205–581 mm, 41 observations) representing high rainfall. A single dummy variable D_1 is used, coded as 0 for the first cluster and 1 for the second cluster, enabling the model to distinguish between two distinct rainfall regimes. or LIBO Estate 3, which comprises three clusters, two dummy variables (D_1 and D_2) are employed. The coding scheme is: cluster 1 (0–120 mm) = D_1:0, D_2:0; cluster 2 (123–241 mm) = D_1:1, D_2:0; cluster 3 (254–481 mm) = D_1:0, D_2:1. This uniquely represents each rainfall category. The same approach is applied to LIBO Estate 7, which comprises eight clusters requiring seven dummy variables (D_1–D_7). The eight rainfall categories range from 0–38 mm (very low) to 564–587 mm (extreme, only 3 observations), indicating highly variable rainfall patterns that require fine-grained categorization.

Table 7. Group Results With K-Means Clustering and Dummy Variables

Time	LIBE Divisi 2	Group	D1
Jan-13	283	2	1
Feb-13	220	2	1
Mar-13	60	1	0
Apr-13	97	1	0
May-13	132	1	0
⋮	⋮	⋮	⋮
Sept-22	1	0	1
Oct-22	331	2	1
Nov-22	301	2	1
Dec-22	330	2	1

An exemplar of dummy implementation for LIBO Estate 2 is presented in Table 7, wherein each rainfall value is linked to the cluster group and the corresponding D1 value. These dummy variables are subsequently employed as supplementary predictors in a polymerase chain reaction (PCR) analysis for each station within the Libo Estate region. The ensuing discussion will elaborate on the modeling results that incorporate dummy variables, as delineated in Table 8.

Table 8. PCR Model with Dummy Variables

Location	Model
LIBE Estate 2	$Y = 113.146 + 1.3825\ w_1 - 8.8584\ w_2 + 174.567\ D_1$
LIBE Estate 3	$Y = 320.032 + 0.774\ w_1 - 6.713\ w_2 - 250.124\ D_1 - 145.727\ D_2$
LIBE Estate 7	$Y = 573.109 + 0.187\ w_1 + 0.437\ w_2 - 557.272\ D_1 - 504.614\ D_2 - 462.129\ D_3 - 418.280\ D_4 - 370.131\ D_5 - 327.464\ D_6 - 253.476\ D_7$

The incorporation of dummy variables has been demonstrated to markedly enhance model performance, as evidenced by an augmentation in the R^2 value and a diminution in the RMSE value, as illustrated in Table 8. As indicated by the findings presented in Table 9, the incorporation of dummy variables within the PCR model results in a substantial enhancement of model performance. The coefficient of determination (R^2) exhibited an increase from 6.97–10.62% to 63.00–98.50%. The RMSE declined from 83.136–115.427 mm to 17.638–45.177 mm, representing a genuine improvement, not merely a numerical change. For LIBO Estate 7, the RMSE decreased by 78.7% (from 48.4% to 10.3% of the average monthly rainfall), demonstrating that prediction errors now fall within acceptable ranges for agricultural applications. The concurrent improvement in correlation (0.235 to 0.924) validates that the model now captures rainfall dynamics accurately.

Table 9. Coefficient of Determination and RMSE Values for The PCR And PCR with Dummy

Model	Station	Predictor Variables	R^2	RMSE	Correlation
PCR	LIBE Estate 2		10.62%	115.427	0.235
	LIBE Estate 3	w_1, w_2	8.90%	92.767	0.186
	LIBE Estate 7		6.97%	83.136	0.343
PCR - *dummy*	LIBE Estate 2	w_1, w_2, D_1	63.0%	45.177	0.924
	LIBE Estate 3	w_1, w_2, D_1, D_2	82.04%	33.320	0.944
	LIBE Estate 7	$w_1, w_2, D_1, D_2, D_3, D_4, D_5, D_6, D_7$	98.50%	17.683	0.984

Table 9 also shows that the correlation between the predicted results and the observed rainfall data for 2022 for the initial PCR model is relatively low (<0.4), indicating weaknesses in the model that does not

include the dummy variable. After adding the dummy variable, the correlation value increased significantly to 0.984. This indicates a very strong relationship between predictions and observations. Additionally, the PCR–Dummy model for LIBO Estate 7 performs the best, with an R^2 value of 98.50% and an RMSE of 17.683, indicating that this model can explain nearly all variation in the rainfall data with very low prediction error.

