Statistical downscaling of GCM using kernel support vector regression for rainfall prediction in Bireuen district

Sanusi and J Husna
Information System Department, Abulyatama University, Aceh Besar, 24415, Indonesia
E-mail: sanusi@abulyatama.ac.id

Abstract. Global climate phenomena, including El-Nino and Southern Oscillation, Indian Ocean Dipole and Madden Julian Oscillation, have affected rainfall behavior in Indonesia. This study attempts to model a statistical downscaling of the General Circulation Model (GCM), which is usually used as learning data for the prediction of climate change, using Support Vector Regression (SVR) to predict rainfall during the dry season in Bireuen, Aceh Province. Data consists of observational rainfall data at 7 (seven) weather stations and 10 models of 7x7 grid-scale hindcasts of GCM data collected within 24 years (1990-2013). The RBF kernel was found to produce the best performance (correlation value: 0.828; RMSE: 24.035) compared to the linear (correlation value: 0.538; RMSE: 27.207) and polynomial kernel (correlation value: 0.639; RMSE: 25.584). Among GCM downscaling output, CMC1-CanCM3 was considered to be the best model for rainfall prediction in Bireuen District, particularly in Pteudada and Gandapura regions. This model is recommended to be used as a rainfall forecasting tool during drought and would help in optimizing the arrangement of cropping strategy in Bireun area to prevent possible crop failure.

1. Introduction
Rainfall in Indonesia is strongly influenced by global phenomena such as El-Nino and Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and Madden Julian Oscillation (MJO). ENSO is known as an anomalous sea surface temperature over the tropical eastern Pacific Ocean, while IOD is defined by the difference in sea surface temperature between two areas, the western Indian Ocean and the eastern Indian Ocean) leading to the occurrence of drought in Indonesia. On the other hand, MJO is a global phenomenon affecting climate diversity during the inter-seasonal period which can generally affect the climate in Indonesia [1-3]. The climate variability caused by all of these phenomena significantly contributes to rainfall variability in various regions of Indonesia, which later affects farming activities, including rice cropping, especially in Bireun District of Aceh Province.

Bireuen is one of the central areas for rice production in Aceh with 7.063 hectares of rice field area; however, the implementation of this area for the enhancement of rice production has encountered many obstacles, including lack of irrigation facilities and water scarcity [4]. In fact, in 2018, as many 260 ha of rice cultivation area in Bireuen experienced drought, leading to deficiency in rice production. Not only presented in Bireun, but the drought has also hit several other regions in Aceh, including Aceh Besar (179 ha) and Gayo Luwes regency (12 hectares). Of the total 451 hectares rice-field areas suffering from drought in Aceh, eight hectares were categorized as heavy, 39 hectares was medium, and 403 hectares was light drought. One of the strategies is by predicting rainfall in this area during
the drought period. On the other hand, Bireuen is located in the tropical rain belt. It has rugged topography and some coastal regions where the interaction of the sea and land processes occurs, causing difficulties in modelling rainfall prediction.

Several studies have been conducted to manage the effects of climate change using the Statistical Downscaling (SD) with different methods and models. Wigena [5] employed an 8x8 grid-scale SD modelling of the General Circulation Model (GCM) with Pursuit Projection Regression (RPP), while a 5x5 grid domain of GCM using Grid Search Support Vector Regression (SVR) was utilized by Agmalaro [6]. Both studies were attempted to predict monthly rainfall in Indramayu. The literature reviews had pointed out challenges when choosing an SD method, due to non-linear rainfall-data and a curse of dimensionality and multicollinearity of GCM. In addition, the selection of highly correlated grids is not easy, potentially omission of contiguous domains. In this case, the RPP and SVR which are non-linear, nonparametric and data-driven methods are considered as a solution. However, the SVR method has been extensively used and could produce better results if combined with more appropriated-optimization methods. Moreover, it is essential to carry out further investigation of grid-scale adjustment in different topographic areas. Therefore, this study is trying to model statistical downscaling using Support Vector Regression (SVR) and brute force search by utilizing 7x7 grid-scale GCM data to predict rainfall during a drought in Bireuen District, Aceh Province.

The GCM output can be used to predict local climate conditions such as rainfall and temperature [7], which then can be widely utilized to analyze the impacts of climate variability on agriculture and biodiversity [8]. This model has been designed to globally simulate certain time series of climate variables and can be implemented for climate and weather prediction [9,10]. The resolution of the GCM model is found within 2.5° or 300 km² on each layer in the atmosphere [11]. Therefore, to bridge the gap between global and regional output (results), a downscaling technique is needed to reduce the spatial scale of the GCM. It is a technique used to obtain local-scale information from a global model (GCM) [12-14]. The statistical downscaling attempts to link a circulation between global and local-scale variables [15]. The process of statistical downscaling is shown in Figure 1.

