Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 17(1) 2024: 1-12
http://ejournal.undip.ac.id/index.php/media_statistika

STACKING ENSEMBLE APPROACH IN STATISTICAL DOWNSCALING USING CMIP6-DCPP FOR RAINFALL ESTIMATION IN RIAU

Dani Al Mahkya[1], Anik Djuraidah[2], Aji Hamim Wigena[2], Bagus Sartono[2]
[1] Actuarial Science Study Program, Institut Teknologi Sumatera, Lampung, Indonesia
[2] Department of Statistics, IPB University, Bogor, Indonesia

e-mail: *dani.almahkya@at.itera.ac.id*

DOI: 10.14710/medstat.17.1.1-12

Article Info:
Received: 3 March 2024
Accepted: 16 August 2024
Available Online: 22 September 2024

Keywords:
Stacking Ensemble; Statistical Downscaling; Rainfall; Regression; Principal Component; LASSO

Abstract: Rainfall modeling and prediction is one of the important things to do. Rainfall has an important relationship and role with various aspects of the environment. One phenomenon that can be associated with rainfall is forest and land fires. Riau is one of the provinces in Indonesia that has a high potential for forest and land fires. This is because Riau has a large area of peatland. One approach that can be used to estimate rainfall is statistical downscaling. The concept of this approach is to form a functional relationship between global and local data. This research uses CMIP6-DCPP output data that will be used to estimate rainfall at 10 observation stations in Riau. The proposed model in this research is Stacking Ensemble with PC Regression and LASSO Regression in the base model and Multiple Linear Regression in the meta model. This research aims to determine the best CMIP6-DCPP model for estimating rainfall in Riau and increasing the accuracy of rainfall estimates using the Stacking Ensemble approach.

1. INTRODUCTION

Rainfall is a measure of the height of rainwater collected in a rain gauge in a flat, non-absorbent, non-permeable, non-flowing place in millimeters (BMKG, 2023). Rainfall is an important factor related to various aspects of the environment. Forest and land fires are closely related to rainfall conditions in an area (Yusuf et al., 2019). Based on the location of the incident, almost 75 percent of forest and land fires are located on peatlands (Yunianto, 2021). One of the provinces in Indonesia that has a large amount of peatland is Riau, which is around 5.09 million hectares or 56.42% of the total peatland in Sumatra (DLHK, 2018). Riau is one of the provinces that has a high potential for forest and land fires in Indonesia. For these reasons, rainfall estimation is urgently needed, especially in the Riau region. However, in reality, many complex factors affect rainfall in a region. This makes rainfall estimation difficult. One strategy that can be used is the statistical downscaling approach.

Statistical downscaling is a technique that can be used to model rainfall in a region based on GCM (General Circulation Model) output information. This process is done by forming a statistical model that states the functional relationship between GCM output data as a predictor variable and rainfall at a point or region as a response variable (Wigena & Djuraidah, 2022). GCMs are numerical models that represent the atmosphere, oceans, cryosphere, and land surface physics used to simulate global climate variables (IPCC, 2023). One provider of GCM output data is CMIP6 (Coupled Model Intercomparison Project Phase

6) with one of the contributing models being the Decadal Climate Prediction Project (DCPP). This model allows for the coordination of climate predictions, predictability, and variability over decades (Boer et al., 2016). The CMIP6-DCPP model has been widely used in scientific research on climate characteristics in the world. Based on the latest information, CMIP6-DCPP has 17 models and 89 experiments (PCMDI, 2024). The CMIP6-DCPP model shows a good ability to predict extreme rainfall in Iran (Asadi-Rahim et al., 2023). Another development used for summer monsoon rainfall prediction (Monerie et al., 2023) and the extreme summer rainfall in India (Konda et al., 2023). In addition to rainfall, the CMIP6-DCPP model has also been developed to predict surface air temperature in the Eurasian region (Huang et al., 2023). This research will use 10 CMIP6-DCPP model output data that will be used in statistical downscaling procedures with a minimum domain size covering the Riau region.

In the statistical downscaling approach, multicollinearity between grids in a domain can occur (Sahriman & Yulianti, 2023). For this reason, the statistical models that will be used in this research are Principal Component Regression (PC) and LASSO Regression. Both models can be used to handle multicollinearity. This research will also propose the use of a stacking ensemble approach to improve the performance and accuracy (Wolpert, 1992) of statistical downscaling in rainfall estimation. In regression models, this approach can also be called stacked regression (Breiman, 1996) with the concept of building a linear combination of multiple predictors to improve accuracy. Meanwhile, the criteria for model goodness used in this research are coefficient of determination (R^2) and Root Mean Square Error (RMSE). Based on the explanation that has been described, this research aims to determine the best CMIP6-DCPP for rainfall estimation in Riau and improve the accuracy of the rainfall estimation model with statistical downscaling using a stacking ensemble.

2. LITERATURE REVIEW

2.1. Statistical Downscaling

GCM output data that has low resolution ranging from $\pm 2.5^0$ or ± 300 km^2 requires an approach that can be used for rainfall estimation. One approach that can be used is Statistical Downscaling. The general form of statistical downscaling is as follows (Wigena & Djuraidah, 2022):

$$\boldsymbol{y}_{t\times 1} = f(\boldsymbol{X}_{t\times \mathrm{g}}) \tag{1}$$

where $\boldsymbol{y}_{t\times 1}$ is the local climate variable and $\boldsymbol{X}_{t\times \mathrm{g}}$ is the GCM output data. The notation t is the length of the time period and g is the number of grids in the GCM domain. One application of statistical downscaling is to estimate rainfall. This estimation can be done using precipitation data from GCM output and rainfall data at a point or region.

