Eduvest – Journal of Universal Studies
Volume 5 Number 6, June, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

Optimizing Machine Learning for Daily Rainfall Prediction in Bogor: A Statistical Downscaling Approach

Fradha Intan Arassah[1], Kusman Sadik[2], Bagus Sartono[3], Parwati Sofan[4]
Faculty of Mathematics and Natural Sciences, IPB University, Indonesia[1,2,3]
Research Center for Geoinformatics, National Research and Innovation Agency, Indonesia[4]
Email: icyarassah@apps.ipb.ac.id

ABSTRACT

This study explores the use of machine learning models as a statistical downscaling technique to predict daily rainfall in Bogor, Indonesia. The general circulation model (GCM) is a leading tool for climate prediction, and this research applied a two-stage machine learning model to improve its predictions. The main objectives were to evaluate different GCM domains and handle missing data using two imputation approaches. The first stage involved constructing datasets with varying methods for addressing missing values, followed by the application of a support vector classification (SVC) model to classify rainy and non-rainy days. In the second stage, a recurrent neural network (RNN) model was developed to predict daily rainfall amounts. The results revealed that using random forest imputation for missing data enhanced model accuracy and reduced the root mean square error (RMSE). Among the different GCM domains, the 5 km resolution GCM data was the most accurate when compared to local station climatology. The SVC model, using a radial basis function kernel, achieved an impressive classification accuracy of 98.5%, while the RNN model achieved an RMSE of 16.19. These findings are valuable for improving rainfall predictions and can provide effective data-driven recommendations for disaster mitigation efforts in the region.

KEYWORDS *GCM; Statistical Downscaling; Support Vector Machine; Recurrent Neural Network; Relevance Vector Machine*

INTRODUCTION

Machine learning techniques are renowned for their excellent predicting accuracy and capacity to model both linear and nonlinear data with little assumptions (Nurhidayat & Fatrianto, 2021). According to Roihan et al. (2019), these techniques are essential for analyzing and modeling data and may be generally divided into supervised learning and unsupervised learning approaches. According to Jiang et al. (2020), supervised learning involves developing statistical models that aim to predict output responses based on input variables. Unsupervised learning, on the other hand, focuses on data without predefined labels in an effort to comprehend the relationships and structure within it (James et al., 2021). When developing rainfall prediction models, machine learning's innate capacity to spot trends and resolve challenging issues is especially helpful. Giving useful advice for disaster prevention initiatives requires a precise assessment of increases or declines in intense rainfall.

Advanced techniques for forecasting climate and weather are numerical models, such as the general circulation model (GCM) (Fadhli et al., 2020). The GCM offers a multitude of data that is crucial for modeling, including detailed climate parameters on a regional scale

(Tan et al., 2020). Precipitation, temperature, wind speed, sunshine length, shortwave and longwave radiation, humidity, and other climate parameters are all included in the GCM dataset. The link between global-scale grid data (predictor variables) and local-scale grid data (response variables) is described by statistical downscaling (SD) approaches, which enable the use of GCM data for local-scale rainfall estimation (Farikha et al., 2021).

Furthermore, technical problems with measurement equipment may result in the absence of traditional weather observations from meteorological stations, which reflect local-scale weather, on a particular day. Consequently, missing data in time series is a frequent issue with weather and climate data derived from ground measurements (Anderson & Gough, 2018). The random forest (RF) approach can be used to solve this problem (Hong & Lynn, 2020). The RF approach is a standard for nonparametric imputation as numerous studies have shown that it performs better than conventional imputation methods. According to a study by Waljee & al. (2013), where the data were missing entirely at random (MCAR), missForest consistently outperformed other imputation techniques such as k-nearest neighbors (kNN) and mice by producing the lowest imputation error.

One useful technique for handling both linear and nonlinear data types is support vector classification (SVC), which is especially useful for identifying rainy and non-rainy days in nonlinear time series data (Saikin et al., 2021). Effective at identifying trends and producing precise outcomes, the SVC model maximizes the distance between groups in order to optimize its hyperplane (Sitepu, 2022). Nevertheless, problems occur when linear separation in the input space is not possible, requiring the employment of kernels to convert data into a higher-dimensional space (Ovirianti et al., 2022). By automatically adjusting the class weights to be inversely proportionate to the class frequency, the SVC can manage unbalanced data.

