BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 0777–0790
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp0777-0790

FORECASTING MONTHLY RAINFALL IN PANGKEP REGENCY USING STATISTICAL DOWNSCALING MODEL WITH ROBUST PRINCIPAL COMPONENT REGRESSION TECHNIQUE

Sitti Sahriman[1*], Anisa[2]

[1,2]*Statistics Department, Faculty of Mathematics and Natural Sciences, Universitas Hasanuddin*
Jln. Perintis Kemerdekaan no 10 Tamalanrea Makassar, Sulawesi Selatan, 90245, Indonesia

*Corresponding author e-mail: * **sittisahrimansalam@gmail.com***

Article History:

Received: 30th June 2024
Revised: 30th January 2025
Accepted: 1st March 2025
Published: 1st April 2025

Keywords:

General Circulation Model;
Minimum Covariance Determinant;
Minimum Vector Variance;
Principal Component Regression;
Statistical Downscaling

ABSTRACT

A General Circulation Model (GCM) is a global climate model commonly used to predict local-scale climate patterns. However, the spatial resolution of GCMs is typically on a global scale, which is inadequate for predicting local climate. Statistical downscaling (SD) is used to transform climate information from a global scale to a smaller scale for local-scale climate predictions. GCM data have large dimensions and high correlations between grids, so principal component regression (PCR) is used in SD. The minimum covariance determinant (MCD) and minimum vector variance (MVV) methods are used in principal component analysis to obtain robust principal components (PCs). The data used in this study were the monthly rainfall data in Pangkep Regency for the period from January 1999 to December 2022 as the response variable, which were obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) Region IV Makassar. The predictor variable data were GCM precipitation data (64 variables) for the same period and three dummy variables. This study aimed to obtain rainfall forecasts in Pangkep Regency for the year 2023 based on a robust PCR model using results from MCD and MVV. The modeling results indicated that both the MCD and MVV methods provided similar model accuracy, with a coefficient of determination of approximately 91%. The PCR model with two PCs from the MVV method and dummy variables was identified as the best model for explaining the variability in rainfall data in Pangkep Regency. Additionally, the 2023 rainfall forecast results showed that both methods yielded relatively similar accuracy. The addition of dummy variables in the PCR model improved both the model accuracy and rainfall forecasts. The PCR model with three PCs from MVV and dummy principal component variables produced accurate rainfall forecasts based on a high correlation value (0.974) and the smallest mean absolute percentage error (7.290).

How to cite this article:

S. Sahriman and Anisa., "FORECASTING MONTHLY RAINFALL IN PANGKEP REGENCY USING STATISTICAL DOWNSCALING MODEL WITH ROBUST PRINCIPAL COMPONENT REGRESSION TECHNIQUE," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 0777-0790, June, 2025.

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

Astronomically, Indonesia is located between 6° N - 11° S latitude and 95° E - 141° E longitude, crossing the equator. This places Indonesia within the tropical climate zone with significant rainfall variability. Changes in rainfall in Indonesia can have significant impacts on various sectors, including salt production. Sulawesi Selatan is one of the national centers for salt production in Indonesia, with one of its regions being Pangkep Regency. In 2019, salt production in South Sulawesi reached its highest level in the last five years, amounting to 140,338 tons. However, salt production experienced a decline in 2020, reaching only 45.31 tons. In 2021, salt production further decreased drastically to only 1.83 tons. Subsequently, production increased again in 2022 to 3,282 tons. There are several factors influencing salt production processes in South Sulawesi, one of which is weather conditions, particularly rainfall.

The research on rainfall is conducted to reduce risks and enhance resilience against weather variations. Therefore, high-resolution climate models continue to be developed, considering global-scale climate circulation, including Global Circulation Models (GCMs). GCM climate models use mathematical models of planetary atmospheric or oceanic circulation based on physical processes to simulate the transfer of energy and matter through the climate system. However, the horizontal resolution of GCMs ranges from 250 to 600 km, thus resulting in low accuracy for predicting local-scale climate **[1]**. To address this issue, the statistical downscaling model was commonly applied because it was cheaper and more efficient in linking global-scale climate variables to local scales **[2]**.

Statistical Downscaling (SD) is a method for converting outputs from global-scale climate models into locally scaled information. SD is a technique that uses local-scale data such as rainfall data from BMKG stations as response variables, and global-scale data such as GCM outputs as predictor variables **[3]**. This model could provide deeper insights into the impacts of climate change at the local level, given the complexity of predictions, especially regarding rainfall and regional topography **[4]**. However, GCM data provided climate data in spatial form available in grid format covering the entire regional domain. This led to multicollinearity issues in SD models, necessitating a statistical technique to handle multicollinearity effects in GCM data **[5]**. Methods that can overcome multicollinearity include principal component regression.

Principal Component Regression (PCR) was a statistical approach that combined linear regression with principal component analysis **[6]**. Principal Component Analysis (PCA) was a statistical method that reduced research variables into smaller dimensions without losing information from the variables. This method was commonly used to reduce the dimensions of GCM output data and address multicollinearity issues **[7]**. The reduced variables, known as principal components (PCs), were linear combinations of the original variables **[8]**. However, the presence of outliers can significantly affect the results of PCA, leading to misleading interpretations **[9]**. Therefore, robust methods could be used to minimize the impact of outliers by replacing classical estimators with robust estimators **[10]**.

There are several robust methods for the sample covariance matrix in PCA to handle outliers, namely the Minimum Covariance Determinant (MCD) and the Minimum Vector Variance (MVV) methods. The MCD method works by identifying a subset of observations whose covariance matrix has the smallest determinant among all possible data combinations. This process aimed to produce a covariance matrix robust to outliers **[11]**. Furthermore, the Fast Minimum Covariance Determinant (FMCD) method was an algorithm more efficient in generating robust covariance matrices **[12]**. Meanwhile, the MVV method involved selecting a small subset of vectors from the main sample with the smallest variance. These vectors were then used to form a robust outlier-resistant matrix **[13]**.

Rainfall modeling had been widely conducted through the application of Statistical Downscaling techniques, such as by **[14]**, who used the least absolute shrinkage and selection operator (LASSO) method and PCR. **[15]** augmented dummy variables based on hierarchical and non-hierarchical clustering techniques in SD modeling for rainfall estimation. **[16]** utilized the PCR method in SD models with missing values to forecast daily rainfall data. **[17]** used projection pursuit regression in SD modeling for daily rainfall forecasting in Kupang City, East Nusa Tenggara. In addition, monthly rainfall estimation from the Bandung, Bogor, Citeko, and Jatiwangi stations was conducted using cluster-wise regression in SD model **[18]**. Furthermore, **[19]** compared classical PCR results with robust PCR from MVV for rainfall forecasting in Pangkep Regency. Additionally, **[20]** compared PCR and Latent Root Regression methods in SD models.

Rainfall forecasting using the SD model was generally focused on methods to address multicollinearity in GCM data. On the other hand, in addition to multicollinearity, the presence of outliers could also affect the

accuracy of rainfall forecasting. Therefore, this study used robust PCR, which could address both multicollinearity and outliers in the rainfall data, thus improving forecasting accuracy. The objective of this study was to estimate rainfall in Pangkep Regency using the SD model and robust PCR method, which was based on the covariance matrix formed from the MCD and MVV methods.

