BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0401–0410
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0401-0410

STATISTICAL DOWNSCALING MODEL WITH PRINCIPAL COMPONENT REGRESSION AND LATENT ROOT REGRESSION TO FORECAST RAINFALL IN PANGKEP REGENCY

Sitti Sahriman [1*], Andi Sri Yulianti [2]

[1,2]*Statistics Department, Faculty of Mathematics and Natural Sciences, Hasanuddin University*
Perintis Kemerdekaan Street KM.10, Makassar City, South Sulawesi 90245, Indonesia

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

ABSTRACT

Article History:

Received: 6th November 2022
Revised: 26th January 2023
Accepted: 9th February 2023

Keywords:

Dummy Variable;
Global Circulation Model;
Statistical Downscaling;
Principal Component;
Principal Component Regression;
Latent Root Regression

Climate information, especially rainfall, is needed by various sectors in Indonesia, including the marine and fisheries sectors. Estimating high-resolution climate models continues to develop by involving global-scale climate variables, one of which is the global circulation model (GCM) output precipitation. Statistical downscaling (SD) relates global scale climate variables to local scales. Principal component regression (PCR) and latent root regression (LRR) techniques are statistical methods used in the SD model to overcome the high correlation between GCM data grids. PCR focuses on the variability in the predictor variables, while the LRR focuses on the variability between the response variables and predictors. This method was applied to Pangkep Regency rainfall data as a local scale response variable and GCM precipitation as a predictor variable (January 1999 to December 2020). This study aimed to obtain the number of principal components (PC) in the SD model and the forecast value of the 2020 rainfall data. In addition, the dummy variable resulting from K-means was used as a predictor variable in PCR and LRR. The result is that using the first 11-15 PC has a cumulative diversity proportion of 98%. Furthermore, by using the data for the 1999-2019 period, adding a dummy variable to the PCR can increase the accuracy of the model (the coefficient of determination is 92.27%-92.43%). However, LRR with and without dummy variables produces the same model accuracy. In general, the LRR model is better at explaining the diversity of the Pangkep District rainfall data than the PCR model. The prediction of rainfall for the 2020 period at LRR with 13 PC is accurate based on the highest correlation value (0.97) and the lowest root mean square error prediction (75.17).

How to cite this article:

S. Sahriman and A. S. Yulianti., "STATISTICAL DOWNSCALING MODEL WITH PRINCIPAL COMPONENT REGRESSION AND LATENT ROOT REGRESSION TO FORECAST RAINFALL IN PANGKEP REGENCY," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0401-0410, March 2023.

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 one of the largest archipelagic countries in the world and has the opportunity to develop potential in the fisheries and marine sector, including salt production. South Sulawesi Province is included in the 14 salt-producing regions in Indonesia. However, salt production in South Sulawesi decreased in 2020, reaching only 45,310 tons, with production in 2019 reaching 140,338 tons. In 2021, the salt production target in South Sulawesi was 46,500 tons. However, South Sulawesi's salt production until the third quarter only reached 466.05 tons. The salt production comes from five South Sulawesi districts: Pangkep, Selayar Islands, Jeneponto, Takalar, and Maros. Nationally, the national salt production from PT Garam (Persero) and Garam Rakyat until January 15, 2021, only reached 1.3 million tons from the targeted 3 million tons in 2020 and 3.1 million tons in 2021. The high rainfall is one of the factors that affect the decline in national salt production.

Rainfall is one of the factors that can affect the increase and decrease in salt production. The average rainfall intensity and rainfall pattern in a year are indicators that are closely related to the length of the dry season. The dry season's distance will affect the water evaporation rate at the salt production site. High levels of rainfall will have a negative impact on salt production. Therefore, the estimation of rainfall is used by salt farmers to determine the right time for making salt to minimize crop failure. Currently, many climate models have been developed to improve climate information, one of which is utilizing the output of the global circulation model (GCM). However, the GCM output climate information is still worldwide, so it has low accuracy in predicting local scale climates. GCM can be used to obtain local scale climate information using statistical downscaling techniques [1].