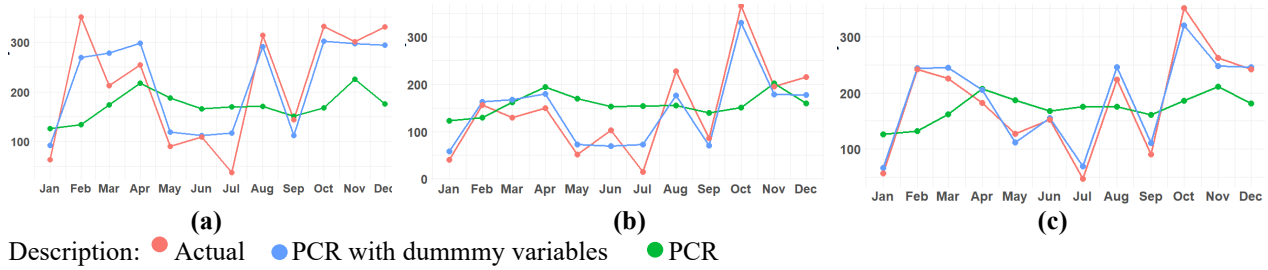


Figure 8. Plot of Observed Rainfall and Estimated Rainfall Using PCR and Dummy Variables for 2022: (a) LIBO Estate 2; (b) LIBO Estate 3; (c) LIBO Estate 7
(*Source: Processed using RStudio 4.3.2*)

As illustrated in Fig. 8, the incorporation of dummy variables into the PCR model results in estimation outcomes that closely mirror the observed rainfall patterns across all weather stations. The model demonstrates a high degree of proficiency in capturing the monthly variations in rainfall, suggesting that it exhibits a notable predictive capacity. Furthermore, the PCR model with dummy variables demonstrates the highest degree of alignment with the observed data, suggesting that the incorporation of dummy variables significantly enhances the model's capacity to accurately represent rainfall patterns in Siak Regency, particularly at each weather station within the LIBO Estate.

4. CONCLUSION

The results of the Principal Component Analysis (PCA) indicate that two principal components account for 99.57% of the variability in precipitation data derived from GCM outputs. The use of the Principal Component Regression (PCR) method, without the inclusion of dummy variables, yields relatively low coefficients of determination (R^2), with a range of 6.97% to 10.62%. However, when PCR is combined with dummy variables—constructed through rainfall data grouping using the K-means clustering method—there is a substantial improvement in model performance. The R^2 values increased significantly, ranging from 63.0% to 98.50%. Furthermore, the Root Mean Square Error (RMSE) exhibits a substantial decrease, transitioning from an initial range of 83.136–115.427 to a significantly lower range of 17.638–45.177. The findings indicate that integrating PCR with dummy variables significantly enhances the model's accuracy and its capacity to capture rainfall variability, particularly in the context of rainfall data modeling in Siak Regency, with a particular emphasis on the Libo Estate region. The findings indicate that integrating PCR with dummy variables significantly enhances the model's accuracy for rainfall modeling in Siak Regency. However, the model's applicability is limited to the research location, as cluster structures, rainfall ranges, and principal components are site-specific and derived from local climate patterns. The PCR-Dummy methodology is transferable to other regions, but parameter values cannot be directly applied to areas with different climate characteristics. Future research should validate this approach in diverse climate zones and investigate its applicability across multiple oil palm plantation regions to establish broader frameworks for water resource management.

