Intechno Journal :
Information Technology Journal
Vol. 8 No. 1 July 2026

Homepage: https://jurnal.amikom.ac.id/index.php/intechno

INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL

e-ISSN: 2655-1438
p-ISSN: 2655-1632

Rainfall Prediction Using Random Forest with Synthetic Minority Over-Sampling Technique (SMOTE)

Sri Bintang Marpaung[1], Ichy Lucya Resta[2], Tugiyo Aminoto[3]
[1,3]Department of Physics, Faculty of Science and Technology, Universitas Jambi, Jambi, Indonesia
[2]Department of Physics Education, Faculty of Teacher Training and Education, Universitas Jambi, Jambi, Indonesia

Article Info

Article history:
Received: 16 July 2026
Revised: 28 July 2026
Accepted: 30 July 2026
Published: 31 July 2026

Keywords:
Rainfall
Random Forest
SMOTE
Machine Learning
Tanjung Perak

ABSTRACT

Rainfall is a crucial meteorological parameter that directly affects maritime activities and port operations, particularly in coastal regions. Tanjung Perak Port, as one of the busiest ports in Indonesia, is highly vulnerable to weather disturbances due to rainfall variability. This study develops a rainfall prediction model based on machine learning using the Random Forest algorithm optimized with the Synthetic Minority Over-Sampling Technique (SMOTE) to address data imbalance. Daily meteorological data for 2013-2024 were obtained from NASA POWER, including rainfall, minimum and maximum temperature, relative and specific humidity, wind speed, and surface pressure. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). SMOTE improved the model's ability to identify high-intensity rainfall events. The combined 2024-2025 evaluation produced an MAE of 1.457 mm, an RMSE of 1.819 mm, and R^2 of 0.861. The model therefore captures seasonal rainfall patterns and has potential as a decision-support tool for weather-risk mitigation at Tanjung Perak Port.

Corresponding Author:
Ichy Lucya Resta
Email: ichylucya@unja.ac.id

1. INTRODUCTION

Rainfall is one of the most important meteorological variables influencing environmental conditions, socio-economic activities, and disaster risk, particularly in tropical coastal regions. In Indonesia, rainfall variability is strongly affected by complex interactions between the atmosphere and ocean and by large-scale climate phenomena such as the monsoon system, El Niño-Southern Oscillation (ENSO), and Madden-Julian Oscillation (MJO). These interactions produce high temporal and spatial rainfall variability, making accurate prediction challenging [1].

Tanjung Perak is one of the busiest ports in Indonesia and plays a crucial role in national and regional logistics. Weather disturbances, especially heavy rainfall accompanied by strong winds, can disrupt port operations, reduce safety, and increase economic losses. Reliable rainfall prediction is therefore essential for operational planning, maritime safety, and risk mitigation in the port area [2].

Conventional rainfall forecasting in Indonesia is generally based on numerical weather prediction models that simulate atmospheric processes using physical equations. Although these models provide valuable information at regional and national scales, their local performance is often limited by high computational cost and difficulty in representing complex local atmospheric processes [3]. Machine-learning approaches have therefore emerged as promising alternatives because they can model non-linear relationships and learn patterns directly from historical data.

Among machine-learning algorithms, Random Forest has attracted substantial attention because of its robustness, predictive accuracy, and ability to estimate feature importance. Random Forest is an ensemble method that combines multiple decision trees trained on randomly selected data subsets, thereby reducing overfitting and improving generalization [4]. Previous studies have also demonstrated its potential for rainfall and weather prediction in coastal and tropical environments [5], [6].

A major challenge in rainfall modelling is data imbalance: extreme rainfall events occur much less frequently than normal conditions. This imbalance can bias the learning process toward the majority class and reduce sensitivity to rare but operationally important events. The Synthetic Minority Over-Sampling Technique (SMOTE) addresses this problem by generating synthetic samples for minority classes and enabling the model to learn from a more balanced representation [7].