At the same time, rainy day data is modeled using the recurrent neural network (RNN), which is widely acknowledged as an appropriate model for rainfall estimation because rainfall data is time series. Previously, Sulaiman et al. (2022) used RF, SVC, and artificial neural network (ANN) models to predict rainfall. These models did not, however, produce outcomes that were satisfactory. Using an RNN, a neural network technique that is better at processing time series data than an ANN, this study presents a novel way by implementing a wide range of techniques. In order to estimate rainfall, a two-stage machine learning method using RF, SVC, and RNN for statistical downscaling is used.

This study presents an innovative approach that integrates machine learning techniques such as RF, SVC, and RNN for rainfall prediction in *Bogor*, *Indonesia*. Previous studies, such as Sulaiman et al. (2022), implemented RF, SVC, and ANN for rainfall forecasting but were unsuccessful in producing satisfactory results. Similarly, Wang et al. (2020) used GCM data for downscaling in other geographical locations but did not integrate machine learning models for post-processing. The novelty of this research lies in its two-stage machine learning method, combining RF, SVC, and RNN to improve the accuracy of rainfall predictions in a region with high rainfall variability like *Bogor*. Furthermore, the study addresses missing data issues—an essential challenge in climate prediction—by utilizing the random forest method for data imputation, which has been shown to outperform traditional imputation methods. This research offers a more precise approach to statistical downscaling and provides an effective model for future climate prediction in similar regions.

The main objective of this work is to use the previously published two-stage machine learning approach to simulate rainfall prediction in the Bogor city region of West Java, Indonesia. The choice of Bogor City as our research site was motivated by the city's 3,500–5,000 mm of yearly rainfall (Setiawan & Wibowo, 2021). This substantial rainfall above the 2,000–3,000 mm annual national average (Hidayat & Empung, 2016), creating particular difficulties for data processing and predictive modeling adapted to Bogor city's particular circumstances. The study offers implications for improving climate forecasting accuracy, particularly in regions with extreme weather patterns, by creating a model that can be applied to other parts of *Indonesia* or regions with similar climatic conditions. The results will provide valuable insights for disaster mitigation and planning.

RESEARCH METHOD

The predictor variable data (X) utilized in this study are GCM output data obtained from the NEX-GDDP-CMIP6 NASA Global platform. The information is arranged on a global grid that spans the 90°N to 60°S and 180°E to 180°W ranges, with a scale of 0.25° latitude x 0.25° longitude. By performing domain cropping on the GCM data, the dataset was limited to the Bogor region of the West Java Province. Spatial resolutions of 10 km, 5 km, and 2.5 km are present in the resulting dataset. The first figure displays the GCM domain.

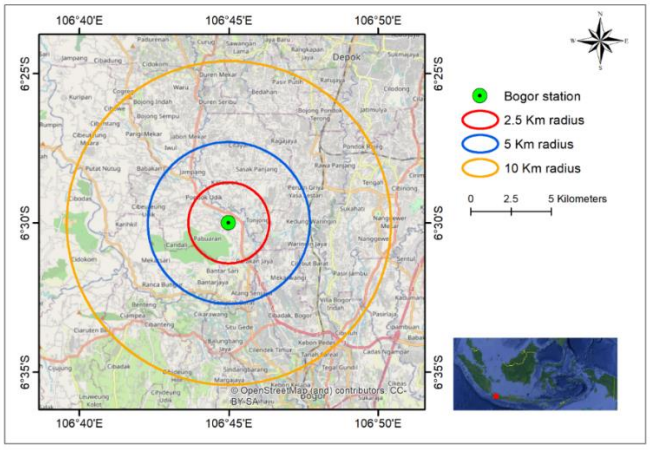


Figure 1. GCM domain for 2.5 km, 5 km, and 10 km radii from Bogor's Meteorological Station

The response variable (Y) in this study is the daily rainfall data (in mm) recorded at a climatology station located in Bogor City, positioned at 6°30'S and 106°45'E, covering the period from January 2015 to March 2023. Furthermore, a description of the predictor factors is provided in Table 1.

Table 1. Predictor variables

Variable	Unit
Minimum temperature	°C
Maximum temperature	°C

Average temperature	℃
Relative humidity (RH)	%
Shortwave radiation	W/m^2
Longwave radiation	W/m^2
Precipitation	mm

Statistical Downscaling

The Coupled Model Intercomparison Project (CMIP) is a collaborative framework of climatology that aims to improve our comprehension of climate change (Kurniadi et al., 2023). To improve climate change scenario modeling, CMIP was created in stages (Meehl et al., 2000). Phase 1 and Phase 6 of the project, which focuses on a global GCM, are the latest stages. According to Try et al. (2022), rainfall data from CMIP6 showed better correlation and lower error coefficients than those from CMIP5, which was explained by Phase 6 correcting the shortcomings of Phase 5.