2.2. Stacking Ensemble

The stacking algorithm proposed by Wolpert (1992) has the concept of combining several models in the training process. This concept aims to improve the accuracy of a model in estimating. The stacking algorithm has been widely applied in various studies in various aspects (Berliana & Bustamam, 2020; Jayapermana et al., 2022; Sunarko et al., 2023; Zhang et al., 2022). Specifically regarding the regression approach, there is an approach proposed by Breiman in 1996 called stacking regression (Breiman, 1996). This approach forms a linear combination of various predictors to improve prediction accuracy. In modeling with the stacking algorithm, there are at least two stages that must be met, namely the level 0 stage and the level 1 stage (Ghasemieh et al., 2023). At level 0, the training data will be modeled

in such a way as to obtain a predicted value and this stage is called the base model. The predicted value will be stored as a new data set that is used as a predictor variable for the meta model at level 1. The stacking algorithm used in this research is described in Figure 1.

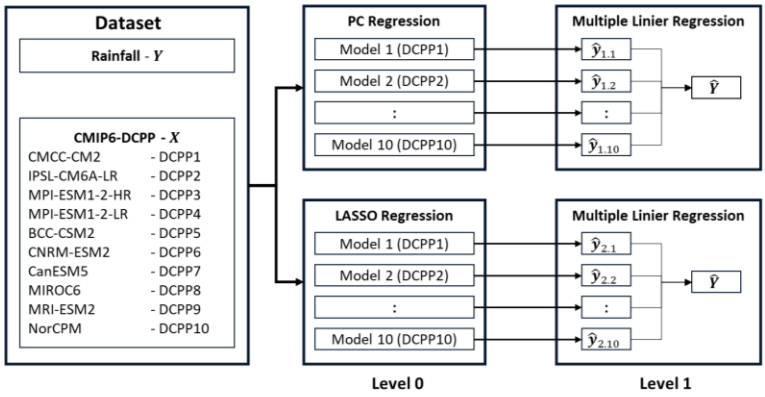

Figure 1. Schematic of the Stacking Ensemble in this Research

This research will focus on statistical downscaling modeling based on 10 CMIP6-DCPP model outputs. Based on Figure 1, each CMIP6-DCPP output (minimum domain) will be used to model rainfall at each station using PC regression and LASSO regression. This stage is performed at level 0 or base model. The results of rainfall modeling in the base model will produce rainfall estimates at each station based on each CMIP6-DCPP. The results of the estimated rainfall will be used in the level 1 or meta model stage by forming a multiple linear regression model to obtain the final estimate of rainfall.

3. METHODOLOGY

3.1. Data

In statistical downscaling, there are two types of data used: global data and local data. The global data used in this research are 10 CMIP6-DCPP model output data. The complete list and characteristics of the CMIP6-DCPP models used can be seen in Table 1. The domain size used for each CMIP6-DCPP model is the minimum domain size that only covers Riau province. The CMIP6-DCPP data used was obtained from the website https://esgf.llnl.gov/nodes.html.

Table 1. List of CMIP6-DCPP Models

Model	Institution	Minimum domain	Notation
CMCC-CM2	Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy	4 × 5	DCPP1
IPSL-CM6A-LR	Institut Pierre Simon Laplace, France	3 × 4	DCPP2
MPI-ESM1-2-HR	Deutsches Klimarechenzentrum, Germany	6 × 6	DCPP3
MPI-ESM1-2-LR	Deutsches Klimarechenzentrum, Germany	4 × 3	DCPP4
BCC-CSM2	Beijing Climate Center, China	5 × 5	DCPP5
CNRM-ESM2	Centre National de Recherches Meteorologiques, France	4 × 4	DCPP6
CanESM5	Canadian Center for Climate Modeling and Analysis, Canada	3 × 3	DCPP7
MIROC6	Japanese modeling community, Japan	4 × 5	DCPP8
MRI-ESM2	Meteorological Research Institute, Japan	5 × 5	DCPP9
NorCPM	NorESM Climate Modeling Consortium	4 × 3	DCPP10

Meanwhile, the local data used is rainfall observation data in Riau obtained directly from the Meteorology, Climatology and Geophysics Agency (BMKG). There are 10 rainfall observation stations that are discussed in this study. The location of the rainfall observation stations is shown in Figure 2 with coordinate details in Table 2. The time series plot of rainfall data at 10 observation stations is shown in Figure 3.

The global and local data that have been described will be used in the statistical downscaling using several models. The length of time period in this study is 2010-2020 with monthly data. The data set will be divided into training data and testing data. The training data uses the period from 2010-2017 and the testing data uses the period from 2018-2020.