This study integrates Random Forest with SMOTE to improve rainfall prediction in the Tanjung Perak area. The objectives are: (8) to develop a Random Forest-based rainfall prediction model using historical meteorological data; (2) to analyse the relative importance of meteorological variables influencing rainfall; and (4) to evaluate model performance using MAE, MSE, and RMSE. The results are expected to support the application of machine learning in meteorology and provide practical information for weather-related decision making in port operations.

2. METHOD

Daily meteorological data were obtained from the NASA POWER (Prediction of Worldwide Energy Resources) database for the Tanjung Perak area for 2013-2024 [10]. Daily rainfall (mm) was used as the target variable, while the predictor variables comprised wind speed (m/s), minimum temperature (°C), maximum temperature (°C), relative humidity (%), specific humidity, and surface pressure (hPa).

Data preprocessing included data cleaning, median imputation for missing values, standardization of variable names, and construction of time-related features. To prevent data leakage, the dataset was split chronologically: data from 2013-2023 were used for training, while 2024 data were reserved for validation.

The Random Forest algorithm was applied because it can represent non-linear relationships in multivariate data. Hyperparameter tuning was performed to determine the optimal number of trees and tree depth. For rainfall-category classification, class imbalance was treated using SMOTE, which generates synthetic minority samples through nearest-neighbour interpolation. SMOTE was applied only to the training subset so that the validation dataset remained independent.

Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Feature-importance analysis was used to identify the meteorological variables with the greatest influence on the Random Forest model. Error values approaching zero indicate better agreement between predictions and observations and provide information on the model's generalization to unseen data [9].

$$\mathrm{MAE} = (1/n)\ \Sigma_{i=1}^{n}\ |y_i - \hat{y}_i| \quad (1)$$

$$\mathrm{MSE} = (1/n)\ \Sigma_{i=1}^{n}\ (y_i - \hat{y}_i)^2 \quad (2)$$

$$\mathrm{RMSE} = \sqrt{[(1/n)\ \Sigma_{i=1}^{n}\ (y_i - \hat{y}_i)^2]} \quad (3)$$

where n is the number of observations, y_i is the observed value, $\hat{y}_i$ is the predicted value, $|y_i - \hat{y}_i|$ is the absolute error, and $(y_i - \hat{y}_i)^2$ is the squared error. RMSE is expressed in the same unit as the original rainfall data.

Table 1. Evaluation criteria based on MAE and RMSE

Range (mm)	Category
< 10	Very Good
10-20	Good
20-30	Fair
30-40	Poor
> 40	Very Poor

The MSE criteria were derived consistently by squaring the RMSE thresholds used in this study.

Table 2. Evaluation criteria based on MSE

Range (mm²)	Category
< 100	Very Good
100-400	Good
400-900	Fair

`

900-1600	Poor
> 1600	Very Poor

Accuracy indicates the closeness of predictions to the observed values and was calculated using Equation (4).

$$\text{Accuracy} = (1 - \text{MAE}) \times 100\% \quad (4)$$

Table 3. Performance classification based on accuracy

Range	Classification
90%-100%	Very Good
80%-<90%	Good
70%-<80%	Fair
60%-<70%	Poor
≤ 60%	Very Poor

3. RESULTS AND DISCUSSIONS

The analysis evaluates the performance of the Random Forest rainfall-prediction model optimized with SMOTE and interprets the numerical results in relation to atmospheric processes in the Tanjung Perak Port area. The discussion focuses on links among meteorological variables, rainfall events, model error, and the research objectives.

3.1 Relationship Between Meteorological Variables and Rainfall

Pearson correlation analysis was used to evaluate the direction and strength of linear relationships among rainfall and the meteorological predictors, including air temperature, humidity, surface pressure, wind speed, and month. Although Random Forest can represent non-linear interactions, the correlation matrix provides an initial physical interpretation of the data structure.