Statistical downscaling (SD) is a model that relates global scale climate variables to local scale climate variables. The SD approach uses a regression model to determine the functional relationship between the response variable in the form of local climate and the predictor variable in the form of global scale climate outputs of GCM [2]. The output of GCM is spatial and temporal data, creating a spatial correlation between grids in one domain. The larger the GCM data domain used, the more predictor variables in the SD model, resulting in a more complex model. Therefore, pre-processing is needed in this case to reduce dimensions and, at the same time, overcome multicollinearity in the data.

Principal component regression (PCR) is a statistical method that can overcome multicollinearity by combining linear regression with principal component analysis (PCA) [3]. PCA focuses on diversity in correlated predictor variables [4]. In addition to PCR, the latent root regression (LRR) method can also overcome multicollinearity in the data. LRR is an extension method of PCR. The difference lies in the formation and selection of the principal component (PC). Data dimension reduction in PCR only involves predictor variables. In contrast, the LRR method combines the matrix of response variables with predictor variables in the formation of PC. The LRR method forms a PC by calculating the relationship between the predictor variable and the response variable so that the PC obtained contains more information than the PCR method [5].

Previous researchers have widely used the PCR and LRR methods, including [6] comparing PCR and partial least squares regression (PLSR) in predicting rainfall in el-nino, la-nina, and normal conditions in Indramayu Regency. [7] predicted extreme rainfall data with functional principal component quantile regression. [8] added dummy variables based on hierarchical and non-hierarchical cluster techniques in SD modeling for rainfall estimation. Furthermore, [9] used LRR to predict car sales in the United States from 1961-1990. [10] applied the application of LRR in dealing with multicollinearity in multiple linear regression models. In addition, [11] also uses LRR to model the factors that affect the JCI in the Indonesia Stock Exchange.

This study compares the SD model with the PCR and LRR methods in estimating the Pangkep Regency rainfall data. In addition, dummy variables from the K-means cluster technique are used to improve the model's accuracy and the rainfall data's prediction results. Prediction of rainfall data using the SD model with dummy variables has a higher accuracy [8].

2. RESEARCH METHODS

2.1 Data

The data used in this research is the output of GCM climate model intercomparison project (CMIP5) precipitation data in mm/month. This data can be obtained from the web http://www.climatexp.knmi.nl/ (issued by KNMI Netherlands). The GCM domain used in this study is several square grids measuring 8×8 grids (2.5°×2.5° for each grid) at 119.57°E to 129.37°E and -14.83°S to 5.17°N above the Pangkep Regency area. . The GCM output data is used as the predictor variable (*X*), and the Pangkep Regency rainfall data is used as a response variable (*Y*). The average rainfall data in Pangkep Regency, South Sulawesi, for 1999-2020 was obtained from the BMKG station IV Makassar. The rainfall data is the average rainfall from 3 rain posts in Pangkep Regency, Bungoro, Ma'rang, and Labakkang. In addition, dummy variables (*D*) based on the K-Means non-hierarchical cluster technique were used in this study to improve the accuracy of the model.

2.2 Analysis Method

The analytical methods used are PCR and LRR. The PCR method begins with PCA to reduce the dimensions of the precipitation data. It produces some PC that is selected based on the eigenvalues and the proportion of diversity. Furthermore, some PC were used as predictor variables in the PCR method. Meanwhile, data reduction in the LRR method, in addition to involving precipitation data as a predictor variable, also involves rainfall data as a response variable so that several PC are obtained in the LRR method.

The stages of analysis carried out in this study are as follows:

1. Determine the group of rainfall data based on the K-Means cluster technique. The elbow method used to determine the optimum number of clusters is based on the value within the sum of squares (WSS) [12].
2. Identifying multicollinearity in precipitation data using variance inflation factors (VIF) with R_j^2 is the coefficient of determination of the j-th predictor variable regression with other predictor variables [13].

$$VIF_j = \frac{1}{1-R_j^2} \qquad , j = 1,2, \dots ,64$$

3. Divide the data into modeling data (1999-2019) and validation data (2020).
4. Apply the SD technique using PCR and LRR with additional dummy variables.

