JURNAL EMACS
e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.8 No.3 September 2026: 277-292 DOI: 10.21512/emacsjournal.v8i3.16835

Machine Learning Estimation of Central Java Monthly Rainfall from Bias-Corrected NASA POWER

Christian Cahyono[1*], Denik Sri Krisnayanti[1,2], Dyah Erny Herwindiati[1,3]

[1]Doctorate Program in Civil Engineering, Faculty of Engineering,
Tarumanagara University, Jakarta 11440, Indonesia
[2]Department of Civil Engineering, Faculty of Science and Engineering,
Nusa Cendana University, Kupang 85001, Indonesia
[3]Department of Informatics, Faculty of Information Technology,
Tarumanagara University, Jakarta 11440, Indonesia
christian.328251003@stu.untar.ac.id; denik.krisnayanti@staf.undana.ac.id; dyahh@fti.untar.ac.id
*Correspondence: christian.328251003@stu.untar.ac.id

Abstract *– Monthly rainfall underpins water availability, reservoir, and irrigation analyses, yet rain gauge records in Indonesia are sparse and often interrupted. This study compares multiple linear regression (MLR), eXtreme Gradient Boosting (XGBoost), and random forest (RF) for estimating monthly rainfall at nine stations in Central Java. NASA POWER daily rainfall was first bias-corrected at each station by fitting bin-wise linear equations between the gauge and satellite probability-duration curves, a piecewise form of empirical quantile mapping, using only data from before 2017. The corrected monthly series was the modeling target, and NASA POWER air temperature, relative humidity, and wind speed were the predictors; no rainfall variable was used as a predictor, although gauge information enters through the corrected target. The models were trained on a pooled sample of Java stations and verified against gauge observations from 2017–2019 and 2024–2026 at the same stations, so the verification is temporal only. The correction raised the mean monthly gauge–satellite correlation in the verification months from 0.81 to 0.84. On 178 station-months at six stations that were not used in training, the pooled RMSE was 126, 111, and 103 mm/month for MLR, XGBoost, and RF, equal to 69%, 61%, and 56% of the mean observed rainfall. Ensemble scores at three Semarang stations were optimistic because 108 of their verification months were also training months. The ensembles outperformed MLR, but the gap between XGBoost and RF was small and untested for significance. Independent and spatial validation is needed before operational use.*

Keywords: *monthly rainfall; NASA POWER; quantile mapping; XGBoost; random forest*

I. INTRODUCTION

Rainfall is the main input for almost every water resources calculation. Design flood analysis requires short-duration data, whereas water availability, reservoir reliability, drought indices, and irrigation water demand are usually computed from monthly totals. Monthly aggregation smooths the irregularity of daily data but maintains the seasonal pattern that governs reservoir storage behavior.

In Indonesia, the data problem is clear. The Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) distributes station records through its Data Online portal (BMKG, 2026a), but only as daily values, with limited download ranges and frequently interrupted series. Users who need monthly totals must sum daily data that are not necessarily complete. A gap of one week removes almost a quarter of the days in a month; unless it is filled, the monthly total is biased low by the rain that fell on those days, which in the wet season can be a large share of the total.

The usual remedy for sparse gauges is satellite and reanalysis rainfall. TRMM data have been used to drive rainfall-runoff models in three Javanese catchments, including the Bengawan Solo, although simulations driven by gauge data remained better (Suroso et al., 2023). GPM IMERG performs adequately for rainfall-runoff modeling in Indonesia (Hidayah et al., 2024), but its skill drops sharply for extreme daily rainfall (Ramadhan et al., 2022). An evaluation of GSMaP version 08 against hundreds of Indonesian gauges gave a similar picture: accuracy improved as the accumulation period lengthened, from low at the hourly scale to good at the monthly scale (Ramadhan et al., 2023). The same pattern appears when products are compared through hydrological simulation (Ougahi & Mahmood, 2022), and even ERA5 still carries its largest random errors in the tropics (Lavers et al., 2022). Central Java lies in the monsoonal rainfall region of

Copyright © 2026

Indonesia, with one wet season peaking around December–February and a dry season around July–August, and the year-to-year departures from this cycle are linked to sea surface temperature anomalies such as those of ENSO (Aldrian & Susanto, 2003).