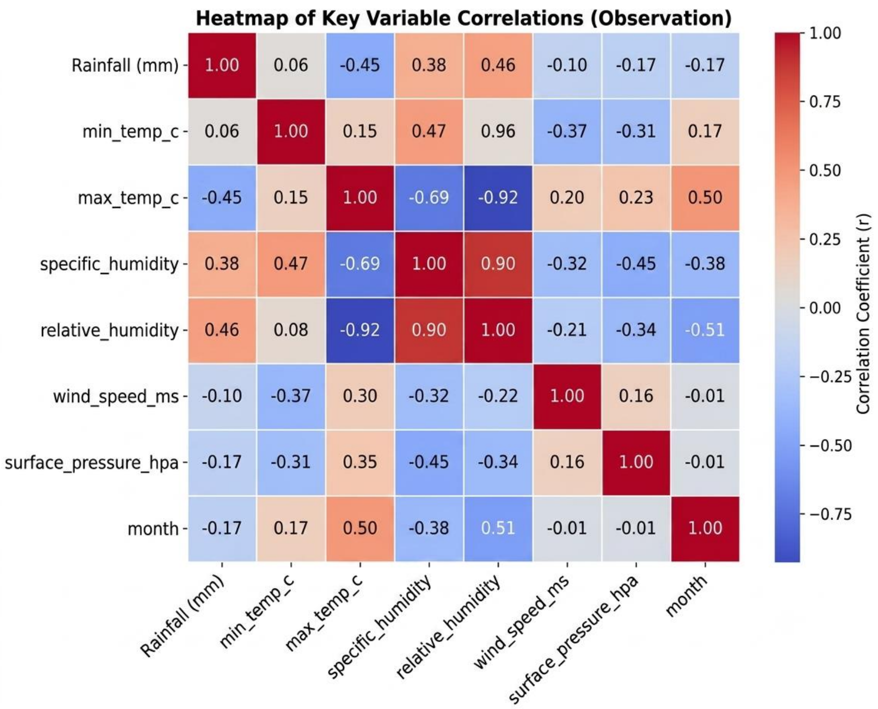

Figure 1. Correlation heatmap among meteorological variables

Relative humidity (r = 0.46) and specific humidity (r = 0.38) show positive relationships with rainfall. Physically, both variables represent the availability of atmospheric water vapour. Higher humidity allows air parcels to reach saturation more rapidly during cooling or uplift, increasing condensation, cloud formation, and precipitation probability. The positive coefficients therefore agree with the fundamental role of moisture supply in rainfall formation.

Maximum temperature and surface pressure show negative relationships with rainfall. High maximum temperature in the dataset is generally associated with clear-sky conditions and stronger solar heating, whereas rainfall events are often linked to enhanced cloud cover. Lower surface pressure favours convergence and upward air motion, which supports adiabatic cooling, saturation, and cloud development. The moderate correlation magnitudes indicate that rainfall emerges from interacting, non-linear processes rather than from a single controlling variable, supporting the use of a machine-learning model.

3.2 Data Imbalance and SMOTE Implementation

The data were divided chronologically into training and testing subsets so that the model was trained on past observations and evaluated on unseen future observations. Maintaining the temporal sequence is essential in time-series prediction because random splitting can transfer future information into the training process and produce unrealistically optimistic performance [10]. The training set comprised approximately 80% of the historical data, including 2013-2023, while 2024 was used as the most recent out-of-sample validation period before forecasting 2025.

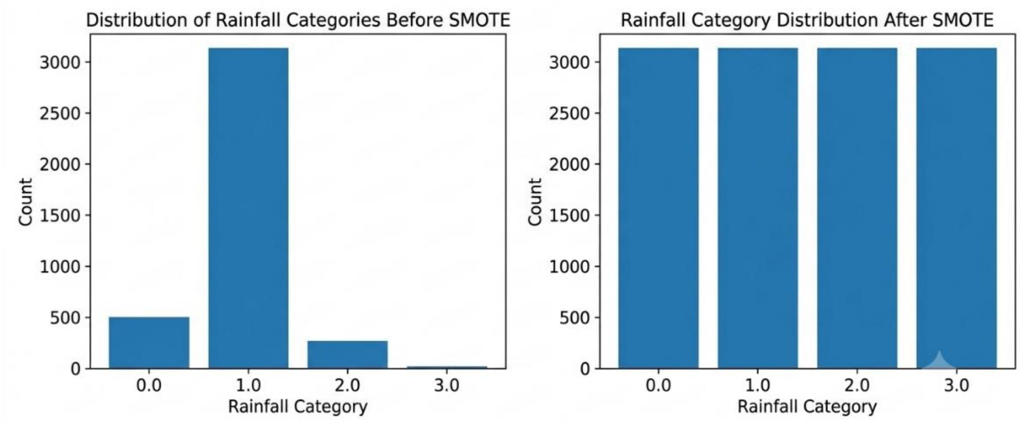

Figure 2. (left) Rainfall categories before oversampling, (right) Rainfall categories after oversampling