The NASA Earth Exchange Global released the NEX-GDDP-CMIP6 dataset, which includes global climate scenarios based on GCM simulations (with a spatial resolution of 27.83 km) inside CMIP Phase 6. For obtaining data from Southeast Asia, EC-Earth3-Veg-LR stands out as the best model among the 35 GCMs in CMIP Phase 6. For the SD technique to work, a domain or large-scale climatic variable (predictor variable X) must be connected to either a local variable (response variable Y) or data from a weather station.

One crucial step in the statistical downscaling (SD) method is defining the domain (Sulaiman et al., 2022). The domain represents the geographic location and spatial extent of the atmospheric surface area within the GCM, which provides the predictor variable values used to estimate rainfall at specific local stations. Selecting the appropriate GCM domain greatly affects the accuracy of the forecast, making it a vital element in the SD process. To refine global GCM data, the domain is cropped through a process known as climatological reduction.

Random Forest

The RF algorithm is a flexible technique for building ensembles of decision trees by utilizing bagging to combine various randomly generated predictors (Ali et al., 2012). t is applicable in both classification and regression tasks and has been effectively employed for imputing missing data (Tang & Ishwaran, 2017). Due to the frequent occurrence of missing values in statistical datasets, imputation methods based on random forests are increasingly being adopted. Each missing entry is regarded as independent of the others, and it is assumed that these missing values are dispersed randomly throughout the data. The random forest (RF) model's capacity to handle such randomly missing data demonstrates its applicability in a variety of analytical scenarios (Emmanuel et al., 2021). The general steps for using the random forest algorithm to impute missing data are as follows:

(1) Detect the locations of missing values, which in this case are assumed to be in the response variables.

(2) Divide the dataset into calibration data (with no missing values) and validation data (containing missing values), where the goal is to predict the missing responses. (3) Train a random forest regressor using the calibration data.

(4) Apply the trained model to the validation data to estimate the missing values.

Support Vector Classification

A support vector machine (SVM) is a machine learning approach that utilizes kernel-based functions to perform both linear and nonlinear modeling. It is applied in two main areas: classification, referred to as SVC, and regression, known as SVR (Stekhoven & Bühlmann, 2012). In the initial step, SVC is used to categorize data into two classes, namely rainy and non-rainy days. For the region of Bogor, Indonesia, a day is considered rainy if it receives more than 1.5 mm of precipitation, marked as "1" in this study. Days with rainfall below 1.5 mm are categorized as non-rainy and labeled as "0". The effectiveness of the SVC model largely depends on the choice of parameters, particularly the regularization parameter C and the type of kernel function used (Al Azies et al., 2019a). SVC commonly utilizes four primary kernel types: radial basis function (RBF), sigmoid, polynomial, and linear. The model's classification process is based on the following function.

$$f(x) = \text{sign}(\sum_{k=1}^{m} \alpha_k y_k K(x_k, x) + b) \quad (1)$$

where

$K(x_k, x) =$ kernel function

value$0 \leq \alpha_k \leq C$

y_k = output

x_k = support vector

m = total support vector

In order to translate nonlinear decision limits into linear equations in higher-dimensional spaces, kernel functions usually work with a calibration dataset (Kowalczyk, 2017). The RBF kernel, sigmoid kernel, linear kernel, and polynomial kernel were among the many kinds of kernel functions used in this investigation (Al Azies et al., 2019b).

Recurrent Neural Network

One type of neural network (NN) architecture created especially for processing continuous or sequential input is the RNN (Petneházi, 2018). For jobs involving time series data, like weather forecasting, an RNN is frequently used (Editor Wolfgang Walz, n.d.). An RNN can efficiently recognize patterns in data and use them to make accurate predictions because of its capacity to store memory through a feedback loop (Editor Wolfgang Walz, n.d.). The temporal aspect of data is especially well-captured by RNNs since information from past or present time points affects subsequent instances (Editor Wolfgang Walz, n.d.).