The stages of data analysis using the PCR method are as follows [14]:

a. Specifies the shape of the Z_j data transformation

$$Z_j = \frac{(X_j-\mu_j)}{\sqrt{\sigma_j{}^2}}$$

The matrix notation can be written as

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

with $\boldsymbol{V}^{1/2} = diag\left(\sqrt{\sigma_1{}^2}, \sqrt{\sigma_2{}^2}, \dots, \sqrt{\sigma_p{}^2}\right), E(\boldsymbol{Z}) = \boldsymbol{0}$ dan $\boldsymbol{Z}$ is the standardized matrix of the original $\boldsymbol{X}$ variables.

b. Forming the variance-covariance matrix of the $\boldsymbol{Z}$ variable, i.e

$$Cov(\boldsymbol{Z}) = \left(\boldsymbol{V}^{1/2}\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{V}^{1/2}\right)^{-1} = \boldsymbol{R}$$

With $\boldsymbol{R}$ is the correlation matrix of the original variable $\boldsymbol{X}$.

c. Calculates the value of the feature root (λ_j) and the feature vector ($\boldsymbol{e}_j$), as well as the PC score ($\boldsymbol{w}_j$) from the variance-covariance matrix of the original variable Z or the correlation matrix of the original variable X.

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

d. Choose a PC with a 98% diversity proportion

e. Regressing Y with selected $\boldsymbol{w}_j$.

f. Transform the regression equation from $\boldsymbol{w}_j$ to $\boldsymbol{z}_j$ and $\boldsymbol{z}_j$ to $\boldsymbol{x}_j$.

The stages of data analysis using the LRR method are as follows [15]:

a. Standardize the data on the response variable and predictor variable using the equation.

$$\boldsymbol{Z}_{y_i} = \frac{(y_i-\bar{y})}{s_Y};\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n};\ S_Y{}^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\ ;\ \boldsymbol{Z}_{x_i} = \frac{(x_i-\bar{x})}{s_X};\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}; S_X{}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

b. Calculating the augmented correlation matrix, which is a correlation matrix that combines response variables and predictors that have been standardized using the equation.

$$\boldsymbol{R} = \boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*} = \begin{bmatrix} 1 & \gamma_{2Y} & \cdots & \gamma_{rY} \\ \gamma_{2Y} & 1 & \cdots & \gamma_{r1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{rY} & \gamma_{r1} & \cdots & 1 \end{bmatrix}$$

c. Calculating the eigenvalues and eigenvectors from the correlation matrix using the equation.

$$\left|\boldsymbol{R} - \lambda_j \boldsymbol{I}\right| = 0 \text{ dan } \left(\boldsymbol{R} - \lambda_j \boldsymbol{I}\right)\boldsymbol{\gamma}_j = 0$$

d. Forming the principal components through principal component analysis based on the eigenvalues and eigenvectors formed. The principal component that will be used is the principal component which has the eigenvalue λ_j > 0.30.

$$\boldsymbol{w}_j = \boldsymbol{Z}^{*}\boldsymbol{\gamma}_j \quad ; \quad \boldsymbol{w}_j = \gamma_{0j}\boldsymbol{Z}_{\boldsymbol{y}} + \boldsymbol{Z}_x\boldsymbol{\gamma}_j^0$$

e. Estimating data free of multicollinearity using the modified least squares method on data that has been standardized using the equation.

$$\widehat{\boldsymbol{\beta}}^{*} = \begin{bmatrix} \hat{\beta}_1{}^{*} \\ \hat{\beta}_2{}^{*} \\ \vdots \\ \hat{\beta}_p{}^{*} \end{bmatrix} = c\sum_{j=0}^{l}\gamma_{0j}\lambda_j{}^{-1}\begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \\ \vdots \\ \gamma_{pj} \end{bmatrix}, j = 1,2,\dots,l; l < p+1$$

f. Estimating the parameters of the original data using the equation.

$$\hat{\beta}_j = \frac{\hat{\beta}_j{}^{*}}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}, j = 1,2,\dots,k$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2 - \cdots - \hat{\beta}_k\bar{X}_k$$

5. Model validation on data for the 2020 period uses correlation statistics, and root mean squared error of prediction (RMSEP).

3. RESULTS AND DISCUSSION

3.1. Formation of Dummy Variables with the K-Means Clustering Method

Adding dummy variables to SD modeling aims to improve model accuracy and data forecast results. The determination of the dummy variable is based on the grouping of rainfall data. K-means clustering is a non-hierarchical method of a group by specifying the number of k centroids as the basis for determining the number of clusters produced. The Elbow method is used to determine the optimum number of clusters by selecting a value within sum of square (WSS), which is not significantly different for the following k clusters.