Before oversampling, light-rain events dominated the dataset, whereas moderate and heavy rainfall occurred relatively rarely. A classifier trained on this distribution tends to minimize total error by favouring the majority class and can consequently miss high-impact events. Applying SMOTE to the training data balanced the class distribution by generating synthetic minority samples. The more balanced representation improved the model's opportunity to learn the meteorological characteristics of moderate and heavy rainfall and strengthened generalization for rare events.

3.3 Random Forest Modelling and 2024 Validation

Automatic hyperparameter tuning was conducted using Optuna. Optuna applies a Bayesian optimization framework that narrows the search space using information from previous trials and is more efficient than exhaustive grid search for combinations such as tree number, tree depth, and related parameters [8]. Training used the 2013-2023 data, and the optimized model was validated with 2024 data after SMOTE had been applied only to the training subset.

Table 4. Model-validation results for 2024

Metric	Value
MAE	1.318 mm
RMSE	1.639 mm
R^2	0.884

The MAE of 1.318 mm indicates an average absolute deviation below 1.5 mm. The RMSE of 1.639 mm is only slightly greater than the MAE, showing that large deviations are limited to a small number of higher-intensity rainfall periods. Both metrics are below the 10 mm threshold and are classified as very good. The R^2 value of 0.884 means that the model explains 88.4% of the seasonal rainfall variation during 2024, indicating strong stability for a statistical model evaluated over a full annual cycle.

Some reduction relative to validation that directly uses observed lag features is expected because hindcast prediction relies recursively on model-generated rainfall values. Each prediction becomes an input for subsequent steps, allowing errors to propagate. Nevertheless, the model preserved the annual pattern without consistent positive or negative bias, indicating that persistent seasonal signals remained identifiable.