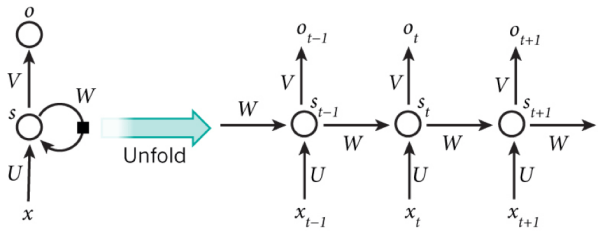


Figure. 2. RNN diagrams

A circuit schematic is shown on the left in Figure 2 (Aksasse et al., 2020), with a black box signifying a one-time step delay. An internal hidden state represented by the letter "s" in an RNN makes it easier for data to move between nodes in the network (Lazzeri, 2020). The unrolling of an RNN into a full network is shown in this diagram (Lazzeri, 2020). Additionally, to guarantee the completeness of the sequence, the image on the shows the RNN unfolded into a whole network (Lazzeri, 2020). The symbols seen in the pictures are described as follows (Lazzeri, 2020):

- There are three sets of weights (W) for every unit, represented by the letters x, s, and o.
- The input is represented by x_t at each time step.
- Each time step's hidden state is denoted by s_t, which is determined using the input in the current state as well as the preceding hidden state.

  $$s_t = f(Ux_t + Ws_{t-1}) \qquad (2)$$

  A nonlinear transformation, like tanh or ReLU, is indicated by the function f. Frequently, the initial hidden state s_{t-1} is set to zero.
- At the present time step, the output is indicated by o_t.

Model Evaluation

1. Accuracy and AUC Values

The SVC model's accuracy and AUC were assessed. The accuracy formula is shown below, and it indicates the total accuracy with which the model can forecast rainy and nonrainy days.

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Ngeative} \qquad (3)$$

The performance of a binary classification model at different threshold values is represented by the AUC. The requirements for calculating the AUC are displayed in Table 2.

Table 2. AUC Value

AUC Value	Description
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
< 0.60	Very Poor

2. Root of The Mean Squared of Error

In this study, the root mean squared error (RMSE) measurement is used to assess the model's accuracy. The average number of mistakes between the expected and actual values is determined by the RMSE (Hodson, 2022). The model's predictions are more accurate when the RMSE is lower.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \quad (4)$$

where
O_i: observed data or actual data

P_i: predicted values

N: number of observations

3. Nash-Sutcliffe Efficiency

The Nash–Sutcliffe Efficiency (NSE) is calculated by taking one minus the ratio between the variance of the model prediction errors and the variance of the observed data. Mathematically, it assesses how well the predicted time series replicates the observed time series. When a model is perfect, meaning it has zero prediction error variance the NSE value reaches its maximum of 1, indicating an ideal match between simulated and observed values.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (5)$$

where
O_i: Observed value.

$\bar{O}$: Average value of the observed data.

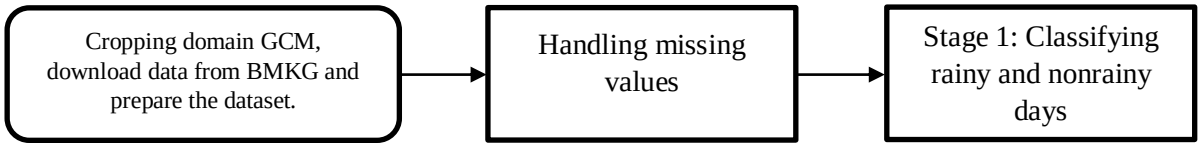
P_i: Predicted value.

Table 3 shows the criteria for determining the NSE value (Lufi et al., 2020):

Table 3. NSE Value

NSE Value	Interpretation
NSE > 0.75	Good
0.36 < NSE< 0.75	Qualified
NSE < 0.36	Not Qualified

Figure 3 shows a flowchart that represents every step of the data processing and analysis procedures used in this investigation. Filling in the blanks, identifying wet and nonrainy days, forecasting rainfall, and assessing the model are the primary tasks included in the flowchart.