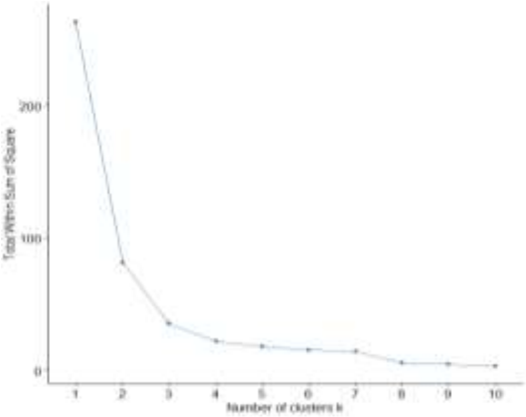

Figure 1. Determination of The Optimum K Cluster Using The Elbow Method

Figure 1 shows that the optimum number is 4 clusters. This is because there is a significant difference in the WSS value in clusters 1, 2, and 3. Meanwhile, in cluster 4 and so on, there is no significant change in the WSS value (the line formed is almost linear). Furthermore, the formation of rainfall data groups using the K-means method with a number of clusters of as many as 4.

The K-means results show that groups 1, 2, 3, and 4 are rainfall with an intensity of 0.00-231.50 mm/month, 231.51-607.00 mm/month, 607.01-1019.00 mm/month, and 1019.01-1540.50 mm /month. Thus three dummy variables are used as predictor variables in SD modeling. Table 1 is the value of the dummy variables $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, and $\boldsymbol{D_3}$. Group 1 with values $\boldsymbol{D_2 = 1}$ and $\boldsymbol{D_1 = D_3 = 0}$ totaled 139 observations. Furthermore, group 2 with values $\boldsymbol{D_1 = 1}$ and $\boldsymbol{D_2 = D_3 = 0}$ had 89 observations. Group 3 has a value of $\boldsymbol{D_3 = 1, D_1 = D_2 = 0}$ and consists of 32 observations. Meanwhile, four observations were included in group 4 with a value of $\boldsymbol{D_1 = D_2 = D_3 = 0}$.

Table 1. Dummy Variables

No	Time	Y	D_1	D_2	D_3
1	Jan-1999	1017.50	0	0	1
2	Feb-1999	427.00	0	1	0
3	Mar-1999	444.00	0	1	0
4	Apr-1999	577.00	0	1	0
5	May-1999	202.50	1	0	0
6	Jun-1999	61.50	1	0	0
7	Jul-1999	96.00	1	0	0
8	Aug-1999	0.00	1	0	0
9	Sep-1999	0.00	1	0	0
10	Oct-1999	248.50	0	1	0
11	Nov-1999	443.00	0	1	0
12	Dec-1999	1019.00	0	0	1
13	Jan-2000	714.50	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
60	Dec-2003	1252.50	0	0	0
61	Jan-2004	526.50	0	1	0
62	Feb-2004	863.50	0	0	1
63	Mar-2004	693.50	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
121	Jan-2009	1540.50	0	0	0
122	Feb-2009	814.50	0	0	1
123	Mar-2009	160.00	1	0	0
124	Apr-2009	194.50	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮
260	Aug-2020	40.30	1	0	0
261	Sep-2020	15.70	1	0	0
262	Oct-2020	197.30	1	0	0
263	Nov-2020	355.70	0	1	0
264	Dec-2020	961.30	0	0	1

3.2. *Variance Inflation Factors*

Multicollinearity shows a strong correlation between two or more predictor variables in a multiple regression model. Multicollinearity must to be detected early in data analysis to improve accuracy in SD modeling. The VIF value can be used to determine multicollinearity in the data. If the VIF value is more than 10, it indicates a significant multicollinearity. Table 2 shows the VIF value of the GCM precipitation variable ($\mathbf{X_1 - X_{64}}$) ranging from 3.54 – 2800.83. Thus, the PCR and LRR methods were used in the SD modeling of rainfall data in the Pangkep Regency.