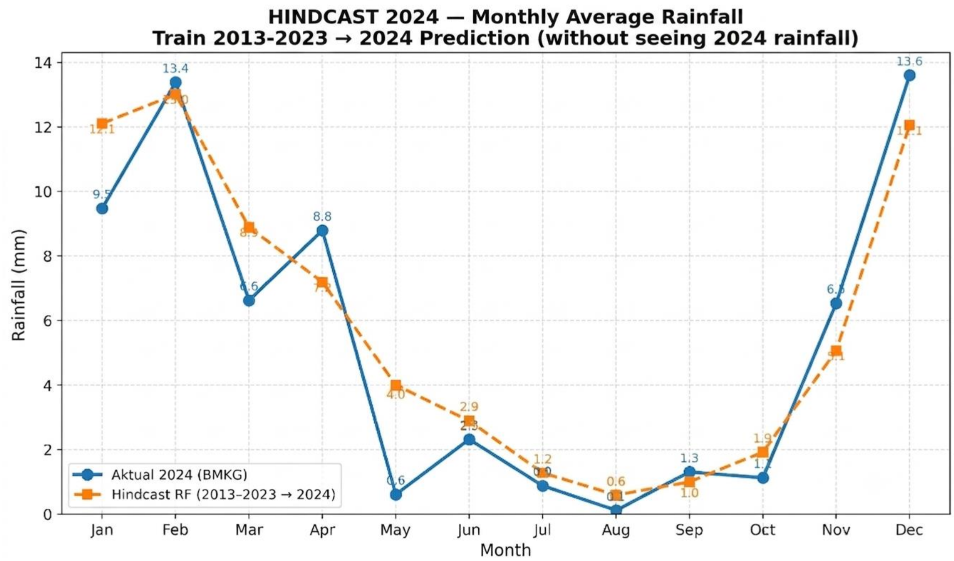

Figure 3. Validation of the 2024 rainfall prediction model

Figure 3 shows that the model reproduces the principal seasonal pattern. Rainfall increases during January-February under the influence of the Asia-Australia monsoon and reaches a peak close to the observed value. During March-April, deviations increase because the monsoon-transition period is more sensitive to interannual disturbances such as the MJO and equatorial waves. The model closely represents the low-rainfall period from May to August, indicating limited recursive error growth when the seasonal signal is stable. Rainfall then increases again from September to December, although November is overestimated. Overall, the timing of the wet and dry phases is maintained.

3.4 Confusion-Matrix Classification Evaluation

The confusion matrix evaluates the ability of Random Forest with SMOTE to classify rainfall into No Rain, Light, Moderate, and Heavy categories. Correct predictions dominate the main diagonal, indicating that the model distinguishes the four classes effectively.

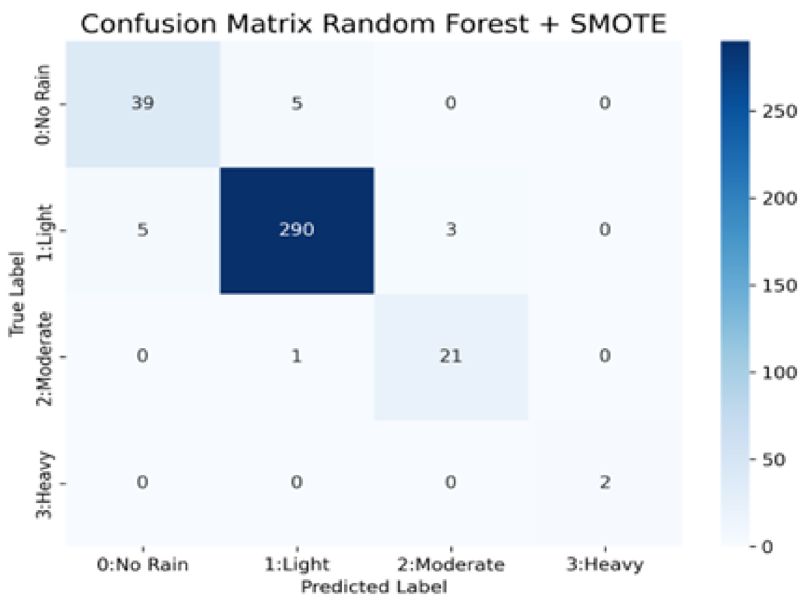

Figure 4. Rainfall-classification confusion matrix

The model correctly identified 39 No Rain cases and 290 Light Rain cases. Twenty-one Moderate cases were classified correctly, with only one case shifted to the adjacent Light category, while the Heavy category was correctly recognized in two cases. Most errors occurred between physically similar adjacent classes, particularly Light and Moderate rainfall, because their predictor conditions can overlap near category thresholds.

SMOTE increased the reported Moderate-class accuracy from approximately 63% to 82% and reduced errors for the Heavy class. The small number of off-diagonal entries indicates that the model generalizes reasonably well rather than merely memorizing the training data. Thus, Random Forest combined with SMOTE improves sensitivity to minority rainfall classes and is potentially useful for operational weather-risk decisions at Tanjung Perak Port.

3.5 Rainfall Prediction for 2025 and Evaluation

The optimized Random Forest Regressor was used to predict 2025 rainfall with a recursive multi-step approach. Observed 2025 rainfall was not used as an input during prediction, producing a realistic forward-forecast scenario without information leakage. Monthly climatological values from 2013-2023 represented slowly varying variables such as temperature, relative humidity, and surface pressure, while lag-1, lag-3, lag-7, 7-day moving-average, and 30-day moving-average rainfall features were updated recursively from preceding predictions.