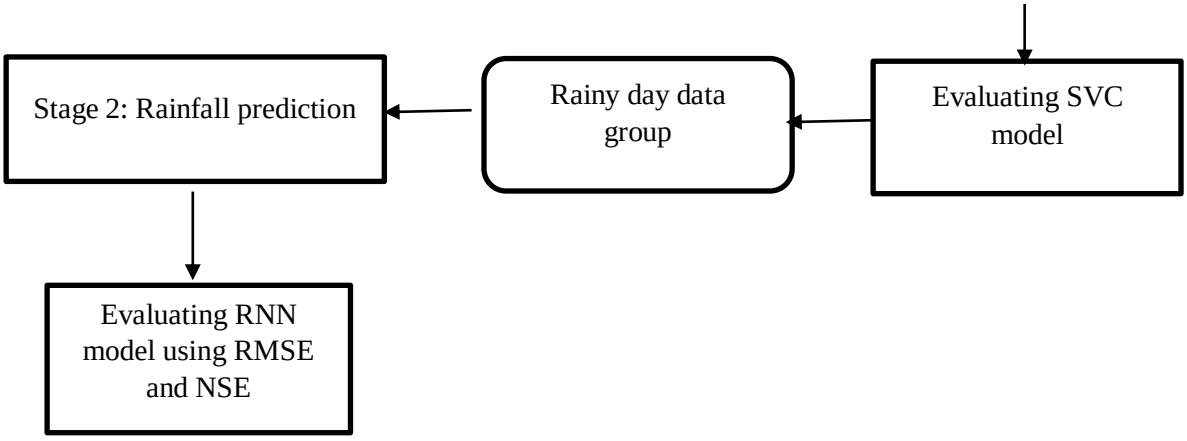


Figure. 3. Flowchart of the data processing and analysis

RESULT AND DISCUSSION

Imputation Missing Values Obtained Using the Random Forest Results

There are 38.3% missing values in the 3,010 rows of rainfall data in the dataset that was acquired from the Bogor Climatology Agency. Missing data can be handled in a variety of ways, including filling in the gaps with zeros or substituting the variable's mean, median, or mode. The effectiveness of random forest (RF) imputation techniques in Python software was evaluated in order to solve these missing variables.

This study's modeling results were compared to data that was imputed by RF imputation (RF data) and data that used zeros to fill in missing values (non-RF data), which are displayed in the next section. In time series data, the complex interactions of meteorological factors, such as minimum temperature, maximum temperature, average temperature, humidity, shortwave radiation, longwave radiation, and precipitation, must be understood by analyzing the correlation between independent variables. Since just one climatological location was covered by the GCM data, there was no multicollinearity among the independent variables.

Support Vector Classification Results

Developing the SVC model to divide the data into two categories, rainy days and nonrainy days was the first step. Eighty percent of the dataset was used for calibration, while twenty percent was used for validation. The best classification model using non-RF data is shown in Table 4 among the SVC models that use an RBF, sigmoid, polynomial, linear kernel, and a set of parameters. The RBF kernel produces the highest accuracy percentage (0.582) and the highest AUC value (0.608) for a 10 km radius of domain GCM. Additionally, Table 4 displays the maximum AUC value of 0.607 and the highest accuracy percentage of 0.578 for domain GCMs with radii of 2.5 and 5 km using the RBF kernel.

Table 4. SVC Results Using Non-RF Data

Radius of Domain GCM	Type of Kernel	Parameter			Evaluation	
		C	γ	d	AUC	Accuracy
10 km	Polynomial	100	0.001	1	0.604	0.577
		100	0.001	2	0.549	0.535

		100	0.001	3	0.604	0.535
	RBF	100	0.001	-	0.608*	0.582*
	Sigmoid	100	0.001	-	0.604	0.577
	Linear	100	0.001	-	0.604	0.577
5 km	Polynomial	100	0.001	1	0.603	0.575
		100	0.001	2	0.459	0.535
		100	0.001	3	0.598	0.535
	RBF	100	0.001	-	0.607*	0.578*
	Sigmoid	100	0.001	-	0.603	0.575
	Linear	100	0.001	-	0.603	0.576
2.5 km	Polynomial	100	0.001	1	0.603	0.575
		100	0.001	2	0.459	0.535
		100	0.001	3	0.598	0.535
	RBF	100	0.001	-	0.607*	0.578*
	Sigmoid	100	0.001	-	0.603	0.575
	Linear	100	0.001	-	0.603	0.576

Table 5 demonstrates that, for a 10 km radius of domain GCM, the SVC model with an RBF kernel and sigmoid, polynomial, or linear kernel models employing RF data produce the maximum accuracy at 0.985 with an AUC of 0.996 from the RBF kernel. The maximum accuracy percentage, 0.985, is likewise displayed in Table 5, but the AUC for domain GCMs with radii of 5 and 2.5 km from the RBF kernel is 0.997. Accordingly, the RBF kernel is used in the preferred SVC model used in this investigation in accordance with these findings.