Author Contributions

Arisman Adnan: Conceptualization, Supervision, Writing-Original Draft, Writing-Review and Editing. Elsa Riesta Alika: Formal Analysis, Software, Visualization, Writing-Original Draft. Divo Dharma Silalahi: Resources, Data Curation, Investigation, Validation. Felia Rizki Aulia: Methodology, Resources, Validation. Gustriza Erda: Project Administration, Visualization, Writing-Review, and Editing. All authors discussed the results, contributed to the interpretation of the data, and approved the final of the manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

We would like to express our sincere gratitude to the Head of the Environmental Statistical Laboratory, Universitas Riau, for providing valuable guidance, support, and access to facilities that significantly contributed to the completion of this research. We are also deeply thankful to the team at PT. SMART Tbk, Div SMART Research Institute, for their collaboration, provision of data, and insightful discussions that enriched the analysis and practical relevance of this study. Their support has been instrumental in the successful execution of this research.

Declarations

The authors declare no competing interests.

Declaration of Generative AI and AI-Assisted Technologies

The authors used generative AI (ChatGPT) only to assist with language polishing and formatting consistency (e.g., improving wording and ensuring uniform terminology). No AI was used to generate research content, perform analysis, or create.

REFERENCES

[1] N. Dharmawati, S. Suntoro, K. Komariah, and H. Hermantoro, "POTENTIAL RAINWATER AVAILABILITY AND CROP WATER REQUIREMENT OF OIL PALM CROPS DUE TO CLIMATE CHANGE," *IOP Conf Ser Earth Environ Sci*, vol. 1482, p., 2025, doi: https://doi.org/10.1088/1755-1315/1482/1/012006.

[2] J. P. Rajakal, V. Andiappan, and Y. K. Wan, "MATHEMATICAL APPROACH TO FORECAST OIL PALM PLANTATION YIELD UNDER CLIMATE CHANGE UNCERTAINTIES," *Chem Eng Trans*, vol. 83, pp. 115–120, 2021, doi: https://doi.org/10.3303/CET2183020.

[3] A. Abubakar, M. Ishak, and A. Makmom, "NEXUS BETWEEN CLIMATE CHANGE AND OIL PALM PRODUCTION IN MALAYSIA: A REVIEW," *Environ Monit Assess*, vol. 194, p., 2022, doi: https://doi.org/10.1007/s10661-022-09915-8.

[4] I. Pradiko, H. Hariyadi, and T. June, "QUANTIFICATION OF CLIMATE FACTORS CONTRIBUTING TO VARIATION OF OIL PALM YIELD," *Jurnal Penelitian Kelapa Sawit*, p., 2023, doi: https://doi.org/10.22302/iopri.jur.jpks.v31i2.222.

[5] Y. Go, Y.-L. Tan, and T.-H. Yiew, "SENSITIVITY OF OIL PALM YIELD IN INDONESIA TO CLIMATE CHANGE: EVIDENCE FROM THRESHOLD COINTEGRATION MODELS," *Environ Dev Sustain*, p., 2024, doi: https://doi.org/10.1007/s10668-024-05635-w.

[6] S. Oktarina, R. Nurkhoiry, and I. Pradiko, "THE EFFECT OF CLIMATE CHANGE TO PALM OIL PRICE DYNAMICS: A SUPPLY AND DEMAND MODEL," *IOP Conf Ser Earth Environ Sci*, vol. 782, p., 2021, doi: https://doi.org/10.1088/1755-1315/782/3/032062.

[7] S. Sahriman, A. Djuraidah, and A. H. Wigena, "APPLICATION OF PRINCIPAL COMPONENT REGRESSION WITH DUMMY VARIABLE IN STATISTICAL DOWNSCALING TO FORECAST RAINFALL," *Open J Stat*, vol. 04, pp. 678–686, 2014, doi: http://dx.doi.org/10.4236/ojs.2014.49063.

[8] L. Safitri, H. Hermantoro, S. Purboseno, V. Kautsar, S. K. Saptomo, and A. Kurniawan, "WATER FOOTPRINT AND CROP WATER USAGE OF OIL PALM (ELEASIS GUINEENSIS) IN CENTRAL KALIMANTAN: ENVIRONMENTAL SUSTAINABILITY INDICATORS FOR DIFFERENT CROP AGE AND SOIL CONDITIONS," *Water (Switzerland)*, vol. 11, no. 35, pp. 1–16, 2018, doi: https://doi.org/10.3390/w11010035.