2. RESEARCH METHODS

2.1 Data Sources

The observational data in this study were the monthly rainfall data in Pangkep Regency (Y) for the period 1999-2023 as the response variable. The predictor variables were GCM precipitation data. The GCM rainfall output (X) was climate simulation data for the Pangkep Regency area produced by the Climate Model Intercomparison Project (CMIP5) through KNMI Netherlands (https://climexp.knmi.nl/start.cgi). The GCM domain used consisted of several square grids measuring 8 × 8 grids (2.5° × 2.5° for each grid) from 119.57°E to 129.37°E and -14.83°S to 5.17°N. Three dummy variables (D_1, D_2, D_3) based on the non-hierarchical K-Means clustering technique were used in this study to improve model accuracy [15]. Thus, there were 67 predictor variables and 1 response variable, as presented in Table 1 below:

Table 1. Data Structure

Time	Y	X_1	X_2	...	$X_{i(j)}$	...	X_{64}	D_1	D_2	D_3
Jan-1999	y_1	$x_{1(1)}$	$x_{1(2)}$	...	$X_{1(j)}$	...	$x_{1(64)}$	$D_{1(1)}$	$D_{1(2)}$	$D_{1(3)}$
Feb-1999	y_2	$x_{2(1)}$	$x_{2(2)}$	...	$X_{2(j)}$	...	$x_{2(64)}$	$D_{2(1)}$	$D_{2(2)}$	$D_{2(3)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
Dec-2023	y_{300}	$x_{300(1)}$	$x_{300(2)}$	...	$X_{300(j)}$	...	$x_{300(64)}$	$D_{300(1)}$	$D_{300(2)}$	$D_{300(3)}$

where $X_{i(j)}$ represents the GCM precipitation at time i and grid j. The results of the K-means analysis grouped the rainfall data of Pangkep Regency into four clusters. Group 1 ($D_1 = 1, D_2 = D_3 = 0$) had 157 observations, Group 2 ($D_2 = 1, D_1 = D_3 = 0$) had 100, Group 3 ($D_3 = 1, D_1 = D_2 = 0$) had 4, and Group 4 ($D_1 = D_2 = D_3 = 0$) had 39 observations. The data was split into training data (Jan 1999-Dec 2022) for modeling and testing data (2023) for validation.

2.2 Analysis Methods

This study used GCM output precipitation data as the response variable, where correlations between grids were high. Multicollinearity can make Ordinary Least Squares (OLS) estimators unreliable. Variance Inflation Factor (VIF) is one method used to detect multicollinearity in data. A VIF value above 10 is often considered an indication of significant multicollinearity among predictor variables [21]. A VIF was calculated using the following Equation (1) where R_j^2 is the determination coefficient of predictor variable j regressed against other predictors [22]:

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = 1,2, \dots, p \tag{1}$$

SD models are used to link global-scale climate variables with local-scale variables. SD models can be defined as follows [23]:

$$\boldsymbol{y} = f(\boldsymbol{X}) \tag{2}$$

where $\boldsymbol{y}_{(n\times 1)}$ represents local climate variables, $\boldsymbol{X}_{(n\times p)}$ denotes GCM output variables, n stands for the number of periods (daily or monthly), and p signifies the number of GCM grid domains. In SD techniques, the PCR method is used to address multicollinearity issues. PCR begins with Principal Component Analysis (PCA) to reduce data dimensionality or address multicollinearity by generating new variables called PCs that retain as much variability from the original data as possible [24]. PCs are obtained from the eigenvalue-eigenvector pairs of the covariance or correlation matrix.

If $\boldsymbol{X}' = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p]$ has a covariance matrix $(\boldsymbol{\Sigma})$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and eigenvector $\boldsymbol{e}_j$, the PC variables $(\boldsymbol{w}_j)$ is obtained, which are linear combination of the original variables, as in **Equation (3) [24]**:

$$\boldsymbol{w}_j = \boldsymbol{e}_j{}'\boldsymbol{X} = e_{j1}\boldsymbol{x}_1 + e_{j2}\boldsymbol{x}_2 + \cdots + e_{jp}\boldsymbol{x}_p \tag{3}$$

The $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_p$ variables are linear combinations of the original variables (X) that are uncorrelated and have maximum variance. Each PC variable has a variance equal to the eigenvalue of the matrix $\boldsymbol{\Sigma}$, so for the j-th PC equation, the variance and covariance are as follows **[24]**,

$$\begin{aligned} Var(\boldsymbol{w}_j) &= \boldsymbol{e}_j{}'\boldsymbol{\Sigma}\,\boldsymbol{e}_j = \lambda_j \\ Cov(\boldsymbol{w}_j, \boldsymbol{w}_{j'}) &= \boldsymbol{e}_j{}'\boldsymbol{\Sigma}\,\boldsymbol{e}_{j'} = 0 \quad ,j \neq j' = 1,2,\dots,p \end{aligned} \tag{4}$$

The covariance matrix of PC $(\boldsymbol{W})$ can be written in **Equation (5)** as follows **[24]**:

$$\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_p) \tag{5}$$

Therefore, the total variance of the original variables equals the total variance explained by the PCs and can be expressed as in **Equation (6)**,

$$\sigma_1{}^2 + \sigma_2{}^2 + \cdots + \sigma_p{}^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_p \tag{6}$$

PCs can also be derived using the correlation matrix by first transforming the original variables (X) into standardized form (Z), as shown in **Equation (7)**:

$$\boldsymbol{Z} = \left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{7}$$

where $\boldsymbol{V}^{1/2} = \mathrm{diag}\left(\sqrt{\sigma_1{}^2}, \dots, \sqrt{\sigma_j{}^2}, \dots, \sqrt{\sigma_p{}^2}\right), \sigma_j{}^2$ and $\boldsymbol{\mu}$ are the variance and the vector containing the mean values of the variables in X, respectively. Meanwhile, $\boldsymbol{Z}$ is the standardized matrix of the original variables X, where mean and variance expressed in **Equation (8)**:

$$E(\boldsymbol{Z}) = \boldsymbol{0} \;\; ; \;\; Cov(\boldsymbol{Z}) = \left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{V}^{\frac{1}{2}}\right)^{-1} = \boldsymbol{R} \tag{8}$$

where $\boldsymbol{R}$ is the correlation matrix of the original variables X. The j-th PC, formed based on the standardized variables $\boldsymbol{Z}' = [\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_p]$, can be determined from the eigenvector obtained through the correlation matrix of the original variables X using the PC formula in **Equation (9)**:

$$\boldsymbol{w}_j = \boldsymbol{e}_j{}'\boldsymbol{Z} = e_{j1}\boldsymbol{z}_1 + e_{j2}\boldsymbol{z}_2 + \cdots + e_{jp}\boldsymbol{z}_p \tag{9}$$

The next step is to regress the selected PCs obtained from PCA against the response variable using PCR. Let $\boldsymbol{P}$ be an orthogonal matrix containing the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ of the original variables X, satisfying the equation $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$. The formation process of PCR from multiple linear regression, with $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{P}$ and $\boldsymbol{\alpha} = \boldsymbol{P}'\boldsymbol{\beta}$, is **Equation (10) [25]**:

$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\boldsymbol{X}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\boldsymbol{W}\boldsymbol{\alpha}+\boldsymbol{\varepsilon}$#10

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{10}$$

The PCR model resulting from reducing to r components is written in **Equation (11)**:

$$\boldsymbol{y} = \alpha_0\boldsymbol{1} + \boldsymbol{W}_r\boldsymbol{\alpha}_r + \boldsymbol{\varepsilon} \tag{11}$$

where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$ is a vector of errors of size $n \times 1$, $\boldsymbol{X}$ is the predictor variable matrix of size $n \times (p+1)$, $\boldsymbol{y}$ is the response variable vector of size $n \times 1$, α_0 is the intercept, $\boldsymbol{1}$ is a vector of ones of size $n \times 1$, $\boldsymbol{W}_r$ is an $n \times r$ of PCs, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_r$ are the parameter vectors of $\boldsymbol{X}$ and $\boldsymbol{W}$, respectively.