Figure 1. Downscaling GCM output

Statistical Downscaling (SD) uses a transfer function that describes a functional relationship between global atmospheric circulation and local climate elements. In general, the SD is formulated as follows [11,15].

\[ Y_{t,p} = f(X_{t,q,s,g}) \]  

(1)

Where;
- \( Y \) = Local climate variable (climate response variable)
- \( X \) = External transformer of General Circulation Model (GCM)
t = Period (e.g: daily, daily, or monthly)
p = The number of Y variables (dimensions of Y)
q = The number of X variables (dimension of X)
s = The number of atmospheric layers (amount of layers in the atmosphere)
g = GCM domain

Support Vector Regression is a method that can overcome overfitting, resulting in better performance. The model generated by Support Vector Regression relies solely on a subset of training data [16]. For instance, if a function is as the following regression lines:

\[ F(x) = w^T \Phi(x) + b \]  (2)

Where \( \Phi(x) \) represents a point in F feature space resulting from mapping \( x \) into the input space, then the coefficients \( w \) and \( b \) are estimated by minimizing the risk function defined in the following equation:

\[
\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{\ell} L_\varepsilon(y_i, f(x_i)) \text{ depends on; } y_i - w\phi(x_i) - b \leq \varepsilon
\]  (3)

\[
w\phi(x_i) + b - y_i \leq \varepsilon, i = 1, 2, 3, ..., \ell
\]  (4)

Where:

\[
L_\varepsilon(y_i, f(x_i)) = \begin{cases} 
|y_i - f(x_i)| - \varepsilon, & |y_i - f(x_i)| \geq \varepsilon \\
0, & \text{for others}
\end{cases}
\]  (5)

The factor \( \| w \|^2 \) is called regularization. Minimizing \( \| w \|^2 \) will highly reduce the function, enabling the function-capacity control. The second factor is an empirical error as measured by the \( \varepsilon \)-insensitive loss function which is required to minimize the norm from \( w \) to get a good generalization for the regression function \( f(x) \). Hence, we need to solve the following optimization problem:

\[
\min \frac{1}{2} \| w \|^2 \text{ depends on; } y_i - w\phi(x_i) - b \leq \varepsilon
\]  (6)

\[
w\phi(x_i) + b - y_i \leq \varepsilon, i = 1, 2, 3, ..., \ell
\]  (7)

Assuming that there is a function \( f(x) \) approximating all points \((x_i, y_i)\), then, using SVR will produce a tube shown in Figure 2.

![Figure 2. \( \varepsilon \)-insensitive loss function](image)

If some points are found out of the range \( f \pm \varepsilon \), a slack variable \( \xi, \xi^* \) should be added to overcome inappropriate limiting factors in the optimization problem. Then, the optimization problem can be solved by the following formula:
\[
\begin{align*}
\min & \frac{1}{2} ||w||^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \quad \text{depends on: } y_i - w^T \phi(x_i) - \xi_i - b \leq \varepsilon, i = 1, 2, 3, \ldots, \ell \quad (8) \\
& w^T \phi(x_i) - \xi_i^* - b \leq \varepsilon, i = 1, 2, 3, \ldots, \ell \text{ and } \xi, \xi^* \geq 0 \quad (9)
\end{align*}
\]

The constant \( C > 0 \) determines the Pareto between the thinness of functions \( f \) and the upper limit of deviation, the value bigger than \( \varepsilon \) will not be tolerated. All deviations greater than \( \varepsilon \) will be subjected to penalties equal to \( C \). Support vector is training data located outside the boundary \( f \) of the decision function. Therefore the number of support vectors decreases when the value of \( \varepsilon \) increases. In the dual formulation, the optimization problems from SVR are as follows:

\[
\begin{align*}
\max & \sum_{\ell=1}^{\ell} y_i (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \quad (10)
\end{align*}
\]

depends on: \( \sum_{\ell=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \) where \( 0 \leq \alpha_i, \alpha_i^* \leq C, \ i = 1, 2, 3, \ldots, \ell \quad (11) \)

\( C \) is determined by the user and \( K(x_i, x_j) \) is a kernel dot product defined as \( K(x_i, x_j) = \phi^T(x_i) \phi^T(x_j) \). By using Lagrange multipliers and optimality conditions, the regression function is explicitly formulated as follows:

\[
(\alpha) = \sum_{\ell=1}^{\ell} (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (12)
\]

There are several kernel SVR functions can be used, such as:

1. The linear kernel function: \( k(x, y) = x^T y + C \)
2. The polynomial kernel function: \( k(x, y) = (\alpha x^T y + C)^d \)
3. The kernel of radial base function (RBF): \( k(x, y) = \exp(-\gamma ||x - y||^2) \)

The following parameters must be determined before SVR training and testing are conducted. These parameters include \( C \) and \( \gamma \) for the Linear Kernel function; \( C, \gamma, d \) and \( \varepsilon \) for the Polynomial Kernel; and parameter \( C, \gamma, \) and \( \varepsilon \) for the RBF kernel.