[9] N. S. Samsuddin, N. F. A. Aziz, B. Balachandran, and J. Ali, "ECONOMIC CLIMATE MODEL ON THE PALM PRODUCTION: EMPIRICAL EVIDENCE FOR MALAYSIA AND INDONESIA," *Malaysian Journal of Consumer and Family Economics*, p., 2024, doi: https://doi.org/10.60016/majcafe.v33.17.

[10] A. Thant and W. Aye, "FUTURE PREDICTIONS OF RAINFALL USING GCMS: A CASE STUDY FOR MANDALAY, MYANMAR," *International Journal of Scientific and Research Publications (IJSRP)*, p., 2019, doi: https://doi.org/10.29322/ijsrp.9.09.2019.p9314.

[11] B. Deepthi and B. Sivakumar, "SHORTEST PATH LENGTH FOR EVALUATING GENERAL CIRCULATION MODELS FOR RAINFALL SIMULATION," *Clim Dyn*, vol. 61, pp. 3009–3028, 2023, doi: https://doi.org/10.1007/s00382-023-06713-x.

[12] E. Rocheta, M. Sugiyanto, F. Johnson, J. Evans, and A. Sharma, "HOW WELL DO GENERAL CIRCULATION MODELS REPRESENT LOW-FREQUENCY RAINFALL VARIABILITY?," *Water Resour Res*, vol. 50, pp. 2108–2123, 2014, doi: https://doi.org/10.1002/2012WR013085.

[13] X. Su, W. Shao, J. Liu, and Y. Jiang, "MULTI-SITE STATISTICAL DOWNSCALING METHOD USING GCM-BASED MONTHLY DATA FOR DAILY PRECIPITATION GENERATION," *Water (Basel)*, vol. 12, pp. 1–21, 2020, doi: https://doi.org/10.3390/w12030904.

[14] M. D. Saputra, A. F. Hadi, A. Riski, and D. Anggraeni, "PRINCIPAL COMPONENT REGRESSION IN STATISTICAL DOWNSCALING WITH MISSING VALUE FOR DAILY RAINFALL FORECASTING," *International Journal of Quantitative Research and Modeling*, vol. 2, no. 3, pp. 139–146, 2021, doi: https://doi.org/10.46336/ijqrm.v2i3.151.

[15] A. Mulyati, A. Wigena, and A. Djuraidah, "STATISTICAL DOWNSCALING USING KERNEL QUANTILE REGRESSION TO PREDICT EXTREME RAINFALL," *Int J Sci Basic Appl Res*, vol. 42, pp. 1–9, 2018, [Online]. Available: https://consensus.app/papers/statistical-downscaling-using-kernel-quantile-djuraidah-mulyati/ed63469d34a65eb28c90670b1cbfaf24/

[16] S. Sahriman and A. S. Yulianti, "STATISTICAL DOWNSCALING MODEL WITH PRINCIPAL COMPONENT REGRESSION AND LATENT ROOT REGRESSION TO FORECAST RAINFALL IN PANGKEP REGENCY," *Barekeng: Journal of Mathematics and Its Applications*, vol. 17, no. 1, pp. 0401–0410, 2023, doi: https://doi.org/10.30598/barekengvol17iss1pp0401-0410.

[17] S. Sahriman and A. Anisa, "FORECASTING MONTHLY RAINFALL IN PANGKEP REGENCY USING STATISTICAL DOWNSCALING MODEL WITH ROBUST PRINCIPAL COMPONENT REGRESSION TECHNIQUE," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, p., 2025, doi: https://doi.org/10.30598/barekengvol19iss2pp777-790.

[18] S. Sahriman, E. L. Randa, S. A. Surianda, M. Z. G. Hisyam, Muh. I. Taufik, and G. D. Putra, "RAINFALL FORECASTING OF SALT PRODUCING AREAS IN PANGKEP REGENCY USING STATISTICAL DOWNSCALING MODEL WITH LINEARIZED RIDGE REGRESSION DUMMY," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 1, pp. 0483–0492, Mar. 2024, doi: https://doi.org/10.30598/barekengvol18iss1pp0483-0492.