This study uses NASA POWER, which distributes meteorological variables from the MERRA-2 reanalysis on a 0.5° × 0.625° latitude–longitude grid as daily series from 1981 onward (NASA Langley Research Center, 2026). Its rainfall variable, PRECTOTCORR, is the MERRA-2 precipitation corrected toward observation-based precipitation products (Reichle et al., 2017), and it is supplied together with air temperature, relative humidity, and wind speed on the same grid, so the predictors and the target are aligned and complete at every station. The weaknesses of the product are also well documented. Against tropical gauges, NASA POWER reproduces the general climate reasonably well, but its rainfall intensity distribution is wrong and its highest values are underestimated, so bias correction is recommended before the data are used for extreme-event analysis (Tan et al., 2023). In a Mediterranean evaluation, its air temperature and relative humidity agreed reasonably well with station observations, whereas its wind speed did not (Rodrigues & Braga, 2021).

Two conclusions recur across these studies. First, gridded rainfall is a workable option where gauges are sparse. Second, it must be corrected against ground observations before use. Most corrections belong to the quantile-mapping family, which matches the distribution of the gridded values to that of the gauge (Themeßl et al., 2011; Gudmundsson et al., 2012); its empirical and parametric variants differ in how well they preserve changes in quantiles and extremes (Cannon et al., 2015). In Indonesia, linear scaling and distribution-based methods provided the most consistent improvements for GPM IMERG (Suhartanto et al., 2025; Tuasikal et al., 2026), and in the Banjir Kanal Timur watershed, Semarang, a random forest correction performed best among the five methods tested (Lufira et al., 2026). The correction used here pairs the gauge and NASA POWER probability-duration curves and fits a linear equation in each 5 mm/day bin, so it is a piecewise-linear form of empirical quantile mapping. A model trained on a target whose distribution is too narrow inherits the same property, which is why the raw rainfall is not modeled directly.

Why, then, estimate rainfall from temperature, humidity, and wind speed when NASA POWER already provides rainfall for the same cell? Two reasons motivate the design. First, the corrected product can be built only where a gauge record exists, because the bin equations need gauge data, and it still inherits the day-to-day timing errors of the reanalysis rainfall. Second, the surface variables are also measured at every BMKG synoptic station, so a model that maps them to monthly rainfall could fill gaps in a station's rainfall record or, after spatial validation, serve stations that report surface meteorology but no reliable rainfall (Cahyono et al., 2026). This study tests only the first use, temporal estimation at gauged stations. Whether the models add skill over the corrected product itself or over a station climatology is not settled here and is treated as a limitation (Section III.J).

Tree-based ensembles have become a common choice for tabular data of this kind. Random forest averages many decorrelated regression trees (Breiman, 2001), whereas XGBoost adds regularized trees one after another (Chen & Guestrin, 2016). The two have been compared for monthly rainfall and maximum 24 h rainfall in Iran, with XGBoost slightly ahead (Kashefi et al., 2026). Tree ensembles have also been used to merge satellite and gauge data (Papacharalampous et al., 2023), while a cross-climate comparison showed that tropical rainfall is harder to predict from surface weather variables than temperate rainfall (Ejike et al., 2025). Simple regression should not be dismissed either: for monthly rainfall at 34 gauges in the Karkheh basin, linear and multiple linear regression performed similarly to a support vector machine tuned by simulated annealing, and simple linear regression was preferred because it has fewer adjustable parameters (Sedghnejad et al., 2024).

NASA POWER variables, including its rainfall, have been used as inputs to hybrid deep-learning models for multi-day rainfall prediction at a single site in Penang (Saleh et al., 2024). In Indonesia, a CHIRPS-based ensemble has been applied in the Bengawan Solo watershed, Central Java, with XGBoost as the best single model (Jumadi et al., 2025), and random forest has served as a bias-correction method for GPM IMERG in Semarang (Lufira et al., 2026). The present study differs from these works and from Kashefi et al. (2026) in three respects: the target is NASA POWER rainfall corrected against the local gauge distribution rather than gauge or raw satellite rainfall; no rainfall variable is used as a predictor; and all three methods are verified against gauge observations in months after the correction period, with the months that also entered training reported separately. The contribution is thus a controlled benchmark of linear and tree-based learners on a gauge-corrected target, not a new algorithm.

This study has four objectives: (1) to correct NASA POWER rainfall against gauge data and measure the effect of the correction on the monthly correlation with the gauges in the verification months; (2) to build MLR, XGBoost, and random

In Press

forest models on the same corrected target and meteorological predictors; (3) to verify the three models with continuous and categorical metrics, separating the verification months that were also used in training from those that were not; and (4) to identify which method gives the most reliable monthly totals and intensity classes at gauged stations, and what limits that conclusion.