2. Method

The data used in this study consisted of rainfall data observed at 7 weather stations, including Juli, Peudada, Balance, Gandapura, Jeumpa, Jeunieb, and Peusangan regencies of Bireuen District. The data was obtained from the Meteorology, Climatology and Geophysics Agency (BMKG), Indrapuri, Aceh Besar. Subsequently, the data contains 10 models of GCM hindcast downloaded from the CLIK APCC website http://clik.apcc21.org/. Their longitude and latitude coordinates were between 40 54’- 50 21 ’LU and 960 21 ’- 970 21 ’BT. The output of the GCM and the developer’s country is described in tables 1 and 2.

| No | GCM data   | Country    | Month          | Year   |
|----|------------|------------|----------------|--------|
| 1  | CMC1-CanCM3 | Columbia   | May-June-July  | 1990-2013 |
| 2  | GCMPS T63T21 | Korea      | May-June-July  | 1990-2013 |
| 3  | GDAPS T106L21 | Korea    | May-June-July  | 1990-2013 |
| 4  | GFDL-CM2P1 | Columbia   | May-June-July  | 1990-2013 |
| 5  | CanCM3-AGCM3 | Canada    | May-June-July  | 1990-2013 |
| 6  | NASA-GSFC L34 | USA       | May-June-July  | 1990-2013 |
| 7  | METRI AGCM L17 | Korea   | May-June-July  | 1990-2013 |
| 8  | PNU       | Korea      | May-June-July  | 1990-2013 |
| 9  | MGO       | Columbia   | May-June-July  | 1990-2013 |
| 10 | BCC       | USA        | May-June-July  | 1990-2013 |
Table 2. Rainfall observational stations' location and times series rainfall data

| Y  | Weather Stations | Latitude | Longitude | Month       | Year     |
|----|------------------|----------|-----------|-------------|----------|
| Y1 | Juli             | 5.073    | 96.530    | May-June-July | 1990-2013 |
| Y2 | Peudada          | 5.051    | 96.425    | May-June-July | 1990-2013 |
| Y3 | Peulimbang       | 5.047    | 96.366    | May-June-July | 1990-2013 |
| Y4 | Gandapura        | 5.226    | 96.846    | May-June-July | 1990-2013 |
| Y5 | Jeumpa           | 5.145    | 96.583    | May-June-July | 1990-2013 |
| Y6 | Jeunieb          | 5.064    | 96.338    | May-June-July | 1990-2013 |
| Y7 | Peusangan        | 5.196    | 96.738    | May-June-July | 1990-2013 |

Observational rainfall data are collected in May-June-July (MJJ) from 1990 to 2013 (24 years) from each observational station. The statistical downscaling data of GCM is taken based on coordinates (longitude and latitude) at each weather station, which later used to determine the GCM output grid of a 7x7 matrix. Since the rainfall data of the GCM model are highly correlated to each other, a 95% spatial dimension reduction using Principal Component Analysis (PCA) is required. The training and testing data are divided using an 8-fold cross-validation technique prior to the construction of the model. Three SVR kernel functions, including linear, polynomial and Radial Basis Function (RBF) were used in this study. Taylor diagram is used to analyze the model’s performance of SVR, along with correlation, standard deviation, and Root Mean Square Error (RMSE) values described on it [17].

3. Results and Discussion
The rainfall data obtained through the implementation of the SVR method varies among its three different kernels employed. Of the 3 SVR different kernels (linear, polynomial, and radial basis function (RBF)). The average values of each SVR kernels at the 7 rainfall observational stations are shown in Table 3.