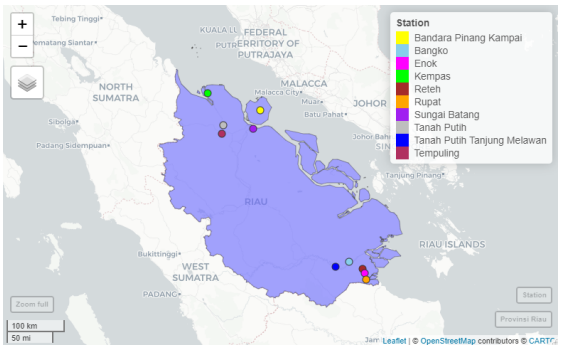

Figure 1. Rainfall Observation Station Location

Figure 2. Time Series Plot of Rainfall at 10 Observation Stations

Table 2. List of Rainfall Observation Stations

Station	Latitude	Longitude
Bangko	2.17000	100.80010
Rupat	1.91127	101.61510
Tanah Putih Tanjung Melawan	1.67861	101.04540
Bandara Pinang Kampai	1.63000	101.49900
Tanah Putih	1.54820	101.02470
Tempuling	-0.43310	102.98330
Kempas	-0.50262	102.77690
Enok	-0.53330	103.20000
Sungai Batang	-0.60561	103.22260
Reteh	-0.69967	103.24460

3.2. Modeling Procedures

This research will begin with a statistical downscaling process based on 10 CMIP6-DCPP output data to estimate rainfall at each selected location. This process is carried out on the base model of the stacking ensemble approach. After obtaining the estimated rainfall of each location with each CMIP6-DCPP, the next step is to form a multiple linear regression model between the estimated rainfall based on each CMIP6-DCPP output data with rainfall observation data. This process is carried out at the meta model stage at each location. The estimation results obtained from each CMIP6-DCPP for each location will be evaluated for model performance and goodness. The process of collecting, preparing and analyzing data in this study used Ms. Excel and R software tools.

The first model to be used in the base model is PC Regression. This model can be used to handle multicollinearity in the statistical downscaling process caused by high correlation between grids in a domain. If the multiple linear regression model equation is defined in equation (2) where $\boldsymbol{y}$ is the response variable, $\boldsymbol{X}$ is the predictor variables, $\boldsymbol{\varepsilon}$ is the error model and $\widehat{\boldsymbol{\beta}}$ is called the least square estimator (Rencher & Schaalje, 2007):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \quad (3)$$

Then the general form of PC Regression is as follows (Jolliffe, 2002):

$$\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \quad (4)$$
$$\boldsymbol{Z} = \boldsymbol{XA} \quad (5)$$

where the (i,k)th element of $\boldsymbol{Z}$ is the value of the kth principal component in the ith observation and $\boldsymbol{A}$ is a matrix of size $p \times p$ with each column being the kth eigenvector of the matrix $\boldsymbol{X'X}$. Based on the form in equation (5), the $\boldsymbol{X\beta}$ component in equation (2) can be rewritten as $\boldsymbol{XAA'\beta} = \boldsymbol{Z\gamma}$ because $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\gamma} = \boldsymbol{A'\beta}$. The estimation of the parameter $\boldsymbol{\gamma}$ in equation (4) will take the following form:

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y} \quad (6)$$

The PC Regression approach has been widely applied in statistical downscaling (Loganathan & Mahindrakar, 2021; Sahriman et al., 2014; Sahriman & Yulianti, 2023), as well as other modeling related to handling multicollinearity (Susilawati & Didiharyono, 2023). The application of PC Regression in statistical downscaling is based on the presence of multicollinearity in the predictor variables, which in this case is a high correlation between grids in a domain (Sahriman et al., 2014).

The next model used in this research is the Least Absolute Shrinkage and Selection Operator (LASSO) Regression. This model is one of the popular methods used to simplify regression models. The concept of simplifying the model is by shrinking some regression coefficients and converting other coefficients to zero (Tibshirani, 1996). This shrinkage is expected to handle the multicollinearity problem in statistical downscaling. Similar to the purpose of using PC Regression, the use of LASSO Regression in this research aims to overcome multicollinearity in the statistical downscaling process. The basic concept of LASSO is to minimize the penalized least squares form (He et al., 2019):

$$\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad (7)$$

where $\lambda > 0$ is the penalty parameter in LASSO regression. The notation $\|\cdot\|_2^2$ means the sum of squared values and the notation $\|\cdot\|_1$ means the sum of absolute values. Equation (7) can be rewritten as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad (8)$$

with $\|\boldsymbol{\beta}\|_1$ can be translated into $\|\boldsymbol{\beta}\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_j|$. Based on equation (7) or equation (8), the selection of the λ value will affect the resulting coefficient value (Jackson, 2023). Several studies have been conducted regarding the application of LASSO regression in the statistical downscaling of rainfall estimates (He et al., 2019; Santri & Hanike, 2020; Yunus et al., 2020).

Based on several models that have been carried out, the next step is to determine the goodness of each model. The criteria used in this research are R^2 and RMSE. The general form of R^2 and RMSE is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

where Y_i is observation data and $\hat{Y}_i$ is prediction data. Meanwhile, $\bar{y}$ is the mean of observation data, and n is the number of observations used in the model. The R^2 and RMSE values in the base model will be used to determine the best CMIP6-DCPP. Meanwhile, at the meta model stage, it is used to measure the improvement in estimation accuracy.

Both PC Regression and LASSO Regression models use the Cross Validation (CV) approach to determine the optimal model that minimizes RMSE. In PC Regression, CV is used to determine the optimal component based on minimum RMSE and estimation of model parameters using equation (6). In LASSO Regression, CV is used to determine the optimum λ value based on minimum RMSE. The best model obtained with CV will be used in the next stage.