II. METHODS

A. Study Area and Data

This study used nine rain gauge stations in Central Java Province (Figure 1 and Table 1). Six are BMKG stations: Ahmad Yani, Banjarnegara, Klimatologi Jawa Tengah, Maritim Tanjung Emas, Maritim Tegal, and Tunggul Wulung. The other three, Gunungan, Pabelan, and Parangjoho, are rain gauges in the upper Bengawan Solo operated by the Bengawan Solo River Basin Organization (BBWS Bengawan Solo). The stations range from the northern coastal lowlands of Semarang and Tegal and the southern coast at Cilacap to the mountainous interior of Banjarnegara and the upper Bengawan Solo. Ahmad Yani, Klimatologi Jawa Tengah, and Maritim Tanjung Emas lie within 6 km of one another in Semarang.

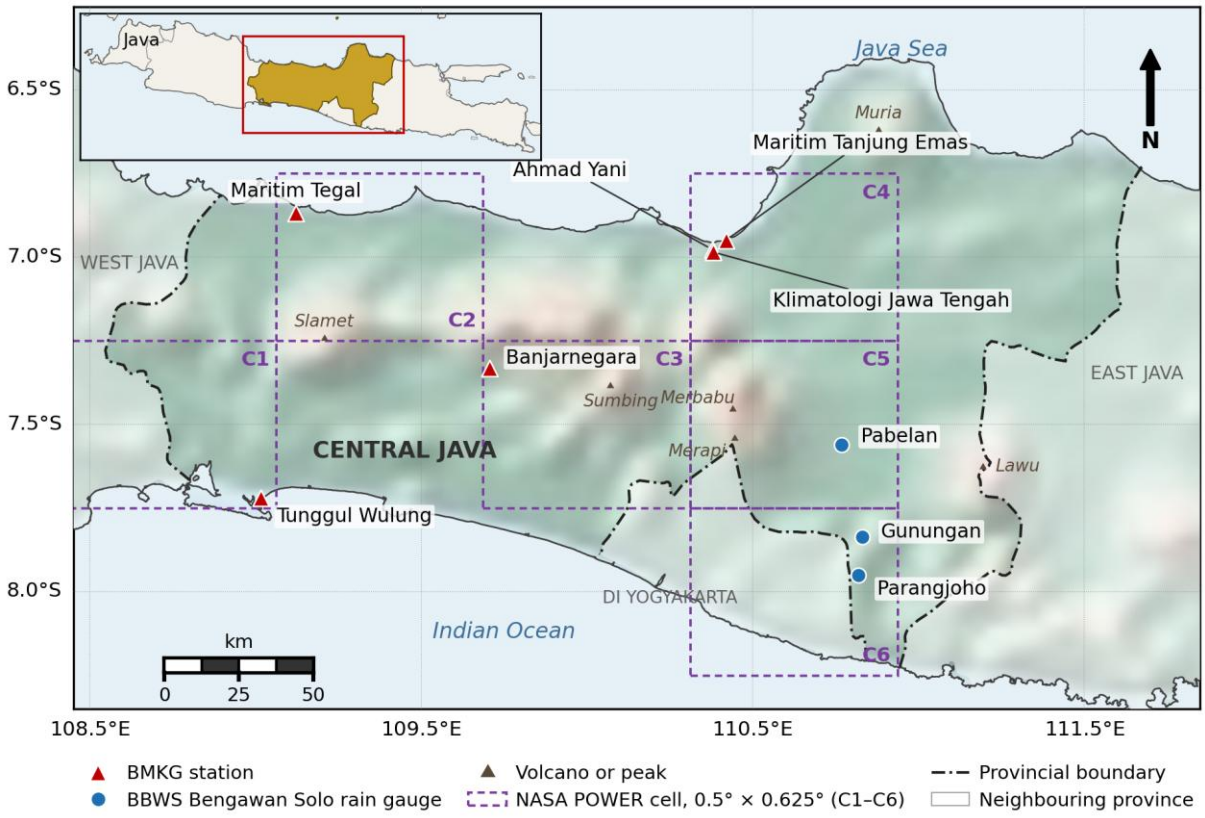


Figure 1. Locations of the Nine Rain Gauge Stations in Central Java. Triangles mark BMKG stations and circles mark BBWS Bengawan Solo rain gauges. Dashed boxes are the NASA POWER grid cells (0.5° × 0.625°) that contain the stations (C1–C6); the three Semarang stations share cell C4, and Gunungan and Parangjoho share cell C6. Boundaries: Natural Earth; relief: shadedrelief.com (both public domain).