Using the maximum likelihood method, the parameter estimators are obtained as follows **[26]**:

$$\hat{\boldsymbol{\alpha}}_r = (\boldsymbol{W}_r{}'\boldsymbol{W}_r)^{-1}\boldsymbol{W}_r{}'\boldsymbol{y} \tag{12}$$

where $\boldsymbol{W}_r = \boldsymbol{X}\boldsymbol{P}_r$ if using the covariance matrix of the variables X, and $\boldsymbol{W}_r = \boldsymbol{Z}\boldsymbol{P}_r$ if using the correlation matrix of the variables X. Here, $\boldsymbol{P}_r$ is an $r \times r$ matrix whose elements are eigenvectors.

To produce a robust covariance matrix ($\boldsymbol{\Sigma}$), the MCD method is used. The objective of MCD is to obtain a subsample of size h from the total n observations, which has the covariance matrix with the smallest determinant among all possible data combinations [27]:

$$h = \frac{n+p+1}{2} \tag{13}$$

The MCD method uses Equation (14) as follows,

$$\bar{\boldsymbol{x}}_{MCD} = \frac{1}{h}\sum_{i\epsilon H} x_i \,; \boldsymbol{S}_{MCD} = \frac{1}{h-1}\sum_{i\epsilon H}(x_i - \bar{\boldsymbol{x}}_{MCD})\,(x_i - \bar{\boldsymbol{x}}_{MCD})' \tag{14}$$

where $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are the mean vector and the sample covariance matrix, respectively.

In addition to the MCD method, the MVV method is also used in this study to construct a robust covariance matrix. The MVV method produces a covariance matrix S_{MVV} with the minimum value of the trace, $Tr(S^2_{MVV})$, among all possible subsets containing h data observations. Consequently, the MVV estimate for the location parameter of the matrix is determined as follows [28]:

$$\bar{\boldsymbol{x}}_{MVV} = \frac{1}{h}\sum_{i\epsilon H} x_i \,; \boldsymbol{S}_{MVV} = \frac{1}{h-1}\sum_{i\epsilon H}(x_i - \bar{\boldsymbol{x}}_{MVV})(x_i - \bar{\boldsymbol{x}}_{MVV})' \tag{15}$$

The MVV method algorithm is as follows [29]:

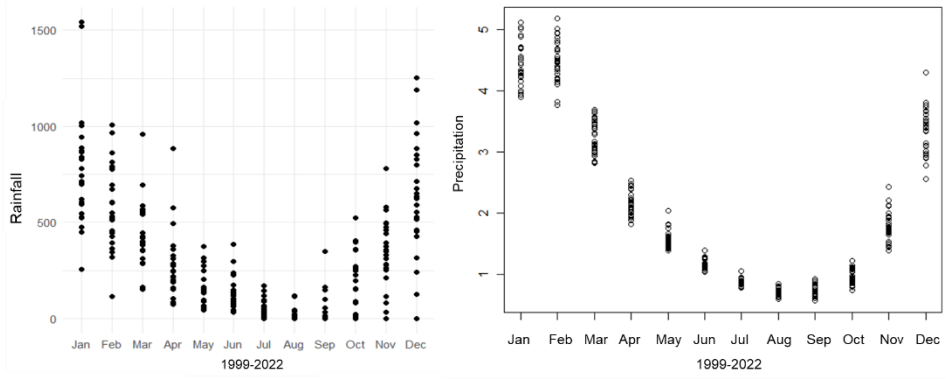
1. Select a data set consisting of $h = \frac{(n+p+1)}{2}$ data observations, referred to as H_{old}. H refers to the set of a selected subset of data with size $h \times p$.
2. Calculate the mean vector $\bar{x}_{Hold}$ and the covariance matrix S_{Hold} for all data in H_{old}. For $i = 1,2,\dots,n$, calculate squared Mahalanobis distance, $d^2_{H_{old}} = d^2_{H_{old}}(x_i, \bar{x}_{Hold}) = (x_i - \bar{x}_{Hold})' S^{-1}_{H_{old}}(x_i - \bar{x}_{Hold})$
3. Sort the results from smallest to largest. This ordering will provide a permutation of observation indices π. For example: $d^2_{H_{old}}(\pi_1) \le d^2_{H_{old}}(\pi_2) \dots \le d^2_{H_{old}}(\pi_n)$
4. Form a new set consisting of h observations with indices $\pi(1), \pi(2), \dots, \pi(h)$ and name it H_{new}
5. Calculate $\bar{\boldsymbol{x}}_{H_{new}}$, $\boldsymbol{S}_{H_{new}}$, and $(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{H_{new}})$ as in step 2
6. If $Tr(\boldsymbol{S}^2_{H_{new}}) = Tr(\boldsymbol{S}^2_{H_{old}})$, the process is complete. If $Tr(\boldsymbol{S}^2_{H_{new}}) < Tr(\boldsymbol{S}^2_{H_{old}})$, continue the process until the k-th iteration reaches $Tr(\boldsymbol{S}^2_{H_{new}}) = Tr(\boldsymbol{S}^2_{H_{old}})$.

Let S_{H_k} be the covariance matrix from the kth iteration. At the end of the k-th iteration, the following inequality will hold: $Tr(\boldsymbol{S}^2_{H_1}) \ge Tr(\boldsymbol{S}^2_{H_2}) \ge \cdots \ge Tr(\boldsymbol{S}^2_{H_{k-1}}) = Tr(\boldsymbol{S}^2_{H_k})$.

3. RESULTS AND DISCUSSION

3.1 Data Exploration

Data exploration was an important step before statistical downscaling modeling as it provided a deep understanding of the characteristics of the data used. Data exploration helped us determine how well the GCM precipitation pattern matched the rainfall pattern in Kabupaten Pangkep, which was key in choosing the best approach for statistical downscaling (SD) modeling.



(a) (b)

Figure 1. Plot of (a) Rainfall in Pangkep Regency and (b) GCM Precipitation

Figure 1 presented the plot of rainfall in Kabupaten Pangkep (y) and GCM precipitation, which in this case referred to the precipitation in the first grid (X_1). The rainfall pattern in Kabupaten Pangkep showed a monsoon pattern, with high rainfall at the beginning and end of the year. The same pattern was also observed in precipitation X_1. This indicated that the rainfall pattern in Kabupaten Pangkep and precipitation X_1 were similar. A similar pattern was also seen in other GCM precipitation, such as X_2 to X_{64}. This finding suggested a similarity in the monsoon pattern between local rainfall data and global-scale GCM precipitation, which is crucial for further analysis in SD modeling.

3.2 Detection of Multicollinearity

Multicollinearity occurs when several predictor variables in a linear regression model are highly correlated with each other. This makes it challenging to accurately determine the individual impact of each predictor variable on the response variable, potentially leading to erroneous conclusions. The Variance Inflation Factor (VIF) is a statistical value used to detect the presence of multicollinearity in a regression model. Based on the calculation results, the VIF values for 64 precipitation variables showed that there were 60 predictor variables (93.75%) with VIF values greater than 10. The VIF values ranged from 4.22 to 3099.68. The high VIF values indicated that these predictor variables were almost a perfect linear combination of other predictor variables. Therefore, principal component analysis (PCA) was used as a pre-processing step in the SD model to address multicollinearity.