4. RESULTS AND DISCUSSION

4.1. Evaluation of Base Model

The evaluation results of the PC Regression model are shown in Table 3 and Table 4. The selection of the number of components in PC Regression is based on the smallest RMSE value through cross-validation. This method causes the number of components used in each model based on CMIP6-DCPP to be different. Based on the results, a large number of components does not guarantee a small RMSE value. Based on the PC Regression model formed from each CMIP6-DCPP, DCPP6 (CNRM-ESM2) dominates by showing good performance at 6 out of 10 observation stations. Other CMIP6-DCPP models that show good performance are DCPP1 (CMCC-CM2) at Tanah Putih Tanjung Melawan station, DCPP3 (MPI-ESM1-2-HR) at Reteh station, DCPP5 (BCC-CSM2) at Sungai Batang station, and DCPP9 (MRI-ESM2) at Rupat station.

Even though DCPP6 (CNRM-ESM2) dominated at 6 of the 10 observation stations, the highest R^2 value were obtained by DCPP1 (CMCC-CM2) at the Tanah Putih Tanjung Melawan station. Meanwhile, the lowest RMSE value was obtained by DCPP3 (MPI-ESM1-2-HR) at the Reteh station. The accuracy of the estimated results in PC Regression has the possibility of being improved again. One way that can be used is to add dummy variables and utilize the lag-GCM (Sahriman et al., 2014). Some possible dummy variables that could be included are such as the relationship with the el nino and la nina phenomena and the altitude factor of each location.

Table 3. R^2 Value of PC Regression Model

Station	R^2									
	DCPP1	DCPP2	DCPP3	DCPP4	DCPP5	DCPP6	DCPP7	DCPP8	DCPP9	DCPP10
Bangko	0.3648	0.0285	0.2436	0.1077	0.2763	0.4231	0.3142	0.2159	0.0908	0.3041
Rupat	0.0134	0.0055	0.0433	0.0070	0.0090	0.0208	0.0096	0.0079	0.0497	0.0016
Tanah Putih Tanjung Melawan	0.4319	0.0371	0.2171	0.1730	0.1111	0.3022	0.1510	0.0802	0.3032	0.0625
Bandara Pinang Kampai	0.0523	0.0099	0.0668	0.0042	0.1027	0.1629	0.0847	0.0849	0.1311	0.0175
Tanah Putih	0.2258	0.0518	0.2346	0.1982	0.0522	0.2583	0.1660	0.0686	0.1289	0.1554
Tempuling	0.0686	0.1654	0.0834	0.0075	0.0045	0.3424	0.1041	0.0907	0.1244	0.0013
Kempas	0.1659	0.0089	0.2004	0.2268	0.0748	0.3737	0.0388	0.1413	0.2255	0.0045
Enok	0.0869	0.0578	0.1929	0.1700	0.4187	0.4222	0.1916	0.1378	0.1670	0.0364
Sungai Batang	0.0906	0.0112	0.1985	0.0017	0.3241	0.0235	0.1467	0.1129	0.1291	0.1140
Reteh	0.0953	0.0848	0.3382	0.1294	0.0617	0.1891	0.0023	0.3174	0.2304	0.3329

Table 4. RMSE Value of PC Regression Model

Station	RMSE									
	DCPP1	DCPP2	DCPP3	DCPP4	DCPP5	DCPP6	DCPP7	DCPP8	DCPP9	DCPP10
Bangko	80.80	99.92	88.17	95.77	86.24	77.00	83.95	89.77	96.67	84.57
Rupat	145.95	146.54	143.72	146.42	146.28	145.41	146.23	146.36	143.24	146.82
Tanah Putih Tanjung Melawan	62.85	81.82	73.77	75.82	78.61	69.65	76.83	79.97	69.60	80.73
Bandara Pinang Kampai	84.74	86.61	84.08	86.86	82.45	79.64	83.27	83.26	81.14	86.28
Tanah Putih	80.42	89.00	79.96	81.84	88.98	78.71	83.46	88.21	85.30	84.00
Tempuling	85.76	81.18	85.08	88.52	88.66	72.06	84.11	84.74	83.15	88.80
Kempas	91.53	99.77	89.61	88.12	96.39	79.31	98.25	92.86	88.19	99.99
Enok	75.54	76.74	71.02	72.03	60.27	60.09	71.08	73.41	72.15	77.60
Sungai Batang	77.67	80.99	72.92	81.38	66.96	80.49	75.24	76.71	76.01	76.67
Reteh	67.69	68.08	57.89	66.40	68.94	64.09	71.09	58.80	62.43	58.13

Table 5. R^2 Value of LASSO Regression Model

Station	R^2									
	DCPP1	DCPP2	DCPP3	DCPP4	DCPP5	DCPP6	DCPP7	DCPP8	DCPP9	DCPP10
Bangko	0.3543	0.0752	0.1901	0.0438	0.1867	0.4141	0.2479	0.1514	0.1433	0.3104
Rupat	0.0359	0.0029	0.0979	0.0122	0.1023	0.0376	0.0344	0.0149	0.0291	0.0119
Tanah Putih Tanjung Melawan	0.1949	0.0594	0.2025	0.1266	0.1331	0.2730	0.1529	0.1698	0.2337	0.1411
Bandara Pinang Kampai	0.0895	0.0218	0.0884	0.0013	0.0793	0.2019	0.0696	0.0897	0.2135	0.0041
Tanah Putih	0.1976	0.1564	0.2245	0.1523	0.0794	0.1099	0.1369	0.1042	0.1297	0.0602
Tempuling	0.0818	0.0969	0.0705	0.0056	0.0216	0.3420	0.0861	0.0983	0.1039	0.0041
Kempas	0.1952	0.0334	0.2195	0.2214	0.1578	0.3068	0.0511	0.1711	0.2503	0.0567
Enok	0.1446	0.0375	0.3201	0.1649	0.3628	0.4237	0.1920	0.1388	0.1920	0.0568
Sungai Batang	0.0763	0.0441	0.2022	0.0740	0.1770	0.0266	0.1279	0.1009	0.1280	0.0913
Reteh	0.0706	0.1334	0.2802	0.1399	0.0553	0.2041	0.0041	0.3730	0.1586	0.2578