Two groups of data were used. The first group was the daily gauge records, used as reference data for the bias correction and for verification. The BMKG records were obtained from BMKG Data Online (BMKG, 2026a) and the upper Bengawan Solo records from BBWS Bengawan Solo. All series were rebuilt on a continuous daily calendar so that no derived values bridged the data gaps, and months with 10% missing day are excluded from training. The second group was NASA POWER daily data retrieved through the POWER API (NASA Langley Research Center, 2026): precipitation (PRECTOTCORR), air temperature at 2 m (T2M), relative humidity at 2 m (RH2M), and wind speed at 2 m (WS2M). The API returns the value of the 0.5° × 0.625° grid cell that contains each station, without spatial interpolation. Daily values were summed (rainfall) or averaged (other variables) into monthly values.

The nine stations fall in six NASA POWER cells (C1–C6 in Figure 1 and Table 1). The three Semarang stations share cell C4, and their monthly predictor values are identical in every verification month; Gunungan and Parangjoho share cell C6, and

their values are identical in 34 of 36 months. For these stations, differences between estimates can arise only from the station coordinates used by the ensembles (Section II.D) and from the target, which is corrected separately at each station.

Verification used the months for which gauge observations were available for this study: January 2017 to December 2019 at the three Semarang stations and the three upper Bengawan Solo gauges, and March 2024 to January or February 2026 at the BMKG stations (Table 1). Gauge data from 2020 to 2023 were not used.

Table 1. Rain Gauge Stations, NASA POWER Cells, Verification Periods, and Station-Months

Station	Operator	Lat. (°S)	Lon. (°E)	Cell	Verification period	n	n_{ind}	Mean obs. (mm/month)
Ahmad Yani	BMKG	6.977	110.378	C4	Jan 2017–Dec 2019; Mar 2024–Jan 2026	59	23	194.4
Banjarnegara	BMKG	7.333	109.707	C3	Mar 2024–Jan 2026	23	23	269.0
Gunungan	BBWS	7.837	110.830	C6	Jan 2017–Dec 2019	36	36	129.5
Klimatologi Jawa Tengah	BMKG	6.985	110.381	C4	Jan 2017–Dec 2019; Mar 2024–Jan 2026	59	23	216.6
Maritim Tanjung Emas	BMKG	6.949	110.420	C4	Jan 2017–Dec 2019; Mar 2024–Jan 2026	59	23	189.2
Maritim Tegal	BMKG	6.868	109.121	C2	Mar 2024–Jan 2026	23	23	168.9
Pabelan	BBWS	7.561	110.769	C5	Jan 2017–Dec 2019	36	36	181.9
Parangjoho	BBWS	7.951	110.818	C6	Jan 2017–Dec 2019	36	36	127.3
Tunggul Wulung	BMKG	7.719	109.015	C1	Mar 2024–Feb 2026	24	24	273.1
All stations						**355**	**247**	**191.1**

Note: BBWS = BBWS Bengawan Solo. Cells are the NASA POWER grid cells in Figure 1. n is the number of verification station-months; n_{ind} is the number of those months that were not part of the training sample (Section II.B). MLR was evaluated on the n_{ind} months only, and XGBoost and random forest on all n months. Mean obs. is the mean observed monthly rainfall over the n months.

Monthly aggregation was the key design decision of this study. At the daily scale, the mismatch between a point gauge and a grid cell is large and nearly random. At the monthly scale, aggregation averages out much of this random error, so the error that remains is mainly systematic, which is the error that a correction can remove (Lufira et al., 2026; Ougahi & Mahmood, 2022).

B. Calibration, Training, and Verification Design

The study has three stages, and each uses a different set of months. (i) Bias correction: the bin equations of each station were fitted on daily gauge and NASA POWER data from years before 2017 and then applied unchanged to the full NASA POWER record, so no verification month entered the correction. (ii) Training: all three models were trained on one pooled monthly sample of Java stations, with 474 station-months for each calendar month. The target is the corrected NASA POWER rainfall, so gauge observations are not used directly in training. This sample, however, includes January 2017 to December 2019 at the three Semarang stations, so 108 of the 355 verification station-months were also training months (Table 1). (iii) Verification: the final models were applied to the verification months and compared with the gauge observations. These are hold-out predictions of the final models, not the out-of-fold predictions of the cross-validation used for tuning.

MLR was evaluated only on the 247 months outside the training sample, whereas XGBoost and random forest were evaluated on all 355 months. To compare the three methods on the same independent sample, the main comparison uses the six stations whose verification months all lie outside the training sample (178 station-months; Section III.D). Because stations can also contribute training months from other years, the verification tests the models in time, not at unseen locations.