3.3 Outlier Detection

Besides multicollinearity, detecting and handling outliers in regression analysis is also crucial because it can significantly impact the interpretation of the model and prediction accuracy. Outliers are data points that significantly differ from the majority of the data in a dataset. Outliers can affect parameter estimation, reduce prediction accuracy, and disrupt the validity of statistical inferences. The presence of outliers can cause residuals of the model to deviate from a normal distribution. Boxplots are one of the effective statistical tools for detecting outliers in a dataset.

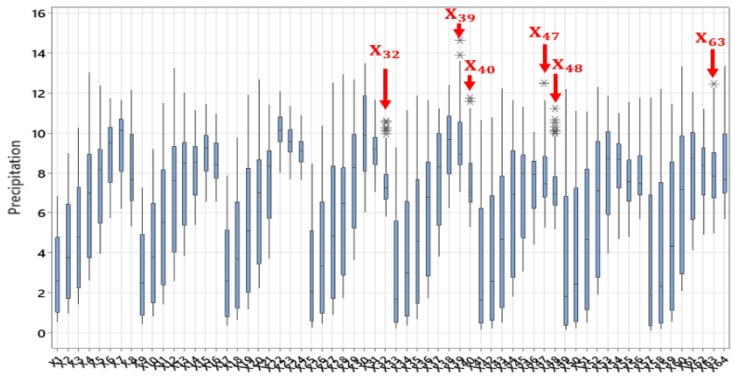


Figure 2. Boxplot of GCM Precipitation, X_1-X_{64}

Figure 2 showed that out of 64 GCM precipitation data points, there were 6 predictor variables containing outliers, namely variables $X_{32}, X_{39}, X_{40}, X_{47}, X_{48}$, and X_{63}. Variable X_{32} had 7 observations

identified as outliers at observations 1, 13, 37, 85, 169, 265, and 277. Variables X_{39} and X_{40} each had 2 outlier observations at observations 169 and 265. Variable X_{47} had 1 observation as an outlier at observation 265, while variable X_{48} had 8 observations identified as outliers at observations 25, 37, 85, 121, 157, 169, 265, and 277. Meanwhile, variable X_{63} had 1 observation as an outlier at observation 120.

Meanwhile, outlier detection using MVV method utilizes the Minimum Volume Ellipsoid (MVE) value. The resulting ellipsoid is the minimum ellipsoid while retaining most of the normal data within the distribution, thus facilitating outlier identification and visualization. The analysis results indicated that there were 8 observations on the predictor variables detected as outliers with a threshold value of 97.5%. These observations included numbers 37, 109, 145, 168, 192, 205, 228, and 240.

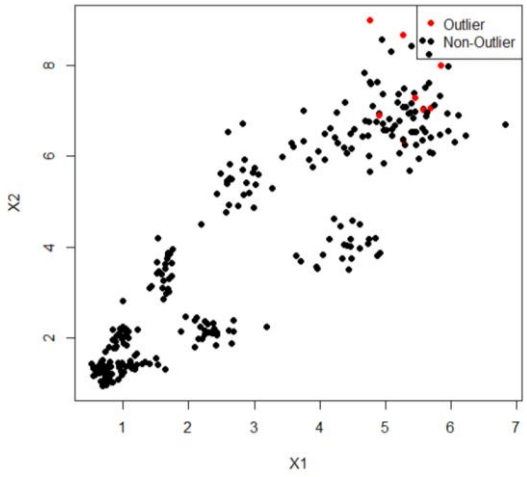


Figure 3. Outlier Detection of GCM using MVE

Figure 3 presented an example plot of GCM precipitation data, namely X_1 and X_2. The points on the graph represented the data, with red dots indicating outliers and black dots indicating non-outliers. In this figure, most of the data clustered in the center, while the outliers were scattered in the upper and right parts of the graph based on MVE. The same approach was applied to other GCM precipitation data. Therefore, MCD and MVV methods were used in PCA to form robust covariance matrices against outliers.

3.4 Principal Component Analysis Using Minimum Covariance Determinant and Minimum Vector Variance

Principal Component Analysis (PCA) is a statistical analysis technique used to reduce the dimensionality of data by extracting new variables, called principal components, that explain most of the variability in the original data. However, typically, the principal components (PCs) are formed based on the covariance matrix of the original data, which is sensitive to outliers. Therefore, the Minimum Covariance Determinant (MCD) and Minimum Vector Variance (MVV) methods are used to form a robust covariance matrix that is resistant to outliers. In this study, the number of observation subsets used in both the MCD and MVV methods was,

$$h = \frac{n+p+1}{2} = \frac{288+64+1}{2} \approx 176$$

The selected observation subsets in both the MCD and MVV methods were then used to calculate the covariance matrix of those subsets. The covariance matrix from the MCD method is presented in **Table 2**,

Table 2. The Covariance Matrix of the MCD Method

Variable	X_1	X_2	X_3	X_4	⋯	X_{63}	X_{64}
X_1	1.918	1.791	2.192	2.919	⋯	0.874	0.502
X_2	1.791	2.648	2.818	3.435	⋯	1.201	0.948
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
X_{63}	0.874	1.201	1.286	1.747	⋯	1.503	0.752
X_{64}	0.502	0.948	0.933	1.029	⋯	0.752	0.986

Meanwhile, the covariance matrix from the MVV method is presented in **Table 3**

Table 3. The Covariance Matrix of the MVV Method

Variable	X_1	X_2	X_3	X_4	⋯	X_{63}	X_{64}
X_1	2.167	2.164	2.522	3.020	⋯	0.989	0.920
X_2	2.164	2.974	3.152	3.566	⋯	1.325	1.383
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
X_{63}	0.989	1.325	1.399	1.697	⋯	1.358	0.854
X_{64}	0.920	1.383	1.400	1.405	⋯	0.854	1.299

Principal components (PCs) were formed based on the correlation matrix. By using the covariance matrix, the correlation matrix obtained is as follows:

$$R_{MCD} = \begin{bmatrix} 1 & 0.795 & \cdots & 0.365 \\ 0.795 & 1 & \cdots & 0.587 \\ \vdots & \vdots & \ddots & \vdots \\ 0.365 & 0.587 & \cdots & 1 \end{bmatrix} \; ; \; R_{MVV} = \begin{bmatrix} 1 & 0.852 & \cdots & 0.548 \\ 0.852 & 1 & \cdots & 0.703 \\ \vdots & \vdots & \ddots & \vdots \\ 0.548 & 0.703 & \cdots & 1 \end{bmatrix}$$

Table 4 presented the eigenvalues and the proportion of variability in the original variables explained by the PCs formed based on the correlation matrix. The table indicated that the first PC (W_1) of the MVV method explained 73% of the data variability, which was better than the result from the MCD method, which was only 67%. Additionally, the second PC (W_2) contributed approximately 10% to 11% in explaining the total data variability for both methods. Collectively, the first four PCs explained about 90% of the variability in the original data. This suggests that by using only four PCs, most of the information present in the original data can be effectively preserved and represented.