Table 6. RMSE Value of LASSO Regression Model

Station	RMSE									
	DCPP1	DCPP2	DCPP3	DCPP4	DCPP5	DCPP6	DCPP7	DCPP8	DCPP9	DCPP10
Bangko	81.46	97.49	91.23	99.13	91.42	77.60	87.92	93.39	93.84	84.19
Rupat	144.28	146.73	139.56	146.04	139.22	144.15	144.39	145.84	144.79	146.06
Tanah Putih Tanjung Melawan	74.81	80.86	74.46	77.92	77.63	71.09	76.74	75.97	72.99	77.27
Bandara Pinang Kampai	83.06	86.09	83.10	86.98	83.52	77.76	83.96	83.05	77.19	86.87
Tanah Putih	81.87	83.94	80.49	84.15	87.69	86.23	84.91	86.50	85.26	88.60
Tempuling	85.15	84.44	85.67	88.61	87.89	72.08	84.95	84.38	84.12	88.68
Kempas	89.90	98.53	88.54	88.43	91.97	83.44	97.62	91.24	86.77	97.33
Enok	73.12	77.56	65.19	72.25	63.11	60.02	71.06	73.36	71.06	76.78
Sungai Batang	78.28	79.63	72.75	78.38	73.89	80.36	76.06	77.23	76.06	77.64
Reteh	68.61	66.25	60.38	66.00	69.17	63.49	71.02	56.35	65.28	61.31

The next discussion is the result of LASSO Regression. The selection process of λ in LASSO Regression uses cross-validation based on the smallest Mean Square Error (MSE) value. This process will cause the shrinkage of variables to be different for each CMIP6-DCPP model. The results shown by the LASSO Regression model in Table 5 and Table 6 show that DCPP6 (CNRM-ESM2) dominates in 5 out of 10 observation stations. Other CMIP6-DCPP models that show good performance are DCPP3 (MPI-ESM1-2-HR) at Tanah Putih and Sungai Batang stations, DCPP5 (BCC-CSM2) at Rupat station, DCPP8 (MIROC6) at Reteh station, and DCPP9 (MRI-ESM2) at Bandara Pinang Kampai station. The highest R^2 value based on LASSO Regression results was obtained based on DCPP6 (CNRM-

ESM2) at Enok station. While the smallest RMSE value is obtained at DCPP8 (MIROC6) with the Reteh station location. Based on the results of PC Regression and LASSO Regression, the R^2 value in the results shows how well a regression model predicts and explains the results of the observed data. While the RMSE explains how much the predicted results differ from the results of the observed data.

4.2. Evaluation of Meta Model

The evaluation results at the meta model stage are shown in Table 7. Based on these results, the meta model formed by stacked multiple linear regression performed very well and improved over the base model results. The results show an increase in R^2 and a decrease the RMSE value at all stations. These results confirm that stacking ensemble is one approach that can be used to improve prediction accuracy (Breiman, 1996; Wolpert, 1992).

Table 7. Comparison of R^2 and RMSE between Base Model and Meta Model

No	Station	R^2				RMSE			
		Base Model (Optimum)		Meta Model		Base Model (Optimum)		Meta Model	
		PCA Reg.	Lasso Reg.	PCA Reg.	Lasso Reg.	PCA Reg.	Lasso Reg.	PCA Reg.	Lasso Reg.
1	Bangko	0.4231	0.4141	0.5988	0.5884	77.00	77.60	64.22	65.04
2	Rupat	0.0497	0.1023	0.0775	0.2008	143.24	139.22	141.13	131.36
3	Tanah Putih Tanjung Melawan	0.4319	0.2730	0.5637	0.4348	62.85	71.09	55.07	62.69
4	Bandara Pinang Kampai	0.1629	0.2135	0.2635	0.3904	79.64	77.19	74.70	67.96
5	Tanah Putih	0.2583	0.2245	0.4192	0.4109	78.71	80.49	69.65	70.15
6	Tempuling	0.3424	0.3420	0.4716	0.4828	72.06	72.08	64.59	63.90
7	Kempas	0.3737	0.3068	0.4983	0.4900	79.31	83.44	70.98	71.56
8	Enok	0.4222	0.4237	0.5829	0.5643	60.09	60.02	51.06	52.18
9	Sungai Batang	0.3241	0.2022	0.4857	0.4183	66.96	72.75	58.41	62.12
10	Reteh	0.3382	0.3730	0.5565	0.5441	57.89	56.35	47.39	48.05