Table 2. VIF Values

No	Predictor	VIF	No	Predictor	VIF
1	X_1	54.76	33	X_{33}	2350.63
2	X_2	562.56	34	X_{34}	2785.64

Table 2. VIF Values

No	Predictor	VIF	No	Predictor	VIF
3	X_3	573.19	35	X_{35}	1596.50
4	X_4	464.27	36	X_{36}	1122.00
5	X_5	74.05	37	X_{37}	147.46
6	X_6	24.13	38	X_{38}	37.42
7	X_7	47.52	39	X_{39}	12.44
8	X_8	34.32	40	X_{40}	107.60
9	X_9	450.85	41	X_{41}	1634.75
10	X_{10}	1327.54	42	X_{42}	2800.83
11	X_{11}	1072.88	43	X_{43}	1462.34
12	X_{12}	1150.48	44	X_{44}	1112.93
13	X_{13}	159.74	45	X_{45}	198.23
14	X_{14}	38.65	46	X_{46}	13.56
15	X_{15}	27.26	47	X_{47}	29.29
16	X_{16}	12.63	48	X_{48}	86.86
17	X_{17}	1135.03	49	X_{49}	758.39
18	X_{18}	1388.34	50	X_{50}	2503.08
19	X_{19}	846.05	51	X_{51}	1458.47
20	X_{20}	1018.00	52	X_{52}	549.92
21	X_{20}	49.58	53	X_{53}	302.09
22	X_{22}	3.54	54	X_{54}	33.21
23	X_{23}	7.63	55	X_{55}	40.74
24	X_{24}	8.19	56	X_{56}	114.69
25	X_{25}	1947.69	57	X_{57}	67.94
26	X_{26}	2320.58	58	X_{58}	995.73
27	X_{27}	1599.96	59	X_{59}	677.30
28	X_{28}	968.32	60	X_{60}	143.15
29	X_{29}	163.18	61	X_{61}	189.48
30	X_{30}	52.65	62	X_{62}	28.90
31	X_{31}	5.37	63	X_{63}	57.14
32	X_{32}	28.28	64	X_{64}	87.92

3.3. Formation of Principal Components

The initial stage in the PCR and LRR method is done by forming a PC based on the eigenvalues and eigenvectors of the covariance variance matrix. PC is a linear combination of the original variables. In the PCR method, PC formation is only based on GCM precipitation data. Meanwhile, PC formation in the LRR method involves GCM precipitation data ($\boldsymbol{X}$ as predictor variable) and rainfall data for Pangkep Regency ($\boldsymbol{Y}$ as response variable). PCR focuses on the variability in the predictor variables, while the LRR focuses on the diversity between the predictor and response variables.

The determination of the number of PC used in the SD model is based on the proportion of total diversity of about 98%. In addition, the selection of the number of PC in the LRR method is based on the eigenvalue $\boldsymbol{\lambda} \geq \mathbf{0.1}$. Table 3 presents the eigenvalues and proportions of variance described by PC in the PCR and LRR methods. The proportion of PC cumulative diversity of 98% was achieved by more than the first 11 PC. In addition, the eigenvalues indicate that the first 15 PC have $\boldsymbol{\lambda} \geq \mathbf{0.1}$. Furthermore, the SD model simulation is based on the number of PC involved in the model, namely 11 PC to 15 PC.

Table 3. Eigenanalysis of PCR and LRR Methods

Component	Method	Principal Component							
		$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	⋯	$\boldsymbol{w_{11}}$	$\boldsymbol{w_{12}}$	$\boldsymbol{w_{13}}$	$\boldsymbol{w_{14}}$	$\boldsymbol{w_{15}}$
Eigenvalue	PCR	51.486	4.646	⋯	0.167	0.15	0.124	0.105	**0.095**
(λ)	LRR	52.099	4.667	⋯	0.186	0.166	0.150	0.124	**0.103**
Proportion	PCR	0.804	0.073	⋯	0.003	0.002	0.002	0.002	0.001
	LRR	0.802	0.072	⋯	0.003	0.003	0.002	0.002	0.002
Cumulative	PCR	0.804	0.877	⋯	**0.981**	0.983	0.985	0.987	0.988
	LRR	0.802	0.873	⋯	**0.979**	0.981	0.983	0.985	0.987

3.4. SD Model with PCR and LRR Methods

The SD model uses the PCR and LRR methods to get the forecast value of the Pangkep Regency rainfall. The PCR method uses PCA to obtain PC as an independent predictor variable. Similar to PCR, the LRR method also produces PC as a predictor variable in the model. The number of PC used in each PCR and LRR method is 11 PC to 15 PC. The PCR1 model is the result of SD modeling using the PCR method, which involves 11 components, namely $\boldsymbol{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}}$ as predictor variables. Likewise, other models correspond to the number of predictors used. In addition, dummy variables $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, and $\boldsymbol{D_3}$ were added to the PCR (PCR-dummy) and LRR (LRR-dummy) methods.