Table 5. SVC Results Using RF Data

Radius of Domain GCM	Type of Kernel	Parameter			Evaluation	
		C	γ	d	AUC	Accuracy
10 km	Polynomial	100	1	1	0.545	0.499
		100	1	2	0.444	0.409
		100	1	3	0.387	0.302
	RBF	100	1	-	0.996*	0.985*
	Sigmoid	100	1	-	0.494	0.495
	Linear	100	1	-	0.545	0.499
5 km	Polynomial	100	1	1	0.546	0.510
		100	1	2	0.441	0.381
		100	1	3	0.342	0.295
	RBF	100	1	-	0.997*	0.985*
	Sigmoid	100	1	-	0.494	0.509
	Linear	100	1	-	0.546	0.510
2.5 km	Polynomial	100	1	1	0.546	0.510
		100	1	2	0.441	0.381
		100	1	3	0.342	0.295
	RBF	100	1	-	0.997*	0.985*
	Sigmoid	100	1	-	0.494	0.509
	Linear	100	1	-	0.546	0.510

Reccurent Neural Network

After an RNN model was built utilizing the rainy-day group's SVC findings using both RF and non-RF data, the data were split into calibration and validation data with proportions of 80% and 20%, respectively. With a low RMSE and a high NSE, the RNN's performance is displayed in Table 6. The RNN results for the RMSE and NSE were comparable for the RF and Non-RF data.

Table 6. RNN Results Using Non-RF and RF for Rainy Data

Radius of Domain GCM	Evaluation (Non-RF)		Evaluation (RF)	
	RMSE	NSE	RMSE	NSE
10 km	18.787*	0.004	17.404	0.003*
5 km	18.877	0.01*	16.187*	0.001
2.5 km	18.877	0.01*	16.187*	0.001

In this study, the performance of a two-stage machine learning framework was evaluated using two different datasets: one where missing values were imputed using the RF method (referred to as RF data), and another where missing values were replaced with zeros (non-RF data). The research also explored three GCM domains to examine how well global climate data could be downscaled.

For the classification phase, which utilized support vector classification (SVC), the study compared the effectiveness of four different kernel functions. Across all GCM domains, the SVC models trained on both RF and non-RF data demonstrated that the radial basis function (RBF) kernel delivered the best performance, achieving the highest values for both the area under the curve (AUC) and overall model accuracy.

Specifically, for the non-RF dataset, the GCM domain with a 10 km radius produced better AUC and accuracy compared to the 5 km and 2.5 km domains. In contrast, for the RF dataset, the domains with 5 km and 2.5 km radii outperformed the 10 km radius in terms of model accuracy and AUC. These results are consistent with findings from Shangzhi Hong and Henry S. Lynn (2020), which showed that RF-imputed data enhanced classification accuracy. They also align with the conclusions drawn by Nurul Ainina Sulaiman et al. (2022), who highlighted the superiority of SVC models that used RBF kernels in combination with RF-imputed data.

After classification into rainy and non-rainy categories, the regression phase was conducted solely on the rainy-day subset to forecast daily rainfall amounts. This phase employed recurrent neural network (RNN) models, trained separately on RF and non-RF datasets. Table 7 presents the comparative performance of the two-stage SVC–RNN models across GCM domains with radii of 10 km, 5 km, and 2.5 km for both data types.

Table 7. Accuracy and RMSE Results of the Two-stage Model Performance

Radius of Domain GCM	Accuracy of SVC		RMSE of RNN	
	Non-RF	RF	Non-RF	RF
10 km	0.582*	0.985*	18.787*	17.404
5 km	0.578	0.985*	18.877	16.187*
2.5 km	0.578	0.985*	18.877	16.187*

Overall, the use of RF-imputed data enhanced the performance of the SVC model by increasing its classification accuracy and reduced the RMSE in the RNN model across all domain sizes. As shown in Table 7, the SVC-RNN model using non-RF data performed best in terms of accuracy and RMSE within the 10 km radius GCM domain, while the model using RF data achieved superior results in the 5 km and 2.5 km radius domains.

When comparing the two-stage models across the three GCM domains, the findings indicate that the 5 km and 2.5 km radius domains yielded identical performance metrics. This implies that the optimal domain size for effective downscaling using local climate station data is 5 km.

In this research, the random forest algorithm was applied to address missing values, and a 5 km GCM domain radius was selected as the most effective. The classification step employed SVC with RBF kernel, which was then integrated with a recurrent neural network (RNN) to forecast rainfall amounts for the rainy-day subset in January 2024.

.

Forecasting Daily Rainfall Using a Two-Stage Model

Predictions were made using simulated data from the NASA Earth Exchange Global-provided NEX-GDDP-CMIP6 dataset, which covered the period from January 1 to January 31, 2024. All things considered, the two-stage model accurately captures the fluctuating daily rainfall pattern in January 2024, as shown in Figure 4.