Table 8. Coefficient of Multiple Linear Regression in Meta Model

Station	Coefficients										
	Intercept	DCPP1	DCPP2	DCPP3	DCPP4	DCPP5	DCPP6	DCPP7	DCPP8	DCPP9	DCPP10
Bangko	10.0564	0.3274	-0.7900	0.0610	0.1417	0.2807	0.3877	0.3249	0.0208	-0.3107	0.4990
Rupat	12.6110	-0.1504	0.2986	0.6197	-0.2361	0.7696	0.4842	-0.6126	-0.3180	0.9133	-0.8343
Tanah Putih Tanjung Melawan	0.0935	0.7007	-0.2779	0.1552	0.3065	-0.1087	0.4259	-0.1238	-0.2098	0.3961	-0.2648
Bandara Pinang Kampai	-496.4693	0.0558	0.4914	-0.2872	1.1298	0.3528	0.8185	0.2436	0.6181	0.2000	0.0642
Tanah Putih	-172.1634	0.2768	-0.0340	0.3255	0.4541	-0.0228	0.4475	0.2680	0.1080	0.0227	0.1471
Tempuling	-699.9000	0.0012	0.5361	0.4645	0.6447	-1.4650	0.8194	-0.3295	0.9272	0.0858	3.8960
Kempas	-3.3795	0.3736	-0.4206	0.3185	0.4524	-0.0514	0.6106	0.1008	0.3814	-0.0230	-0.7232
Enok	-72.9957	0.1942	-0.1325	0.3233	0.2714	0.6072	0.5222	0.2003	0.4447	-0.5535	-0.4609
Sungai Batang	-20.9558	0.1394	-0.8697	0.3102	-0.0383	0.8320	-0.4557	0.7016	0.2494	-0.1365	0.4283
Reteh	-235.0429	0.2447	0.3084	0.3959	0.1588	-0.1157	-0.2543	1.0370	0.5018	0.3306	0.5107

Based on Table 7, rainfall estimation with stacking ensemble using PC Regression shows better performance than LASSO Regression. This is shown by the performance of PC Regression which dominates 7 out of 10 locations compared to LASSO Regression. Multiple linear regression coefficients on the meta model using PC Regression are shown in Table 8. Based on this result, the stacking ensemble approach with PC Regression (base model) and Multiple Linier Regression (meta model) will be used in the testing data evaluation stage.

4.3. Stacking Ensemble in Testing Data

Based on the training data, the stacking ensemble is proven to improve the performance of the estimation model. But the next question is, how does this model perform on the testing data for each location. Therefore, the R^2 of the training data will be compared

with the R^2 of the testing data and the RMSE of the training data will be compared with the RMSE of the testing data. Performance evaluation results will be obtained based on this comparison. Figure 4 shows the comparison results of modeling results on training data and testing data based on R^2 and RMSE criteria for each location. Red dots represent performance on training data and blue dots represent performance on testing data.

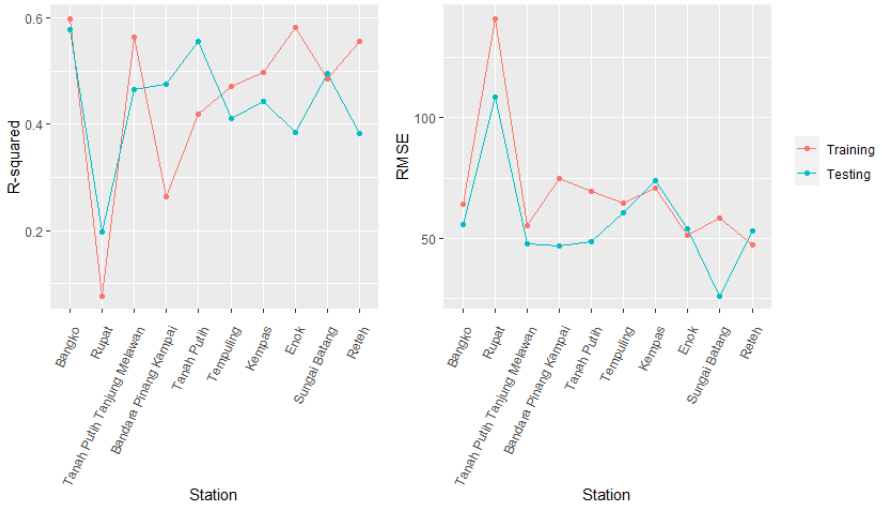

Figure 3. Comparison of R^2 and RMSE in Meta Model between Training and Testing Data