Table 4 presents the coefficient of determination ($\boldsymbol{R^2}$) and root mean squared error (RMSE) of various SD models based on the number of PC and dummy variables. An accurate model produces the highest $\boldsymbol{R^2}$ value with the smallest RMSE value. Table 4 explains that the PCR1-PCR5 model with a dummy resulted in a higher coefficient of determination ($\boldsymbol{R^2}$) (range 92.27%-92.43%) than the model without a dummy (range 62.43%-63.02%). Adding the dummy variable can also reduce the RMSE value to 83.98-84.82. Meanwhile, the LRR model shows a relatively similar model with or without a dummy. The resulting $\boldsymbol{R^2}$ the value ranges from 94.79%-99.92%, and the RMSE ranges from 8.52-69.51. In general, based on the obtained $\boldsymbol{R^2}$ value, the LRR method can explain the diversity of rainfall data better than the PCR method.

Table 4. Values of $\boldsymbol{R^2}$ and RMSE of PCR and LRR Methods

Method	Model	Predictor Variable	R^2	RMSE
PCR	PCR1	$w_1 - w_{11}$	62.71%	186.33
	PCR2	$w_1 - w_{12}$	62.57%	186.69
	PCR3	$w_1 - w_{13}$	62.43%	187.05
	PCR4	$w_1 - w_{14}$	63.02%	185.57
	PCR5	$w_1 - w_{15}$	62.88%	185.93
	PCR1-*dummy*	$w_1 - w_{11}, D_1, D_2, D_3$	92.29%	84.74
	PCR2-*dummy*	$w_1 - w_{12}, D_1, D_2, D_3$	92.29%	84.71
	PCR3-*dummy*	$w_1 - w_{13}, D_1, D_2, D_3$	92.27%	84.82
	PCR4-*dummy*	$w_1 - w_{14}, D_1, D_2, D_3$	92.33%	84.52
	PCR5-*dummy*	$w_1 - w_{15}, D_1, D_2, D_3$	92.43%	83.98
LRR	LRR1	$w_1 - w_{11}$	99.66%	17.68
	LRR2	$w_1 - w_{12}$	99.81%	13.12
	LRR3	$w_1 - w_{13}$	99.82%	12.89
	LRR4	$w_1 - w_{14}$	99.82%	12.8
	LRR5	$w_1 - w_{15}$	99.92%	8.52
	LRR1-*dummy*	$w_1 - w_{11}, D_1, D_2, D_3$	94.79%	69.51
	LRR2-*dummy*	$w_1 - w_{12}, D_1, D_2, D_3$	98.21%	40.71
	LRR3-*dummy*	$w_1 - w_{13}, D_1, D_2, D_3$	98.70%	34.78
	LRR4-*dummy*	$w_1 - w_{14}, D_1, D_2, D_3$	99.22%	26.83
	LRR5-*dummy*	$w_1 - w_{15}, D_1, D_2, D_3$	99.42%	23.23

3.5. Forecasting Rainfall Data with PCR and LRR Methods

This section is the validation phase of the SD model using the PCR and LRR methods. The SD model obtained is then used to predict the Pangkep Regency rainfall data for January to December 2020. The measure of the goodness of the forecast results uses the highest correlation value and the lowest RMSEP. Table 5 shows the correlation and RMSEP values of the SD model with PCR and LRR methods. The addition of variables in the PCR model can increase the correlation value to around 0.168-0.189 and reduce the RMSEP value to about 96.77-104.11. The addition of dummy variables in the PCR model can increase the accuracy of the forecast results of rainfall data. In the LRR model, the results of the forecasted rainfall data show relatively similar accuracy to the model with or without a dummy variable. The resulting correlation values ranged from 0.77 to 0.97, and the RMSEP ranged from 75.17 to 288.71. In general, the LRR3-dummy model involving the variables $\boldsymbol{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}, w_{12}, w_{13}, D_1, D_2, D_3}$ is the best model based on the highest correlation value (0.97) and the lowest RMSEP (75.17).