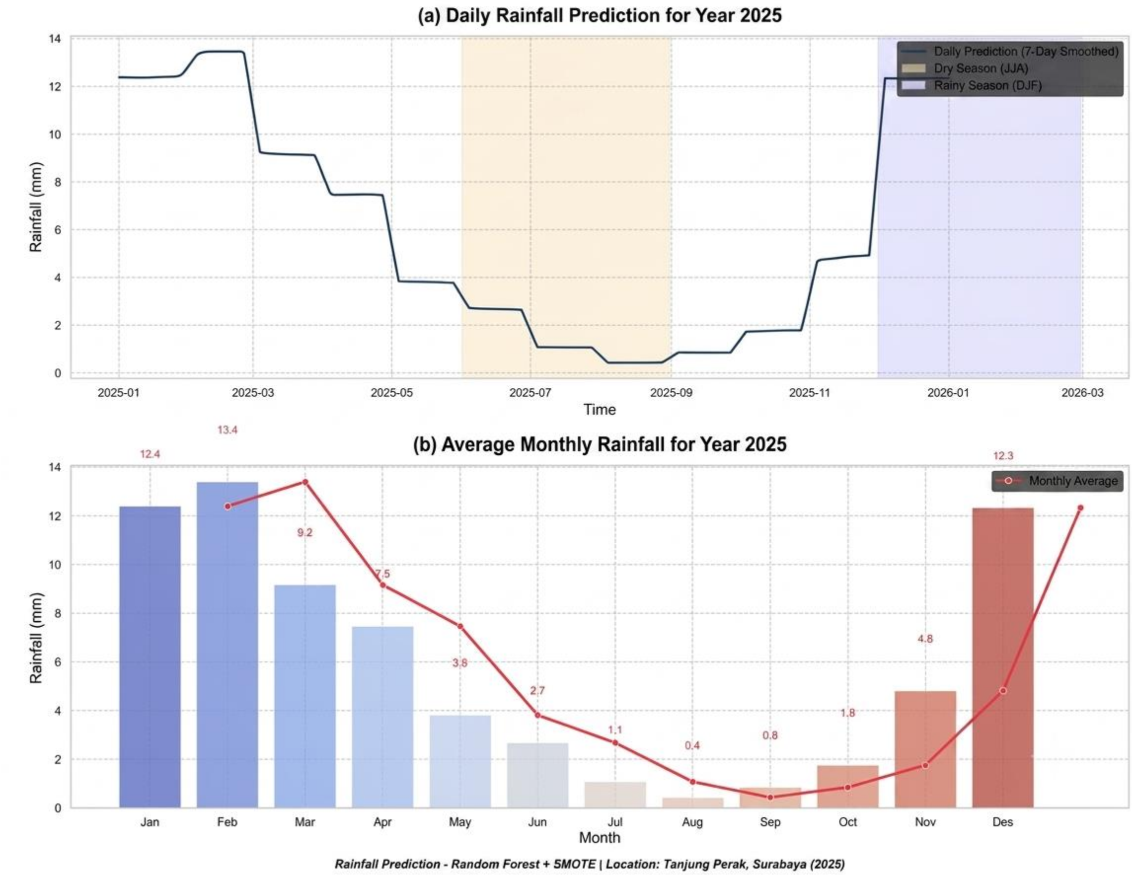

Figure 5. Daily and monthly rainfall forecasts for 2025

The forecast follows the Asia-Australia monsoon cycle: the highest rainfall occurs in December-February at approximately 10-14 mm per day, while June-August is drier at approximately 0-2 mm per day. Monthly rainfall peaks around January-February, decreases below 4 mm from May to August, and rises toward the end of the year. This annual pattern is consistent with the regional climatological cycle reported by BMKG [2].