.

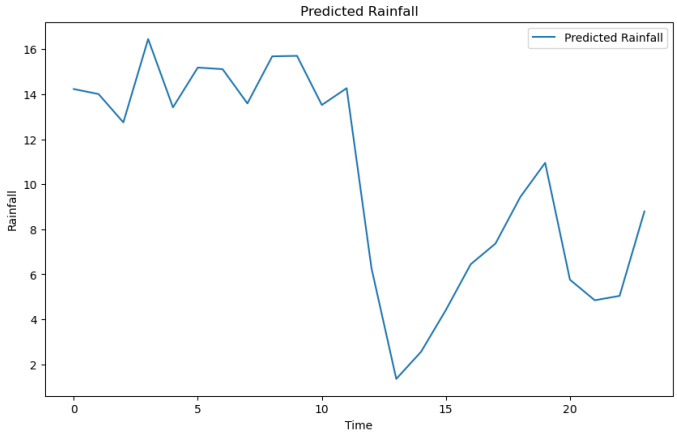


Figure. 4. Rainfall Prediction Period January 2024 in Bogor

Over the 31 days of January, the optimal SVC model using the RBF kernel predicted a total of 25 rainy days, followed by rainfall estimation through the RNN model. As illustrated in Figure 4, rainfall trends downward in the middle of the month before rising again the following day.

CONCLUSION

This study highlights the critical role of GCM domain selection in the downscaling process. By comparing multiple GCM domain configurations, the most effective domain for improving the accuracy of local rainfall predictions was identified. Incorporating RF imputation to handle missing data provided key benefits, such as reducing bias, minimizing classification errors, and boosting overall model performance. Although both the 10 km and

5 km GCM domain radii yielded strong results, the 5 km radius showed greater accuracy and lower RMSE when paired with RF-based imputation. The most effective SVC model utilized the RBF kernel. Overall, the implementation of the two-stage SVC-RNN model proved effective for forecasting future rainfall. As a result, forecasts indicating substantial changes in rainfall, whether increases or decreases can help local authorities take early and informed action to mitigate disaster risks.

.

REFERENCES

Aksasse, H., Aksasse, B., & Ouanan, M. (2020, June 1). Developing Good Habits Using Deep Learning Techniques. *2020 International Conference on Intelligent Systems and Computer Vision, ISCV 2020*. https://doi.org/10.1109/ISCV49265.2020.9204069

Al Azies, H., Trishnanti, D., & Mustikawati, E. (2019a). Comparison of Kernel Support Vector Machin (SVM) in Classification of Human Development Index (HDI). *Journal of Proceedings Series*, *6*, 53–57.

Al Azies, H., Trishnanti, D., & Mustikawati, E. (2019b). ISSN (2354-6026) 53 The 1 st International Conference on Global Development-ICODEV. In *IPTEK Journal of Proceedings Series* (Issue 6).

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*, 272–278.

Anderson, C. I., & Gough, W. A. (2018). Accounting for missing data in monthly temperature series: Validation rule-of-thumb omission of months with missing values. *International Journal of Climatology*, *38*, 4990–5002. https://doi.org/10.1002/joc.5801

Editor Wolfgang Walz, S. (n.d.). *Machine Learning for Brain Disorders*. http://www.springer.com/series/7657

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A Survey on Missing Data in Machine Learning. *J Big Data*. https://doi.org/10.1186/s40537-021-00516-9

Fadhli, N., Wigena, A. H., & Djuraidah, A. (2020). *Determination of General Circulation Model Domain Using LASSO to Improve Rainfall Prediction Accuracy in West Java*. https://doi.org/10.4108/eai.2-8-2019.2290466

Farikha, E. F., Hadi, A. F., Anggraeni, D., & Riski, A. (2021). Projection pursuit regression in statistical downscaling model using artificial neural network for rainfall prediction. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1872/1/012021

Hidayat, A. K., & Empung. (2016). Analisis Curah Hujan Efektif dan Curah Hujan Dengan Berbagai Periode Ulang Untuk Wilayah Kota Tasikmalaya dan Kabupaten Garut. *Jurnal Siliwangi*. https://doi.org/10.37058/jssainstek.v2i2.99

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. In *Geoscientific Model Development* (Vol. 15, Issue 14, pp. 5481–5487). Copernicus GmbH. https://doi.org/10.5194/gmd-15-5481-2022

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of nonnormality, nonlinearity, and interaction. *BMC Med. Res. Methodology*, *20*, 1–12. https://doi.org/10.1186/s12874-020-01080-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer Second Edition.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behav Ther.*, *51*, 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kowalczyk, A. (2017). *Foreword by Daniel Jebaraj*. www.syncfusion.com.