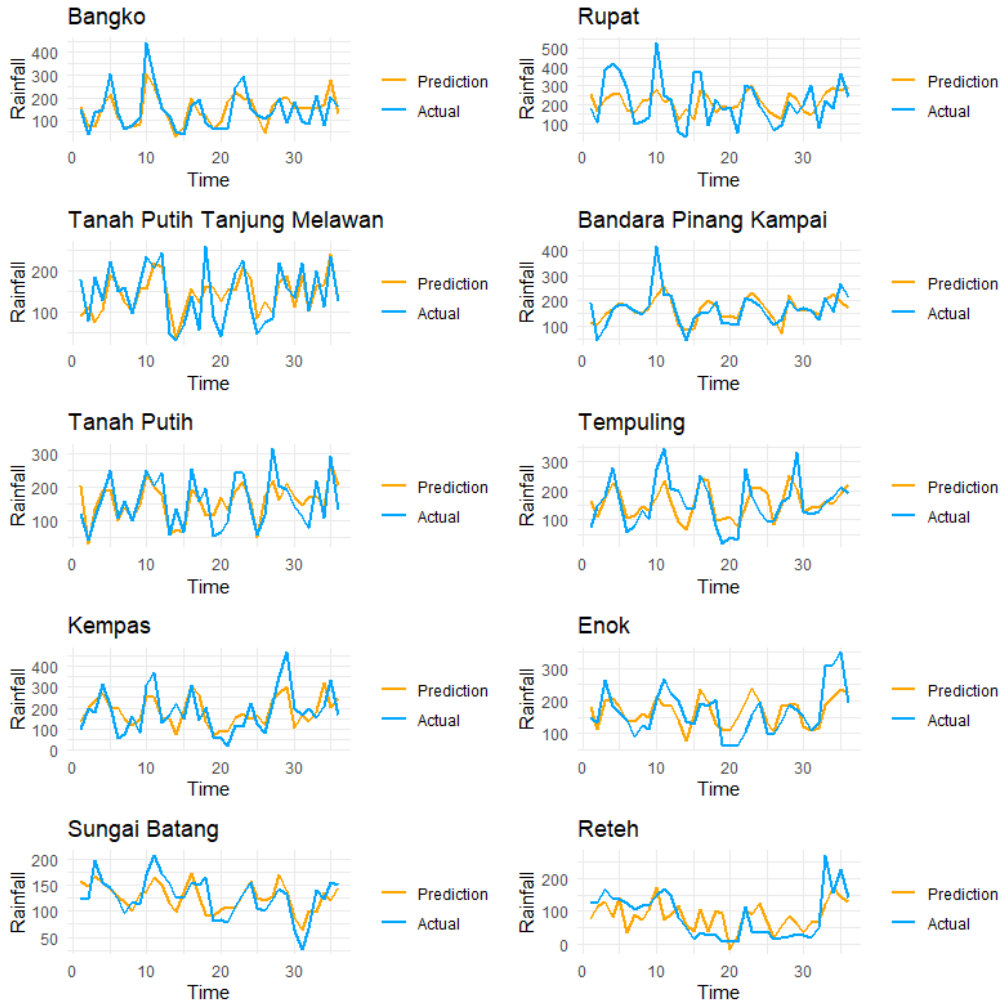

Figure 4. Comparison of Actual and Predicted Rainfall

Based on the results, the locations that have increased R^2 values in the testing data are Rupat, Bandara Pinang Kampai, Tanah Putih, and Sungai Batang. This means that in all four locations, the proposed model has been able to explain the rainfall value in terms of prediction. However, when viewed from the comparison of RMSE in training and testing data, there are three locations where the RMSE value increases in the testing data. These locations are Kempas, Enok, and Reteh. This indicates that the three locations have quite different values between observed and predicted rainfall. The comparison of actual rainfall and predicted rainfall for 10 observation stations is shown in Figure 5.

Several other studies on statistical downscaling with a stacking ensemble approach use many prediction models in the base model (Gu et al., 2022; Zhang et al., 2024). Furthermore, the results of several prediction models are used in the meta model process. However, in this study, the base model building component comes from several GCM outputs. The results obtained will be used in the meta model process. Although the base model building components are different, the results obtained in this study and other studies are in line. The stacking ensemble approach provides improved performance in statistical downscaling.

5. CONCLUSION

Based on the results and discussion, several conclusions can be drawn from the modeling performed. At stacking ensemble level 0 or base model, PC Regression and LASSO Regression show that DCPP6 (CNRM-ESM2) dominates at several observation stations based on both R^2 and RMSE values. In the stacking ensemble level 1 or meta model based on multiple regression models, PC Regression and LASSO Regression showed an increase in the accuracy of estimates compared to the base model or statistical downscaling modeling with each CMIP6-DCPP. This is indicated by the increase in R^2 and the decrease in RMSE in the meta model. Meanwhile, when comparing the results between meta models, PC Regression shows better results than LASSO Regression. This is shown by the dominance of PC Regression in 7 out of 10 locations. At the modeling stage with testing data, there are 4 locations that show an increase in the R^2 value in the testing data. Meanwhile, based on RMSE, there are 6 locations that show a decrease in the RMSE value in the testing data.

ACKNOWLEDGMENT

This research was supported by Research Funds from the ITERA Scholarship Program.

REFERENCES

Asadi-Rahim, B. N., Zarrin, A., Mofidi, A., & Dadashi-Roudbari, A. A. (2023). The Prediction of the Precipitation Extremes over Iran for the Next Decade (2021-2028) using eht Decadal Climate Prediction Project contribution to CMIP6 (CMIP6-DCPP). *Journal of the Earth and Space Physics*, *49*(3), 707–725. https://doi.org/10.22059/jesphys.2023.351678.1007474

Berliana, A. U., & Bustamam, A. (2020). Implementation of Stacking Ensemble Learning for Classification of COVID-19 using Image Dataset CT Scan and Lung X-Ray. *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, 148–152. https://doi.org/10.1109/ICOIACT50329.2020.9332112

BMKG. (2023). *Daftar Istilah Klimatologi*. Balai Besar MKG Wilayah III.

https://balai3.denpasar.bmkg.go.id/daftar-istilah-musim#

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., & Eade, R. (2016). The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/bf00117832

DLHK, R. (2018). *Rapat Koordinasi TIM Restorasi Gambut Daerah Prov. RIAU Tahun 2018*. Dinas Lingkungan Hidup dan Kehutanan Provinsi Riau. https://dislhk.riau.go.id/lihat_berita.php?id_berita=6

Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A Novel Machine Learning Model with Stacking Ensemble Learner for Predicting Emergency Readmission of Heart-Disease Patients. *Decision Analytics Journal*, *7*(May). https://doi.org/10.1016/j.dajour.2023.100242