[19] P. Loganathan and A. B. Mahindrakar, "STATISTICAL DOWNSCALING USING PRINCIPAL COMPONENT REGRESSION FOR CLIMATE CHANGE IMPACT ASSESSMENT AT THE CAUVERY RIVER BASIN," *Journal of Water and Climate Change*, vol. 12, no. 6, pp. 2314–2324, 2021, doi: https://doi.org/10.2166/wcc.2021.223.

[20] Lawrence Livermore National Laboratory, "CMIP6." [Online]. Available: https://aims2.llnl.gov/search/cmip6/

[21] A. H. Wigena, "PEMODELAN STATISTICAL DOWNSCALING DENGAN REGRESI PROJECTION PURSUIT UNTUK PERAMALAN CURAH HUJAN BULANAN," Institute Pertanian Bogor, 2006.

[22] W. Surianto, "PEMODELAN STATISTICAL DOWNSCALING MENGGUNAKAN COMBINE CLUSTERWISE REGRESSION UNTUK PENDUGAAN CURAH HUJAN HARIAN," Institute Pertanian Bogor, 2022.

[23] A. Gwelo, "PRINCIPAL COMPONENTS TO OVERCOME MULTICOLLINEARITY PROBLEM," *Oradea Journal of Business and Economics*, p., 2019, doi: https://doi.org/10.47535/1991OJBE062.

[24] M. Raheem, N. Udoh, and A. T. Gbolahan, "CHOOSING APPROPRIATE REGRESSION MODEL IN THE PRESENCE OF MULTICOLINEARITY," *Open J Stat*, p., 2019, doi: https://doi.org/10.4236/OJS.2019.92012.

[25] R. A. Johnson and D. W. Wichern, *APPLIED MULTIVARIATE STATISTICAL ANALYSIS*, 6 Ed. United States of America: Pearson Prentice Hall, 2007.

[26] D. N. Gujarati and D. C. Porter, *BASIC ECONOMETRICS*, 5th ed. United States: The McGraw-Hill Companies, Inc., 2009.

[27] D. A. Setiady and H. Leong, "IMPLEMENTATION OF K-MEANS ALGORITHM ELBOW METHOD AND SILHOUETTE COEFFICIENT FOR RAINFALL CLASSIFICATION," *Proxies : Jurnal Informatika*, p., 2024, doi: https://doi.org/10.24167/proxies.v4i1.12433.

[28] G. Erda, C. Gunawan, and Z. Erda, "GROUPING OF POVERTY IN INDONESIA USING K-MEANS WITH SILHOUETTE COEFFICIENT," *Parameter: Journal of Statistics*, vol. 3, no. 1, pp. 1–6, 2023, doi: https://doi.org/10.29040/ijcis.v6i1.218.

[29] M. Riasetiawan, A. Ashari, and P. Wahyu, "THE PERFORMANCE EVALUATION OF K-MEANS AND AGGLOMERATIVE HIERARCHICAL CLUSTERING FOR RAINFALL PATTERNS AND MODELLING," *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 431–436, 2022, doi: https://doi.org/10.1109/ICITISEE57756.2022.10057729.

[30] J. Feng and X. Wang, "IMPACT OF TWO TYPES OF LA NIÑA ON BOREAL AUTUMN RAINFALL AROUND SOUTHEAST ASIA AND AUSTRALIA," *Atmospheric and Oceanic Science Letters*, vol. 11, pp. 1–6, 2018, doi: https://doi.org/10.1080/16742834.2018.1386538.

[31] S. Zhong, Y. Zhang, and L. Jiang, "IMPACT OF DIFFERENT TYPES OF LA NIÑA DEVELOPMENT ON THE PRECIPITATION IN THE MARITIME CONTINENT," *Atmosphere-Ocean*, vol. 62, pp. 254–267, 2024, doi: https://doi.org/10.1080/07055900.2024.2326611.