Table 3. Means of Correlation and RMSE values from observational rainfall and GCM output

| Stations | Linear Kernel | Polynomial Kernel | RBF Kernel |
|----------|---------------|------------------|------------|
|          | Correlation   | RMSE             | Correlation| RMSE     | Correlation| RMSE     |
| Juli     | 0.580         | 67.600           | 0.677      | 31.428   | 0.694      | 72.593   |
| Peudada  | 0.468         | 28.942           | 0.496      | 28.284   | 0.745      | 23.043   |
| Peulimbang| 0.538        | 35.483           | 0.487      | 35.713   | 0.713      | 27.419   |
| Gandapura| 0.668         | 40.193           | 0.672      | 21.106   | 0.751      | 31.166   |
| Jeumpa   | 0.629         | 30.801           | 0.594      | 25.811   | 0.655      | 26.165   |
| Jeunieb  | 0.529         | 27.795           | 0.475      | 31.407   | 0.697      | 22.469   |
| Peusangan| 0.413         | 45.566           | 0.642      | 35.521   | 0.709      | 33.696   |

The results presented in Table 3 revealed that the highest correlation value of rainfall prediction was obtained by using the RBF kernel. It occurred at all weather stations in Bereuen District. On the other hand, RMSE values yielded by the 3 SVR models vary in each observed regions. Concerning the implementation of the RBF model, the highest correlation value (0.751) was found in Gandapura, then followed by Peudada (0.745); however, RMSE value obtained in the former region (31.116) was also higher than that of Peudada (23.043). Based on these RMSE values, it indicates that, compared to Gandapura, GCM showed a better estimation at the Peudada rainfall station.
The performance of RBF kernel in rainfall prediction by using GCM and observational rainfall data was plotted using a Taylor diagram. This diagram was constructed only for Gandapura and Peudada stations since they possessed a higher correlation with relatively low RMSE values compared to other tested stations. Figures 3 and 4 describe the values of correlation magnitude, standard deviation, and RMSE of each model along with its GCM and rainfall observational data.

Figure 3. Peudada Station  
Figure 4. Gandapura Station

Figure 3 showed the standard deviation of Peudada station (SD ± 31). The best GCM model, as illustrated in figure 3, is model 1 (CMC1-CanCM3) and model 3 (GDAPS T106L21). The correlation and standard deviation value of both models were close to the rainfall data from observation stations. The correlation and standard deviation values of model 1 are 0.74 and ± 31, while those of model 3 are ± 0.52 and ± 33, respectively. Based on this finding, the first model was stated to be more capable of predicting rainfall in Peudada station during the dry season. Similarly, model 1 was also found to be compatible to predict rainfall in Gandapura station as demonstrated in Figure 4. The result obtained in this study was similar to that attained in Agmalaro’s investigation [6], in which, GCM data was discovered to produce good rainfall estimation. Hence some outliers were found in that study since different grid-scale (5x5), as well as different models of GCM, were implemented.

Further, the examination of the kernel function is carried out by calculating the mean value of correlation and RMSE data of 10 GCM models. The results are shown in Figures 5 and 6 below.
Figures 5 and 6 provide bar charts describing the correlation and RMSE mean values in each kernel function. The highest correlation (0.828) and the lowest RMSE (24.035) mean value is shown by RBF, which confirms that about 83% of observational values are similar to the predicted values. Whereas, the other two kernels exhibited a lower correlation and higher RMSE mean values. Based on this finding, the RBF kernel was revealed to be the most compatible model to predict rainfall. The SVR kernel is considered effective for rainfall prediction when it exhibits high correlation value, yet low RMSE result.

The study emphasized that a good SVR performance on rainfall prediction with high accuracy and efficiency is affected by several parameters, including C, γ, d and ε. These parameters can be optimized by changing their value from 0.01 to 100. The performance of the SVR model obtained in this study by establishing the value of the kernel's parameters using brute force is in line with that of [19] which implemented an SVR model to predict rainfall in Indramayu region, by using the Southern Oscillation Index and 3.4-Nino Sea surface temperature data. Adhani [18] found that RBF kernels also exhibited the highest correlation (0.76) and RMSE values (1.73) compared to linear and polynomial kernels. However, these values are incommensurability due to the different models and grid-scale employed in the studies.

4. Conclusion
The GCM output, integrated with a 7x7 grid-scale statistical downscaling technique, had been used to predict rainfall in Bireun District. Among the three kernel functions of the statistical downscaling technique implemented in this study, the RBF kernel function was found to produce the best model performance on rainfall prediction in Bireun District compared to linear and polynomial kernel functions. The compatibility of the RBF kernel for rainfall prediction was due to its low correlation (0.828) and RMSE (24.035) mean values. Compared to other GCM models, CMC1-CanCM3 exhibited a very good estimation performance which was very close to the observational rainfall data in some regions of Bireun District, especially Peudada (correlation value: 0.745; RMSE value: 23.043) and Gandapura (correlation value: 0.751; RMSE value: 31.166) regions. Hence, CMC1-CanCM3 is recommended to be utilized as a rainfall forecasting tool during the drought in Bireun District and could contribute to a better cropping strategy arrangement in the area to prevent crop failure. A further investigation on rainfall forecasting by using a different GCM data integrated with different Statistical Downscaling techniques, such as Statistical Dynamical Downscaling, along with the implementation of the Ant Colony Optimization algorithm to obtain optimum parameter value is highly suggested.