Table 5. Correlation and RMSEP Values of PCR and LRR Methods

Method	Model	Predictor Variable	Correlation	RMSEP
PCR	PCR1	$w_1 - w_{11}$	0.789	180.55
	PCR2	$w_1 - w_{12}$	0.786	181.97
	PCR3	$w_1 - w_{13}$	0.787	181.79
	PCR4	$w_1 - w_{14}$	0.759	195.11
	PCR5	$w_1 - w_{15}$	0.760	195.14
	PCR1-*dummy*	$w_1 - w_{11}, D_1, D_2, D_3$	0.955	85.21
	PCR2-*dummy*	$w_1 - w_{12}, D_1, D_2, D_3$	0.953	86.62
	PCR3-*dummy*	$w_1 - w_{13}, D_1, D_2, D_3$	0.953	87.06
	PCR4-*dummy*	$w_1 - w_{14}, D_1, D_2, D_3$	0.948	91.03
	PCR5-*dummy*	$w_1 - w_{15}, D_1, D_2, D_3$	0.957	83.78
LRR	LRR1	$w_1 - w_{11}$	0.87	136.8
	LRR2	$w_1 - w_{12}$	0.81	163.66
	LRR3	$w_1 - w_{13}$	0.80	166.31
	LRR4	$w_1 - w_{14}$	0.81	166.24
	LRR5	$w_1 - w_{15}$	0.77	182.48
	LRR1-*dummy*	$w_1 - w_{11}, D_1, D_2, D_3$	0.53	288.71
	LRR2-*dummy*	$w_1 - w_{12}, D_1, D_2, D_3$	0.93	116.11
	LRR3-*dummy*	$w_1 - w_{13}, D_1, D_2, D_3$	0.97	75.17
	LRR4-*dummy*	$w_1 - w_{14}, D_1, D_2, D_3$	0.92	104.04
	LRR5-*dummy*	$w_1 - w_{15}, D_1, D_2, D_3$	0.94	94.44

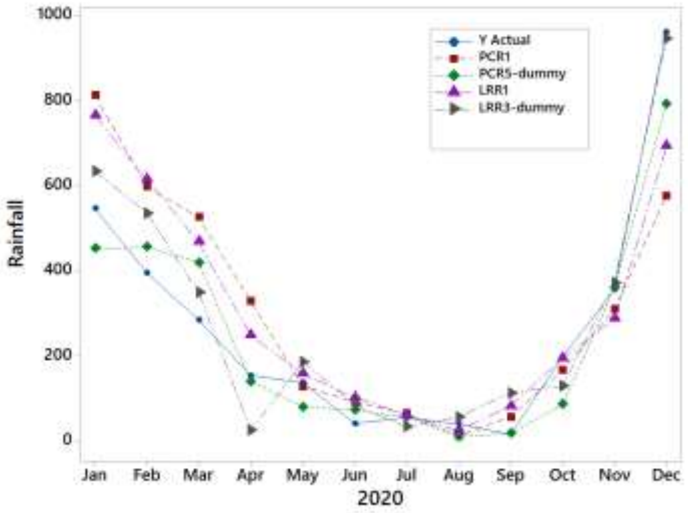

Figure 2. Plot of Actual Rainfall and Estimated Rainfall of PCR1, PCR5-dummy, LRR1, LRR3-dummy Models

Figure 2 presents the actual data plot and the forecast results of the PCR1, PCR5-dummy, LRR1, and LRR3-dummy models. Rainfall data for January, February, and March produce a higher estimated value than the actual value. Meanwhile, the estimated rainfall for the November and December periods is lower than the actual value. Furthermore, the predicted value of low-intensity rainfall (May to September) generally approaches the actual value. The LRR model can capture actual rainfall data patterns better than the PCR model. However, the PCR model can produce forecasts of rainfall data close to the actual value, especially for the January to April period. The LRR model involving 13 PC and a dummy variable (LRR3-dummy) is the best model because it can produce more accurate rainfall estimates.