`

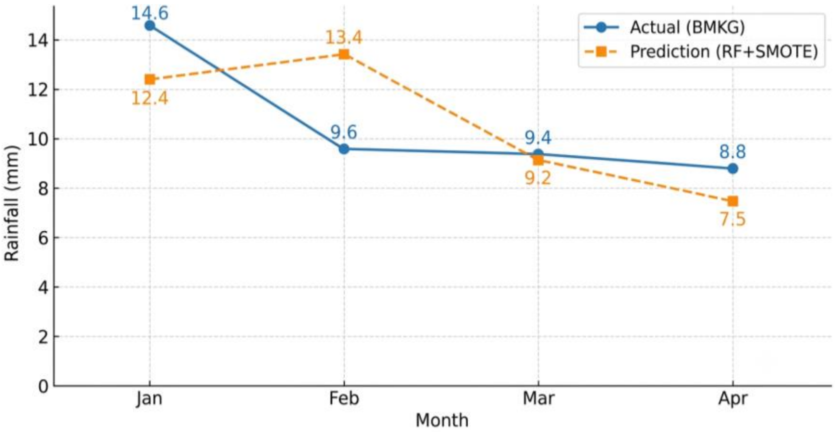

Figure 6. Comparison of observed and predicted rainfall, January-April 2025

Partial validation against BMKG observations for January-April 2025 shows that January rainfall was underestimated: the observed value was approximately 14.6 mm and the model predicted approximately 12.4 mm. February and March were slightly overestimated; in February, the observation was approximately 9.6 mm while the prediction was approximately 13-14 mm. Despite these magnitude differences, the model reproduced the overall decrease from January to April.

Table 5. Evaluation for January-April 2025

Metric	Value
MAE	1.872 mm
RMSE	2.273 mm
MSE	5.168 mm²

The MAE of 1.872 mm, MSE of 5.168 mm², and RMSE of 2.273 mm remain in the very good category. The errors are small relative to the seasonal rainfall range. The higher errors compared with 2024 are physically reasonable because short validation over a transition period is more sensitive to sub-seasonal atmospheric variability that is not fully represented by local predictors. Even with this limitation, the model reproduces the 2025 seasonal tendency and has potential as a decision-support tool.

3.6 Combined Analysis of 2024-2025 Predictions

A combined evaluation from January 2024 to April 2025 was performed to test whether the model maintained seasonal continuity and accuracy across years not included in training. Inter-period evaluation is important because a model may fit one validation year but become unstable when the atmospheric regime changes.

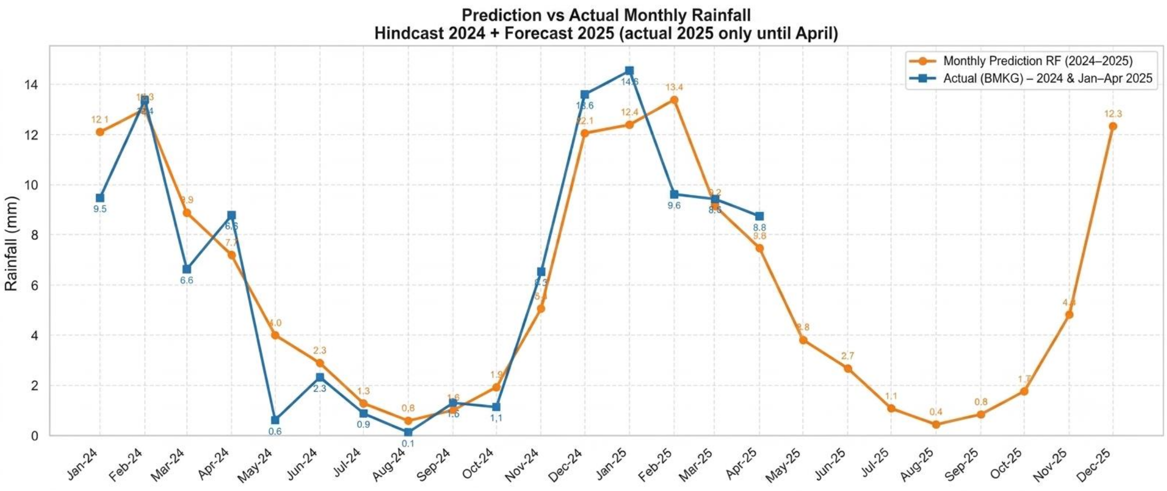

Figure 7. Comparison of predicted and observed monthly rainfall for 2024-2025

The predicted and observed series both show rainfall peaks in January-February 2024, a decline toward the May-August dry season, and an increase in November-December. The decreasing trend from January to April 2025 is also retained. This consistency indicates that the model captures the dominant annual monsoon phase, although individual monthly anomalies remain imperfectly represented.