Table 4. Eigen analysis of the Correlation Matrix

Component		W_1	W_2	W_3	W_4	⋯	W_{64}
Eigenvalue	MCD	43.100	7.607	4.847	**2.235**	⋯	0.000076
	MVV	46.765	6.325	3.748	**1.907**	⋯	0.000066
Proportion	MCD	**0.673**	0.119	0.076	0.035	⋯	0.00000119
	MVV	**0.731**	0.099	0.058	0.029	⋯	0.00000103
Cumulative	MCD	0.673	0.792	0.868	**0.903**	⋯	1
	MVV	0.731	0.829	0.888	**0.918**	⋯	1

The equation of the PCs formed from the MCD method can be written as follows:

$$\begin{aligned} \boldsymbol{w}_1 &= -0.141\boldsymbol{z}_1 - 0.137\boldsymbol{z}_2 - 0.144\boldsymbol{z}_3 - 0.143\boldsymbol{z}_4 + 0.089\boldsymbol{z}_5 + \cdots - 0.104\boldsymbol{z}_{63} - 0.097\boldsymbol{z}_{64} \\ \boldsymbol{w}_2 &= \;\; 0.035\boldsymbol{z}_1 - 0.129\boldsymbol{z}_2 - 0.084\boldsymbol{z}_3 - 0.082\boldsymbol{z}_4 - 0.259\boldsymbol{z}_5 + \cdots - 0.093\boldsymbol{z}_{63} + 0.004\boldsymbol{z}_{64} \\ \boldsymbol{w}_3 &= -0.131\boldsymbol{z}_1 - 0.034\boldsymbol{z}_2 - 0.076\boldsymbol{z}_3 - 0.091\boldsymbol{z}_4 - 0.022\boldsymbol{z}_5 + \cdots + 0.214\boldsymbol{z}_{63} + 0.289\boldsymbol{z}_{64} \\ \boldsymbol{w}_4 &= -0.079\boldsymbol{z}_1 + 0.109\boldsymbol{z}_2 + 0.050\boldsymbol{z}_3 - 0.035\boldsymbol{z}_4 + 0.159\boldsymbol{z}_5 + \cdots - 0.236\boldsymbol{z}_{63} + 0.175\boldsymbol{z}_{64} \end{aligned}$$

Meanwhile, for the MVV method, the equation of the PCs can be written as follows:

$$\begin{aligned} \boldsymbol{w}_1 &= -0.138\boldsymbol{z}_1 - 0.136\boldsymbol{z}_2 - 0.139\boldsymbol{z}_3 - 0.136\boldsymbol{z}_4 + 0.102\boldsymbol{z}_5 + \cdots - 0.104\boldsymbol{z}_{63} - 0.112\boldsymbol{z}_{64} \\ \boldsymbol{w}_2 &= \;\; 0.019\boldsymbol{z}_1 - 0.115\boldsymbol{z}_2 - 0.079\boldsymbol{z}_3 - 0.100\boldsymbol{z}_4 - 0.242\boldsymbol{z}_5 + \cdots - 0.129\boldsymbol{z}_{63} + 0.035\boldsymbol{z}_{64} \\ \boldsymbol{w}_3 &= -0.136\boldsymbol{z}_1 - 0.036\boldsymbol{z}_2 - 0.083\boldsymbol{z}_3 - 0.114\boldsymbol{z}_4 - 0.018\boldsymbol{z}_5 + \cdots + 0.217\boldsymbol{z}_{63} + 0.264\boldsymbol{z}_{64} \\ \boldsymbol{w}_4 &= -0.074\boldsymbol{z}_1 + 0.120\boldsymbol{z}_2 + 0.057\boldsymbol{z}_3 - 0.028\boldsymbol{z}_4 + 0.186\boldsymbol{z}_5 + \cdots - 0.238\boldsymbol{z}_{63} + 0.168\boldsymbol{z}_{64} \end{aligned}$$

The MCD and MVV methods were used to select PCs in the SD model with PCR. Therefore, this study utilized one to four PCs (w_1, w_2, w_3, w_4) as predictor variables in SD modeling. Additionally, dummy variables D_1, D_2, and D_3 were also employed as predictor variables.

3.5 Statistical Downscaling Model Using Robust Principal Component Regression

Principal component regression (PCR) was used as the SD model, regressing selected PCs as predictor variables based on the results of the MCD and MVV methods on rainfall in Pangkep District as the response variable. The analysis results with the MCD and MVV methods showed that using the first four PCs could explain the variability of the original variables well, namely more than 90%. Thus, the SD model developed in this study depended on the number of PC variables and dummy variables used in the analysis. The model's accuracy was evaluated based on the coefficient of determination (R^2) and the root mean squared error (RMSE) values.

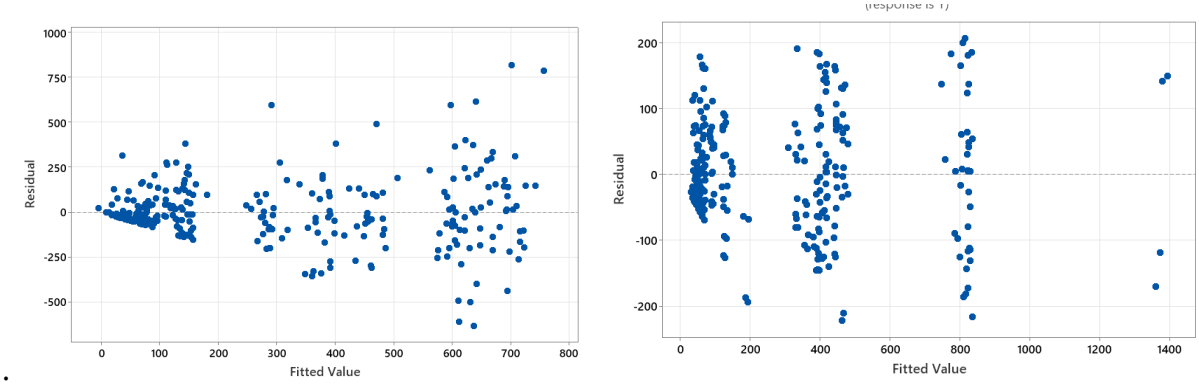
Table 5. R^2 and RMSE Values of the Robust PCR Model

Model	Component	R^2	RMSE
	without dummy variables		
PCR1-MVV	W_1^*	63.39%	181.681
PCR2-MVV	W_1^*, W_2	63.27%	181.994
PCR3-MVV	W_1^*, W_2, W_3	63.20%	182.155
PCR4-MVV	W_1^*, W_2, W_3, W_4	63.08%	182.451
PCR1-MCD	W_1^*	63.32%	181.858
PCR2-MCD	W_1^*, W_2	63.21%	182.141
PCR3-MCD	W_1^*, W_2, W_3	63.20%	182.157
PCR4-MCD	W_1^*, W_2, W_3, W_4	63.08%	182.450
	with dummy variables		
PCRD1-MVV	$W_1^*, D_1^*, D_2^*, D_3^*$	91.53%	87.394
PCRD2-MVV	$W_1^*, W_2^{**}, D_1^*, D_2^*, D_3^*$	**91.59%**	**87.103**
PCRD3-MVV	$W_1^*, W_2^{**}, W_3, D_1^*, D_2^*, D_3^*$	91.62%	86.920
PCRD4-MVV	$W_1^*, W_2^{**}, W_3, W_4, D_1^*, D_2^*, D_3^*$	91.59%	87.073
PCRD1-MCD	$W_1^*, D_1^*, D_2^*, D_3^*$	91.53%	87.382
PCRD2-MCD	$W_1^*, W_2, D_1^*, D_2^*, D_3^*$	91.58%	87.137
PCRD3-MCD	$W_1^*, W_2^{**}, W_3, D_1^*, D_2^*, D_3^*$	91.62%	86.913
PCRD4-MCD	$W_1^*, W_2^{**}, W_3, W_4, D_1^*, D_2^*, D_3^*$	91.59%	87.067