Gu, J., Liu, S., Zhou, Z., Chalov, S. R., & Zhuang, Q. (2022). A Stacking Ensemble Learning Model for Monthly Rainfall Prediction in the Taihu Basin, China. *Water (Switzerland)*. https://doi.org/10.3390/w14030492

He, R. R., Chen, Y., Huang, Q., & Kang, Y. (2019). LASSO as a Tool for Downscaling Summer Rainfall over the Yangtze River Valley. *Hydrological Sciences Journal*, *64*(1), 92–104. https://doi.org/10.1080/02626667.2019.1570210

Huang, Y., Huang, N., & Zhao, Q. (2023). Decadal Prediction Skill for Eurasian Surface Air Temperature in CMIP6 models. *Atmospheric and Oceanic Science Letters*, *17*(May 2023), 0–4. https://doi.org/10.1016/j.aosl.2023.100377

IPCC. (2023). *What is a GCM?* IPCC Data Distribution Centre. https://www.ipcc-data.org/guidelines/pages/gcm_guide.html

Jackson, S. (2023). *Machine Learning*. Bookdown Write HTML, PDF, EPub, and Kindle Books with R Markdown. https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/

Jayapermana, R., Aradea, A., & Kurniati, N. I. (2022). Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topics on Twitter. *Scientific Journal of Informatics*, *9*(1), 8–15. https://doi.org/10.15294/sji.v9i1.31648

Jolliffe, I. T. (2002). *Principal Component Analysis* (Second Edi). Springer Series in Statistics.

Konda, G., Chowdary, J. S., Gnanaseelan, C., & Parekh, A. (2023). Improvement in the Skill of CMIP6 Decadal Hindcasts for Extreme Rainfall Events over the Indian Summer Monsoon Region. *Scientific Reports*, *13*(1), 1–13. https://doi.org/10.1038/s41598-023-48268-1

Loganathan, P., & Mahindrakar, A. B. (2021). Statistical downscaling using principal component regression for climate change impact assessment at the Cauvery river basin. *Journal of Water and Climate Change*, *12*(6), 2314–2324. https://doi.org/10.2166/wcc.2021.223

Monerie, P. A., Robson, J. I., Ndiaye, C. D., Song, C., & Turner, A. G. (2023). CMIP6 Skill at Predicting Interannual to Multi-Decadal Summer Monsoon Precipitation

Variability. *Environmental Research Letters*, *18*(9). https://doi.org/10.1088/1748-9326/acea96

PCMDI (2024). *ESGF CMIP6 DCPP Data Holdings*. DCPP Model List. https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/DCPP/index.html

Rencher, A. C., & Schaalje, G. B. (2007). Linear Models in Statistics. In *Linear Models in Statistics*. https://doi.org/10.1002/9780470192610

Sahriman, S., Djuraidah, A., & Wigena, A. H. (2014). Application of Principal Component Regression with Dummy Variable in Statistical Downscaling to Forecast Rainfall. *Open Journal of Statistics*, *04*(09), 678–686. https://doi.org/10.4236/ojs.2014.49063

Sahriman, S., & Yulianti, A. S. (2023). Statistical Downscaling Model With Principal Component Regression and Latent Root Regression To Forecast Rainfall in Pangkep Regency. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(1), 0401–0410. https://doi.org/10.30598/barekengvol17iss1pp0401-0410

Santri, D., & Hanike, Y. (2020). Pemodelan Statistical Downscaling dengan Regresi Kuantil Komponen Utama Fungsional untuk Prediksi Curah Hujan Ekstrim. *MAP Journal*, *2*(1), 47–57.

Sunarko, B., Hasanah, U., Hidayat, S., Muhammad, N., Ardiansyah, M. I., Hakiki, M. K., Ananda, B. P., & Baroroh, T. (2023). Penerapan Stacking Ensemble Learning untuk Klasifikasi Efek Kesehatan Akibat. *Edu Komputika Journal*, *10*(1), 55–63. https://doi.org/https://doi.org/10.15294/edukomputika.v10i1.72080

Susilawati, S., & Didiharyono, D. (2023). Application of Principal Component Regression in Analyzing Factors Affecting Human Development Index. *Jurnal Varian*, *6*(2), 199–208. https://doi.org/10.30812/varian.v6i2.2366

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Wigena, A. H., & Djuraidah, A. (2022). *Monograph Pengembangan Statistical Downscaling untuk Peningkatan Akurasi Prediksi Curah Hujan*. IPB Press.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yunianto, A. S. (2021). Pemetaan Permasalahan Kebakaran Hutan dan Lahan Kasus di Provinsi Riau. *Jurnal Penelitian Kehutanan Sumatrana*, *2*(1), 25–37. https://doi.org/10.20886/jpks.2021.2.1.25-37

Yunus, M., Saefuddin, A., & Soleh, A. M. (2020). *Pemodelan Statistical Downscaling dengan LASSO dan Group LASSO untuk Pendugaan Curah Hujan*. 649–660.

Yusuf, A., Hapsoh, H., Siregar, S. H., & Nurrochmat, D. R. (2019). Analisis Kebakaran Hutan dan Lahan di Provinsi Riau. *Dinamika Lingkungan Indonesia*, *6*(2), 67. https://doi.org/10.31258/dli.6.2.p.67-84

Zhang, Y., Li, J., & Liu, D. (2024). Spatial Downscaling of ERA5 Reanalysis Air Temperature Data Based on Stacking Ensemble Learning. *Sustainability (Switzerland)* . https://doi.org/10.3390/su16051934

Zhang, Y., Liu, J., & Shen, W. (2022). A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178654