Table 6. Combined evaluation for 2024-2025

Metric	Value
MAE	1.457 mm
RMSE	1.819 mm
R^2	0.861

The combined MAE of 1.457 mm and RMSE of 1.819 mm indicate an average error below 2 mm. R^2 = 0.861 means that approximately 86.1% of monthly rainfall variability is explained by the model, while the remaining variability is likely associated with large-scale or transient atmospheric processes absent from the predictor set. Performance is slightly lower than the 2024 hindcast but remains very good and confirms strong generalization across the combined period.

4. CONCLUSION

1. A daily rainfall-prediction model was successfully developed for Tanjung Perak using Random Forest and a chronological split, with 2013-2023 as training data and 2024 as validation data. The workflow preserved temporal order, avoided data leakage, and used feature engineering to represent non-linear relationships between meteorological variables and rainfall.
2. Humidity was the dominant predictor: relative humidity contributed a feature importance of approximately 0.67, followed by specific humidity at approximately 0.10 and wind speed at approximately 0.09. This result is physically consistent with the role of water-vapour availability and atmospheric saturation in cloud and rainfall formation, while wind contributes to moisture transport.
3. Validation for 2024 produced MAE = 1.318 mm, RMSE = 1.639 mm, and R^2 = 0.884. Partial validation for January-April 2025 produced MAE = 1.872 mm, RMSE = 2.273 mm, and MSE = 5.168 mm^2. The combined 2024-2025 evaluation produced MAE = 1.457 mm, RMSE = 1.819 mm, and R^2 = 0.861. These values indicate low error and stable representation of the dominant seasonal rainfall pattern.

Future research should incorporate regional-scale variables such as ENSO and MJO indices or sea-surface temperature so that the model can respond more effectively to anomalies not captured by local meteorological variables, particularly during monsoon-transition periods. The model should also be updated regularly as new observations become available.

ACKNOWLEDGEMENTS

The authors thank NASA for providing meteorological data through the NASA POWER database or supporting this research.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Sri Bintang Marpaung: Conceptualization, data curation, methodology, analysis, and writing - original draft. Ichy Lucya Resta: Supervision, validation, and writing - review and editing. Tugiyo Aminoto: Supervision, methodology, and review.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

DATA AVAILABILITY

Data will be made available on request. NASA POWER data are publicly accessible through the POWER Data Access Viewer.

REFERENCES

[1] C. P. Chang, Z. Wang, J. McBride, and C. H. Liu, “Annual cycle of Southeast Asia-Maritime Continent rainfall and the asymmetric monsoon transition,” Journal of Climate, vol. 30, pp. 1-21, 2017.

[2] BMKG, Informasi Iklim dan Prakiraan Musim di Indonesia. Jakarta, Indonesia: Badan Meteorologi, Klimatologi, dan Geofisika, 2023.

[3] A. Kristianto, Y. Prasetyo, and S. Nugroho, “Evaluasi model numerik cuaca untuk prediksi curah hujan di wilayah tropis,” Jurnal Meteorologi dan Geofisika, vol. 19, no. 3, pp. 145-154, 2018.

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5-32, 2001.

[5] G. Fibarkah, M. A. Tondang, N. W. Yulistyaningrum, and M. Afrad, “Prediksi curah hujan menggunakan model Random Forest di wilayah pesisir,” Jurnal Meteorologi dan Geofisika, vol. 24, no. 2, pp. 87-98, 2024.

[6] H. N. Irmanda, Ermatita, M. K. Awang, and M. Adrezo, “Enhancing weather prediction models through the application of Random Forest method,” International Journal on Informatics Visualization, vol. 8, no. 3, pp. 1506-1514, 2024.

`

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.

[9] M. Taqiyuddin and T. B. Sasongko, "Prediksi cuaca menggunakan algoritma Random Forest," Jurnal Media Informatika Budidarma, vol. 8, no. 3, pp. 1683-1692,

[10] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," International Journal of Forecasting, vol. 39, pp. 388-427, 2023.