* Significant at $\alpha = 5\%$; ** significant at $\alpha = 10\%$

The analysis results of the robust PCR model using components from the MVV and MCD methods showed a similar level of accuracy (**Table 5**). The recorded R^2 values ranged from 63.08% to 63.39%. This meant that the robust PCR model with PC variables from MVV or MCD could only explain 63% of the data variability. Additionally, both methods resulted in relatively high RMSE values, ranging from 181.681 to 182.451. Furthermore, the use of dummy variables in the model increased the R^2 value to between 91.53% and 91.62%. Moreover, adding dummy variables as predictor variables also reduced the RMSE value by about 52%. Thus, the addition of dummy variables in the robust PCR model significantly improved model accuracy. Overall, the robust PCR model using the first two PCs from the MVV method and three dummy variables (PCRD2-MVV) was the best model in explaining the rainfall data variability in Pangkep District. This conclusion was based on the high R^2 value, low RMSE value, and the significance of all parameters in the model.

Diagnostic checks were conducted on the robust PCR model involving the first two PCs from MVV and dummy variables. **Figure 4** depicted the plot of residuals against fitted values for the PCR2-MVV model, both with and without dummy variables. The spread of residual values in both models indicated heteroscedasticity. The residual variance of the PCR2-MVV model changed with increasing fitted values. Particularly at higher fitted values, there was a larger variation in residuals observed. However, overall, the PCR2-MVV model with dummy variables (PCRD2-MVV) exhibited a relatively more homogeneous spread of residual values. The residuals from this model were distributed relatively evenly, indicating consistent residual variance both at low and high fitted values.



(a) **(b)**

Figure 4. Residual Plot of Model (a) PCR2-MVV dan (b) PCRD2-MVV

The parameter estimation of the PCRD2-MVV model was presented in **Table 6**. This regression model indicated that component W_1 and dummy variables D_1, D_2, D_3 were significant in influencing the response variable (Y) at the 5% significance level. Meanwhile, component W_2 was significant at the 10% significance level. The VIF values for all variables were below 10, indicating no serious multicollinearity. The highest VIF was 5.37 for D_3, which was still within acceptable limits. This model could be considered sufficiently effective in predicting the response variable (rainfall) using the existing predictor variables.

Table 6. Parameter Estimation of the PCRD2-MVV Model

Variable	Coefficient	S.E Coefficient	t-value	p-value	VIF
Constant	734.90*	18.60	39.61	0.000	
W_1	-8.02*	1.15	-6.96	0.000	2.67
W_2	-5.04**	2.96	-1.70	0.090	1.04
D_1	552.30*	45.90	12.04	0.000	1.09
D_2	-355.80*	18.60	-19.12	0.000	2.94
D_3	-623.70*	23.80	-26.19	0.000	5.37

* Significant at $\alpha = 5\%$; ** significant at $\alpha = 10\%$

Based on the parameter estimation results in Table 5.4, the PCRD2-MVV regression model could be written as follows:

$$\hat{Y} = 734.90 - 8.02W_1 - 5.04W_2 + 552.30D_1 - 355.80D_2 - 623.70D_3$$

Next, parameter estimates from components W_1 and W_2 were transformed into parameter estimates for variables X (**Table 7**),

Table 7. Parameter Estimation of the PCRD2-MVV Model

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
Constant	645.518	X_{17}	0.674	X_{34}	0.402	X_{51}	0.478
$\mathbf{X_1}$	0.537	X_{18}	0.537	X_{35}	0.459	X_{52}	-0.136
$\mathbf{X_2}$	0.709	X_{19}	0.469	X_{36}	0.657	X_{53}	-0.179
$\mathbf{X_3}$	0.594	X_{20}	0.607	X_{37}	-0.354	X_{54}	0.074
$\mathbf{X_4}$	0.551	X_{21}	0.095	X_{38}	-0.366	X_{55}	1.207
$\mathbf{X_5}$	0.183	X_{22}	0.094	X_{39}	0.486	X_{56}	0.041
$\mathbf{X_6}$	-0.718	X_{23}	1.653	X_{40}	-0.496	X_{57}	0.113
$\mathbf{X_7}$	-0.225	X_{24}	1.467	X_{41}	0.238	X_{58}	0.226
$\mathbf{X_8}$	0.294	X_{25}	0.524	X_{42}	0.356	X_{59}	0.441
$\mathbf{X_9}$	0.839	X_{26}	0.454	X_{43}	0.445	X_{60}	-0.232
$\mathbf{X_{10}}$	0.637	X_{27}	0.428	X_{44}	-0.075	X_{61}	-0.388
$\mathbf{X_{11}}$	0.545	X_{28}	0.565	X_{45}	-0.356	X_{62}	1.207

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
X_{12}	0.574	X_{29}	0.729	X_{46}	-0.707	X_{63}	0.918
X_{13}	-0.014	X_{30}	0.737	X_{47}	0.585	X_{64}	0.390
X_{14}	-0.917	X_{31}	-0.777	X_{48}	-0.545	D_1	552.300
X_{15}	-0.351	X_{32}	-0.780	X_{49}	0.154	D_2	-355.800
X_{16}	0.012	X_{33}	0.369	X_{50}	0.304	D_3	-623.700

The PCRD2-MVV regression model could be rewritten as,

$$\hat{Y} = 645.518 + 0.537X_1 + 0.709X_2 + 0.594X_3 \ldots + 0.389X_{64} + 552.30D_1 - 355.80D_2 - 623.70D_3$$

3.6 Forecasting Rainfall Data Using Statistical Downscaling Models

The statistical downscaling (SD) model based on principal component regression (PCR) that was developed was further used to forecast rainfall in Pangkep District for the period from January to December 2023. The forecasted rainfall data were validated using several evaluation metrics, namely correlation coefficient, root mean square error of prediction (RMSEP), and mean absolute percentage error (MAPE). The correlation coefficient indicated how well the prediction model captured patterns and trends in the new data. Conversely, RMSEP and MAPE described the accuracy and reliability of the predictions generated by the model. Overall, the correlation coefficient, RMSEP, and MAPE provided a comprehensive overview of the model's performance.

Table 8 presented the performance of the rainfall forecast for Pangkep District in 2023 using the PCR model with components formed based on MCD and MVV results. The table indicated that both MCD and MVV methods yielded similar accuracy in rainfall forecasts. The correlation values between actual and forecasted data ranged from 0.857 to 0.866, with relatively large RMSEP values ranging from 147.421 to 151.878 for the PCR model without dummy variables. This suggested that despite relatively high correlations, the model without dummy variables had a relatively high prediction error rate. It was observed that the models significantly improved prediction performance with the addition of dummy variables. This improvement was evidenced by higher correlation values and lower RMSEP and MAPE values compared to models without dummy variables. The inclusion of dummy variables reduced RMSEP by approximately 47%, highlighting their crucial role in enhancing prediction accuracy. Moreover, increasing the number of components generally enhanced model performance. Overall, the PCRD3-MVV model provided more accurate forecast results with a correlation value of 0.974 and the lowest MAPE value of 7.290.