Acknowledgment
We express our gratitude to the Directorate of Research and Community Service (DRPM) of the Ministry of Research and Technology of Higher Education that provided funding to this research through a program called "Skim hibah penelitian dosen pemula tahun 2018 didanai tahun 2019"

Reference
[1] Chen S T, Yang T C, Kuo C M, Kuo C H and Yu P S 2013 Probabilistic Drought Forecasting in Southern Taiwan Using El Niño-Southern Oscillation Index Terrestrial, Atmospheric & Oceanic Sciences 24(5)
[2] Evana, L, Effendy S and Hermawan E 2008 Pengembangan Model Prediksi Madden Julian Oscillation (Mjo) Berbasis Pada Hasil Analisis Data Real TIME Multivariate Mjo (Rmm1 Dan Rmm2)(Prediction Model Development Madden Julian Oscillation (Mjo) Based on the Results of Data Analysis Jurnal Agromet Indonesia 22(2) 144
[3] Islam M, Chan A, Ashfold M, Ooi C and Azari M 2018 Effects of El-Niño, Indian Ocean Dipole, and Madden-Julian Oscillation on Surface Air Temperature and Rainfall Anomalies over Southeast Asia in 2015 Atmosphere 9(9) 352
[4] Hikmatullah, Yayen A, Soleh, Lili M dan Sudiarjo 2015 Atlas Peta Pengembangan Kawasan Padi Kabupaten Bireuen Provinsi Aceh (Jakarta: Kementrian Pertanian)
[5] Wigena HA 2006 Pemodelan statistical downscaling dengan regression projection pursuit untuk peramalan curah hujan (kcas curah hujan bulanan di Indramayu) Disertasi (Bogor: Institut Pertanian Bogor)

[6] Agmalaro AM 2011 Pemodelan statistical downscaling data GCM menggunakan support vector regression untuk memprediksi curah hujan bulanan Indramayu Tesis (Bogor: Institut Pertanian Bogor)

[7] Tisseuril C, Vrac M, Lek S and Wade AJ 2010 Statistical downscaling of river flows J. of Hydrology 385 279–291

[8] Ullah K, Shrestha S and Guha S 2012 A Decision Support Tool for Selection of Suitable General Circulation Model and Future Climate Assessment J. Earth Sci Climate Change 3 117

[9] Salvi K, Kannan S and Ghosh S 2011 Statistical downscaling and bias-correction for projections of Indian rainfall and temperature in climate change studies In 4th International Conference on Environmental and Computer Science 19 pp 16-18

[10] Ghosh S and Mujamdar PP 2006 Future rainfall scenario over Orissa with GCM projections by statistical downscaling J. Current Science 90(3)

[11] Zhang P and Karori AM 2008 Downscaling NCC CGCM output for Seasonal precipitation prediction over Islamabad – Pakistan Pakistan Journal of Meteorology 4 8

[12] Buono A et al 2010 A Neural Network Architecture for Statistical Downscaling Technique: A Case Study in Indramayu District Proc. AFITA 2010 at International Conference (Bogor: Bogor Agriculture University) p 99

[13] Wilby RL and Wigley TML 2000 Precipitation Predictors For Downscaling: Observed and General Circulation Model Relationships International Journal Of Climatology 20 641 – 661

[14] Aksornsingchai A and Srinilta C 2011 Statistical Downscaling for Rainfall and Temperature Prediction in Thailand Proceedings of the International MultiConference of Engineers and Computer scientists (Hong Kong) 1 pp 16-18

[15] Sutikno 2008 Statistical Downscaling Luaran GCM dan Pemanfaatannya untuk Peramalan Produksi Padi Disertasi (Bogor: Institut Pertanian Bogor)

[16] Priyadarshini D, Acharya M and Mishra A P 2011 Link Load Prediction using Support Vector Regression and Optimization International Journal of Computer Applications 24(7)

[17] Taylor K E 2001 Summarizing multiple aspects of model performance in a single diagram Journal of Geophysical Research: Atmospheres 106 (D7) 7183-7192

[18] Adhani G 2013 Pendugaan Curah Hujan Musim Kemarau Menggunakan Data Southern Oscillation Index dan Suhu Permukaan Laut Nino3.4 dengan Metode Support Vector Regression Skripsi (Bogor: Institut Pertanian Bogor)