Kurniadi, A., Weller, E., Kim, Y. H., & Min, S. K. (2023). Evaluation of Coupled Model Intercomparison Project Phase 6 model-simulated extreme precipitation over Indonesia. *International Journal of Climatology*, *43*, 174–196. https://doi.org/10.1002/joc.7744

Lazzeri, F. (2020). *Machine Learning for Time Series Forecasting with Python*. Wiley.

Lufi, S., Ery, S., & Rispiningtati, R. (2020). Hydrological Analysis of TRMM (Tropical Rainfall Measuring Mission) Data in Lesti Sub Watershed. *Civil and Environmental Science*, *003*(01), 018–030. https://doi.org/10.21776/ub.civense.2020.00301.3

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The Coupled Model Intercomparison Project (CMIP). *Bull. Am. Meteorol. Soc.*, *81*, 313–318. https://doi.org/10.1175/1520-0477(2000)081<0313:tcmipc>2.3.co;2

Nurhidayat, A. I., & Fatrianto, D. (2021). Prediksi Kinerja Akademik Mahasiswa Menggunakan Machine Learning dengan Sequential Minimal Optimization untuk Pengelola Program Studi. *Journal Information Engineering and Educational Technology*, *5*, 84–91. https://doi.org/10.26740/jieet.v5n2.p84-91

Ovirianti, N. H., Zarlis, M., & Mawengkang, H. (2022). Support Vector Machine Using A Classification Algorithm. *Jurnal Dan Penelitian Teknik Informatika*, *7*, 2103–2107. https://doi.org/10.33395/sinkron.v7i3.

Petneházi, G. (2018). *Recurrent Neural Networks for Time Series Forecasting*. http://arxiv.org/abs/1901.00069

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2019). Pemanfaatan Machine Learning dalam Berbagai Bidang. *Indonesian Journal on Computer and Information Technology Review Paper*, *5*, 75–82.

Saikin, S., Fadli, S., & Ashari, M. (2021). Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results. *Journal of Informatics and Science*, *4*, 22–27. https://doi.org/10.31326/jisa.v4i1.881

Setiawan, H., & Wibowo, A. (2021). Geomedia Majalah Ilmiah dan Informasi Kegeografian Pembuatan Peta Curah Hujan Untuk Evaluasi Kesesuaian Rencana Tata Ruang Kawasan Hutan Kabupaten Bogor. *Journal UNY*, *19*, 113–121. https://doi.org/10.21831/gm.v19i2.43227

Sitepu, R. (2022). The Analysis of Support Vector Machine (SVM) on Monthly Covid-19 Case Classification. *International Journal on Information and Communication Technology (IJoICT)*, *8*, 40–52. https://doi.org/10.21108/ijoict.v8i2.671

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Nonparametric Missing Value Imputation for Mixed-type Data, Bioinformatics. *Bioinformatics*, *28*, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sulaiman, N. A. F., Shaharudin, S. M., Ismail, S., Zainuddin, N. H., Tan, M. L., & Jalil, Y. A. (2022). Predictive Modeling of Statistical Downscaling Based on Two stage Machine Learning Model for Daily Rainfall in East-Coast Peninsular Malaysia. *Symmetry (Basel)*, *14*, 1–30. https://doi.org/10.3390/sym14050927.

Tan, Y., Guzman, S. M., Dong, Z., & Tan, L. (2020). Selection of effective GCM bias correction methods and evaluation of hydrological response under future climate scenarios. *Climate*, *8*, 1–21. https://doi.org/10.3390/cli8100108

Tang, F., & Ishwaran, H. (2017). Random Forest Missing Data Algorithms. *Stat. Anal. Data Min.*, *10*, 363–377. https://doi.org/10.1002/sam.11348

Try, S., Tanaka, S., Tanaka, K., Sayama, T., Khujanazarov, T., & Oeurng, C. (2022). Comparison of CMIP5 and CMIP6 GCM performance for flood projections in the Mekong River Basin. *Journal of Hydrology*, *40*, 1–19. https://doi.org/10.1016/j.ejrh.2022.101035

Waljee, A. K., & al., et. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*, 1–7. https://doi.org/10.1136/bmjopen-2013-002847