Table 8. Performance of Rainfall Forecast for Pangkep District in 2023 from Robust PCR Model

Model	Component	Correlation	RMSEP	MAPE
	without dummy variables			
PCR1-MVV	W_1^*	0.859	150.785	10.385
PCR2-MVV	W_1^*, W_2	0.860	150.678	10.598
PCR3-MVV	W_1^*, W_2, W_3	0.864	148.323	9.469
PCR4-MVV	W_1^*, W_2, W_3, W_4	0.866	147.421	9.747
PCR1-MCD	W_1^*	0.857	151.878	10.245
PCR2-MCD	W_1^*, W_2	0.858	151.646	10.810
PCR3-MCD	W_1^*, W_2, W_3	0.864	148.502	9.279
PCR4-MCD	W_1^*, W_2, W_3, W_4	0.866	147.508	9.599
	with dummy variables			
PCRD1-MVV	$W_1^*, D_1^*, D_2^*, D_3^*$	0.971	80.986	9.939
PCRD2-MVV	$W_1^*, W_2^{**}, D_1^*, D_2^*, D_3^*$	0.972	79.899	8.585
PCRD3-MVV	$W_1^*, W_2^{**}, W_3, D_1^*, D_2^*, D_3^*$	0.974	77.658	7.290
PCRD4-MVV	$W_1^*, W_2^{**}, W_3, W_4, D_1^*, D_2^*, D_3^*$	0.974	77.569	7.339
PCRD1-MCD	$W_1^*, D_1^*, D_2^*, D_3^*$	0.971	81.062	9.925
PCRD2-MCD	$W_1^*, W_2, D_1^*, D_2^*, D_3^*$	0.972	79.887	8.671
PCRD3-MCD	$W_1^*, W_2^{**}, W_3, D_1^*, D_2^*, D_3^*$	0.974	77.654	7.320

*Significant at $\alpha = 5\%$; **significant at $\alpha = 10\%$

Figure 5 presented a comparison between actual rainfall data and prediction results using the PCR model with components formed based on MVV and MCD results for Pangkep district in 2023. The first graph

displayed the PCR model equipped with dummy variables (PCRD), while the second graph depicted the PCR model without dummy variables. The PCDR model accurately forecasted rainfall, particularly in January, February, April, May, and June. This indicated that adding dummy variables aided the model in capturing seasonal patterns and fluctuations in rainfall data. Moreover, the PCDR model also demonstrated better performance in predicting high-intensity rainfall, as observed in January and February, which closely matched the actual rainfall amounts.

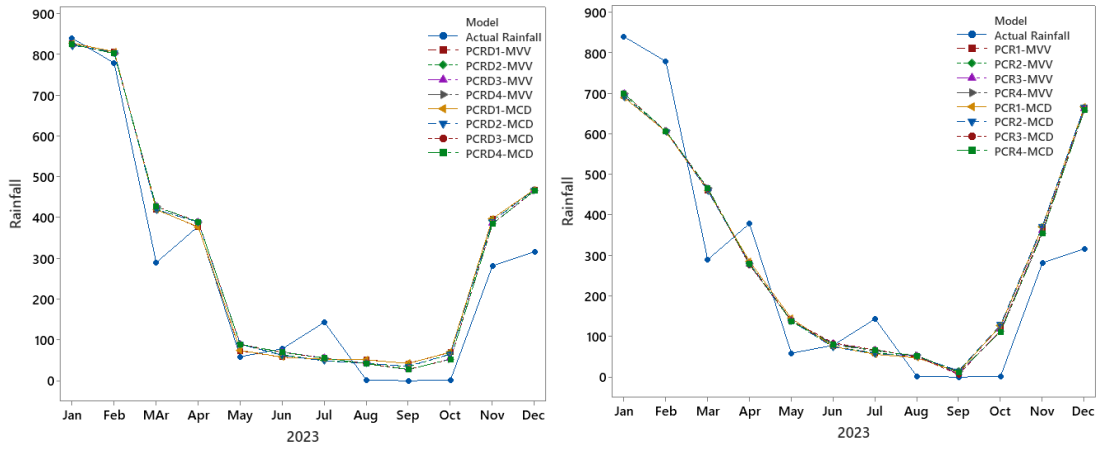


Figure 5. Comparison of Rainfall Forecast in Pangkep District 2023 between Model (a) PCRD and (b) PCR

Figure 5 also illustrates that the PCR model without dummy variables could only predict rainfall with lower accuracy. This model made accurate predictions only at a few points, such as in June and September. However, it failed to predict high-intensity rainfall, especially in January and February. The use of dummy variables in the PCR model significantly improved the accuracy of rainfall predictions, especially during periods of high rainfall. Adding dummy variables allowed the model to better understand and replicate complex rainfall patterns, thereby providing more reliable prediction results. Meanwhile, the use of MVV and MCD methods in both PCR and PCRD models yielded relatively similar performance in forecasting rainfall in Pangkep district.

Similar to previous research findings, the inclusion of dummy variables in the model demonstrated better forecasting accuracy compared to models without dummy variables, as it allowed the model to handle data variation more effectively. Additionally, in this research, the MVV and MCD methods helped manage outliers that could significantly impact the model's accuracy. In PCR models, outliers could have a substantial effect on the principal components used in regression, so handling outliers with robust methods such as MVV or MCD contributed to producing more accurate predictions.

4. CONCLUSIONS

The statistical downscaling model with principal component regression (PCR) effectively predicted rainfall in Pangkep District, addressing multicollinearity and using robust methods like MCD and MVV to handle outliers in principal components. The analysis results indicated that both MVV and MCD methods performed equally well. Both methods effectively explained the variability in rainfall data with coefficient of determination values ranging from 91.53% to 91.62%, and relatively low root mean square error values compared to the model without dummy variables. The addition of dummy variables significantly enhanced the model accuracy. Rainfall predictions for Pangkep District in 2023 showed that both methods also provided similar prediction accuracy. With the inclusion of dummy variables, both methods could predict rainfall patterns closely resembling actual rainfall events, especially during high-intensity rainfall occurrences. Furthermore, the low root mean square error of prediction (between 77.569 and 81.062) indicated minimal deviation between predicted and actual rainfall values, confirming the model's ability to provide predictions close to actual conditions.

ACKNOWLEDGMENT

We would like to thank the leadership of Hasanuddin University through LP2M (Institute for Research and Community Service) at Hasanuddin University for funding this internal research. This research was funded through the "Penelitian Dosen Pemula Unhas" (PDPA) for the fiscal year 2024. We also extend our thanks to all staff members who provided facilities to complete this research.

REFERENCES

[1] E. Zorita and H. Storch, "THE ANALOG METHOD AS A SIMPLE STATISTICAL DOWNSCALING TECHNIQUE: comparison with more complicated methods," *Journal of Climate*, vol. 12, no. 8, pp. 2474-2489, August 1999.

[2] R. L. Wilby and C. W. Dawson, "THE STATISTICAL DOWNSCALING MODEL: INSIGHTS FROM ONE DECADE OF APPLICATION," *International Journal of Climatology*, vol. 33, no. 7, pp. 1707-1719, June 2013.

[3] R. I. Rakhmalia, A. M. Soleh and B. Sartono, "RAINFALL ESTIMATION WITH STATISTICAL DOWNSCALING TECHNIQUE USING TWEEDIE SCATTER CLUSTERWISE REGRESSION," *Indonesian Journal of Statistics and Its Applications*, vol. 4, no. 3, pp. 473 - 483, 2020.