4. CONCLUSIONS

Statistical downscaling is a model that connects global scale climate variables from global circulation model data with local scale climate variables in Pangkep Regency. Principal component and latent root regression methods are used in statistical downscaling models to overcome multicollinearity problems in precipitation data. The latent root regression method produces 11 principal components with a proportion of variance of 98%. In addition, the first 15 principal components produce eigenvalues greater than 0.1. Thus, statistical downscaling model simulation is carried out based on the number of principal components in the model, namely 11-15 PC. Data modeling using data for the period January 1999 to December 2019 shows that adding a dummy variable can improve the accuracy of the principal component and latent root regression models. In addition, adding a dummy variable results in better predictions of rainfall data for the 2020 period than without a dummy variable. The latent root regression model involving 13 principal components and dummy variables is the best model because it produces the estimated rainfall value with the best accuracy based on the highest correlation value (0.97) and the lowest error (75.17).

ACKNOWLEDGEMENT

We thank the leadership of Hasanuddin University through LP2M (Institute for Research and Community Service) Hasanuddin University as the donor of Unhas internal research. This research received funding through the "Penelitian Dosen Penasehat Akdemik (PDPA) 2022”. We also thank all staff who have provided facilities to complete this research.

REFERENCES

[1] E. Zorita and H. V. Storch, “The analog method as a simple statistical downscaling technique: comparison with more complicated methods,” *J Clim,* vol. 12, pp. 2474-2489, 1999.

[2] A. H. Wigena, “Pemodelan statistical downscaling dengan regresi projection persuit untuk peramalan curah hujan bulanan (Ph.D. dissertation),” Bogor Agricultural University, Bogor, ID, Indonesia, 2006.

[3] D. C. Montgomery and E. A. Peck, Introduction To Linear Regression Analysis, Second Edition, New York: John Willey and Sons Inc,, 1992.

[4] Sutikno, Setiawan and H. Purnomoadi, “Statistical Downscaling Output GCM Modeling with Continuum Regression and Pre-Processing PCA Approach,” *IPTEK, The Journal for Technology and Science,* vol. 21, no. 3, 2010.

[5] E. Vigneau and E. M. Qannari, “A New Algorithm for Latent Root Regression Analysis,” *Computational Statistics & Data Analysis,* pp. 231-242, 2002.

[6] W. Estiningtyas and A. H. Wigena, “Teknik statistical downscaling dengan regresi komponen utama dan regresi kuadrat terkecil parsial untuk prediksi curah hujan pada kondisi el nino, la nina, dan normal,” *Jurnal Meteorologi dan Geofisika,* vol. 12, no. 1, pp. 65-72, 2011.

[7] W. J. Sari, “Pemodelan Statistical Downscaling dengan Regresi Kuantil Komponen Utama Fungsional Untuk Prediksi Curah Hujan Ektrim (thesis),” Bogor Agricultural University, Bogor, ID, Indonesia, 2015.

[8] S. Sahriman, Anisa and V. Koerniawan, “Pemodelan Statistical Downscaling dengan Peubah Dummy Berdasarkan Teknik Cluster Hierarki dan Nonhierarki untuk Pendugaan Curah Hujan,” *Indonesian Journal of Statistics and Its Applications,* vol. 3, no. 3, pp. 295-309, 2019.

[9] E. Purwanto, N. Herrhyanto and M. Suherman, “Penggunaan Regresi Akar Laten Untuk Memprediksi Penjualan Mobil di Amerika Serikat Tahun 1961-1990,” *Eureka Matika,* vol. 2, no. 1, pp. 34-42, 2014.

[10] D. L. Riyantini, M. Susilawati and K. Sari, “Penerapan Regresi Akar Laten Dalam Menangani Multikolinearitas Pada Model Regresi Linier Berganda,” *Jurnal Matematika,* vol. 3, no. 1, pp. 8-16, 2014.

[11] D. P. Untari and M. Susanti, “Latent Root Regression Untuk Mengatasi Multikolinearitas,” *Jurnal Pendidikan Matematika,* vol. 12, no. 2, pp. 23-32, 2017.

[12] R. Y. Sari, H. Oktavianto and H. W. Sulistyo, “Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia,” *Jurnal Aplikasi Sistem Informasi dan Elektronika,* vol. 6, no. 48, pp. 65-86, 2021.

[13] Gujarati and N. Damodar, Basic Econometrics, New York: McGraw-Hill, 1995.

[14] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, Six Edition, New Jersey (NJ) : Pearson Prentice Hall, 2007.

[15] N. R. Draper and H. Smith, Applied Regression Analysis, Second Edition, New York:: John Wiley & Sons, 1981.