[4] E. K. Siabi, A. T. Kabobah, K. Akpoti, G. K. Anornu, M. Amo-Boateng and E. Nyantakyi, "STATISTICAL DOWNSCALING OF GLOBAL CIRCULATION MODELS TO ASSESS FUTURE CLIMATE CHANGES IN THE BLACK VOLTA BASIN OF GHANA," *Environmental Challenges*, vol. 5, no. 100249, pp. 1-17, December 2021.

[5] R. N. Rachmawati, I. Sungkawa and A. Rahayu, "EXTREME RAINFALL PREDICTION USING BAYESIAN QUANTILE REGRESSION IN STATISTICAL DOWNSCALING MODELING," *Procedia Computer Science*, vol. 157, pp. 406-413, 2019.

[6] D. C. Montgomery, E. A. Peck and G. G. Vining, INTRODUCTION TO LINEAR REGRESSION ANALYSIS, Sixth ed., New York: John Willey and Sons Inc, 2021.

[7] Sutikno, Setiawan and H. Purnomoadi, "STATISTICAL DOWNSCALING OUTPUT GCM MODELING WITH CONTINUUM REGRESSION AND PRE-PROCESSING PCA APPROACH," *Journal for Technology and Science*, vol. 21, no. 3, pp. 109-118, 2010.

[8] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira and R. M. Manggio, "USE OF PRINCIPAL COMPONENT ANALYSIS (PCA) AND HIERARCHICAL CLUSTER ANALYSIS (HCA) FOR MULTIVARIATE ASSOCIATION BETWEEN BIOACTIVE COMPOUNDS AND FUNCTIONAL PROPERTIES IN FOODS: A CRITICAL PERSPECTIVE," *Trends in Food Science & Technology*, vol. 72, pp. 83-90, February 2018.

[9] X. Chen, B. Zhang, T. Wang, A. Bonni and G. Zhao, "ROBUST PRINCIPAL COMPONENT ANALYSIS FOR ACCURATE OUTLIER SAMPLE DETECTION IN RNA-SEQ DATA," *BMC Bioinformatics*, vol. 21, no. 269, pp. 1-20, 2020.

[10] M. Ahsan, M. Mashuri, M. H. Lee, H. Kuswanto and D. D. Prastyo, "ROBUST ADAPTIVE MULTIVARIATE HOTELLING'S T2 CONTROL CHART BASED ON KERNEL DENSITY ESTIMATION FOR INTRUSION DETECTION SYSTEM," *Expert Systems with Applications*, vol. 145, no. 113105, pp. 1-30, May 2020.

[11] C. Leys, O. Klein, Y. Dominicy and C. Ley, "DETECTING MULTIVARIATE OUTLIERS: USE A ROBUST VARIANT OF THE MAHALANOBIS DISTANCE," *Journal of Experimental Social Psychology*, vol. 74, pp. 150-156, January 2018.

[12] P. J. Rousseeuw and K. V. Driessen, "A FAST ALGORITHM FOR THE MINIMUM COVARIANCE DETERMINANT ESTIMATOR," *Technometrics*, vol. 41, no. 3, p. 212–223, March 2012.

[13] E. Polat and H. Ali, "ADAPTIVE REWEIGHTED MINIMUM VECTOR VARIANCE ESTIMATOR OF COVARIANCE USED FOR AS A NEW ROBUST APPROACH TO PARTIAL LEAST SQUARES REGRESSION," *Journal of Science (Gazi University)*, vol. 33 , no. 3, pp. 872-890, December 2020.

[14] A. M. Soleh, A. H. Wigena, A. Djuraidah and A. Saefuddin, "STATISTICAL DOWNSCALING TO PREDICT MONTHLY RAINFALL USING LINEAR REGRESSION WITH L1 REGULARIZATION (LASSO)," *Applied Mathematical Sciences*, vol. 9, p. 5361–5369, 2015.

[15] S. Sahriman, Anisa and V. Koerniawan, "STATISTICAL DOWNSCALING MODELING WITH DUMMY VARIABLES BASED ON HIERARCHICAL AND NON-HIERARCHICAL CLUSTERING TECHNIQUES TO FORECAST RAINFALL," *Indonesian Journal of Statistics and Its Applications*, vol. 3, no. 3, pp. 295-309, October 2019.

[16] 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, September 2021.

[17] R. P. Putra, D. Anggraeni and A. F. Hadi, "PROJECTION PURSUIT REGRESSION (PPR) ON STATISTICAL DOWNSCALING MODELING FOR DAILY RAINFALL FORECASTING," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 326-332, 2021.

[18] V. P. Butar-butar, A. M. Soleh and A. H. Wigena, "CLUSTERWISE REGRESSION MODELING IN STATISTICAL DOWNSCALING FOR MONTHLY RAINFALL ESTIMATION," *Indonesian Journal of Statistics and Its Applications* , vol. 3, no. 3, pp. 236-246, 2019.

[19] S. Sahriman, A. J. Jaya and A. M. A. Siddik, "A COMPARISON OF THE PRINCIPAL COMPONENT REGRESSION METHODS AND THE ROBUST PRINCIPAL COMPONENT REGRESSION WITH MINIMUM VECTOR VARIANCE IN STATISTICAL DOWNSCALING MODELS," in *AIP Conference Proceedings*, 2022.

[20] S. Sahriman and A. S. Yulianti, "STATISTICAL DOWNSCALING MODEL WITH PRINCIPAL COMPONENT REGRESSION AND LATENT ROOT REGRESSION TO FORECAST RAINFALL IN PANGKEP REGENCY," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0401-0410, April 2023.

[21] D. A. Belsley, Conditioning Diagnostics: COLLINEARITY AND WEAK DATA IN REGRESSION, New York : John Wiley & Sons Inc, 1991.

[22] N. Shrestha, "DETECTING MULTICOLLINEARITY IN REGRESSION ANALYSIS," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39-42, June 2020.

[23] A. H. Wigena, "STATISTICAL DOWNSCALING MODELING WITH PROJECTION PURSUIT REGRESSION FOR MONTHLY RAINFALL FORECASTING: A CASE STUDY OF MONTHLY RAINFALL IN INDRAMAYU," IPB University, Bogor, 2006.

[24] R. A. Johnson and D. W. Wichern, APPLIED MULTIVARIATE STATISTICAL ANALYSIS, Sixth ed., New Jersey (NJ): Pearson Prentice Hall, 2007.

[25] I. T. Jolliffe, PRINCIPAL COMPONENT ANALYSIS, Second Edition ed., New York (NY): Springer-Verlag, 2002.

[26] S. Sahriman, "STATISTICAL DOWNSCALING MODEL WITH TIME LAG OF GLOBAL CIRCULATION MODEL DATA FOR RAINFALL FORECAST," IPB Univiversity, Bogor, Indonesia, 2014.

[27] M. Hubert and m. Debruyne, "MINIMUM COVARIANCE DETERMINANT," *wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 1, pp. 36-43, january 2010.

[28] E. T. Herdiani, "MODIFICATION OF ROBUST ESTIMATORS IN MULTIVARIATE OUTLIER LABELING," *Jurnal Matematika Statistika dan Komputasi*, vol. 14, no. 1, pp. 46-53, July 2017.

[29] D. E. Herwindiati and S. M. Isa, "THE ROBUST PRINCIPAL COMPONENT USING MINIMUM VECTOR VARIANCE," in *Proceedings of the World Congress on Engineering*, 2009.