C. Bias Correction with the Probability-Duration Curve

The correction acts on the distribution of the data rather than the mean and is carried out in three steps. In the first step, over the calibration years, the daily gauge series and the daily NASA POWER series at a station are sorted from largest to smallest, and their exceedance probabilities are computed. The result is two probability-duration curves on the same axis. The two curves lie close together in the light-rain part, where most days fall, and then move apart toward small exceedance probabilities, where the satellite curve lies below the gauge curve. This gap is the visible form of the underestimation of heavy-rain days reported for NASA POWER in the tropics (Tan et al., 2023).

In the second step, the rainfall range is divided into 5 mm/day bins because the gap between the two curves is not uniform and a single multiplier is not sufficient. In each bin, a linear relation is fitted

between values of the two curves paired at equal exceedance probability,

$$y = a \cdot x + b \quad (1)$$

where x is the raw daily NASA POWER rainfall (mm/day), y is the daily gauge rainfall at the same exceedance probability (mm/day), a is the slope of the bin, and b is its intercept. The top bin is left open so that the correction also applies above the last fitted bin. Pairing sorted values in this way makes the method a piecewise-linear form of empirical quantile mapping (Themeßl et al., 2011; Gudmundsson et al., 2012).

In the third step, the bin equations are applied to the raw daily series, corrected values below zero are set to zero, and the results are summed to obtain monthly rainfall, which is the modeling target. The fitted equations for the three upper Bengawan Solo stations are listed in Table 2, and those of the other six stations are given in the Supplementary Material. Because each bin is fitted separately, the transfer function is not forced to be continuous or monotonic at the bin edges; this is examined in Section III.B.

Table 2. Bias-Correction Equations for Each Daily Rainfall Bin at Gunungan, Pabelan, and Parangjoho

Bin (mm/day)	Gunungan	Pabelan	Parangjoho
0–5	$y = 0$	$y = 0.1681x - 0.0744$	$y = 0$
5–10	$y = 0.7917x - 4.7135$	$y = 1.3369x - 5.5913$	$y = 0.752x - 4.3852$
10–15	$y = 1.4584x - 10.409$	$y = 2.0845x - 12.108$	$y = 1.418x - 9.8122$
15–20	$y = 2.5029x - 25.754$	$y = 1.8478x - 8.4091$	$y = 2.1176x - 20.004$
20–25	$y = 1.6903x - 11.044$	$y = 1.1083x + 8.4911$	$y = 2.0648x - 20.108$
25–30	$y = 1.4109x - 1.5178$	$y = 1.0334x + 11.302$	$y = 0.9256x + 10.016$
30–35	$y = 3.1034x - 54.177$	$y = 2.4066x - 29.628$	$y = 3.2883x - 62.391$
35–40	$y = 3.9414x - 86.752$	$y = 3.6857x - 77.483$	$y = 2.6665x - 41.554$
40–45	$y = 7.2222x - 225.59$	$y = 2.095x - 14.516$	–
45–50	–	$y = 0.0672x + 74.941$	–
Open bin	> 45: $y = 0.2866x + 85.129$	> 50: $y = 1.4364x + 6.069$	> 40: $y = 1.9551x + 0.7161$

Note: x is the raw daily NASA POWER rainfall and y is the corrected daily rainfall (mm/day). Each equation applies to x within its bin; corrected values below zero are set to zero. The equations were fitted on daily data from the years before 2017. The equations for the other six stations are given in the Supplementary Material.

D. Predictors and Modeling Methods

MLR used three monthly predictors from NASA POWER: mean air temperature (T_{avg}, °C), mean relative humidity (RH_{avg}, %), and mean wind speed at 2 m (FF_{avg}, m/s). XGBoost and random forest used the same three variables plus the station latitude and longitude. No rainfall variable is used as a predictor, although gauge information enters through the corrected target (Section II.B). These variables are related to rainfall formation in the tropics, where warm, humid air favors convection, and they are routinely available from both BMKG synoptic stations and NASA POWER (Cahyono et al., 2026).

The first method is least-squares regression fitted separately for each calendar month, which provides 12 equations, each fitted on the 474 station-months of that month. Splitting by month handles the seasonal effect directly, without extra variables, and keeps every coefficient easy to interpretFor monthly rainfall RR at observation i,

$$RR_i = \beta_0 + \beta_1 T_{\text{avg},i} + \beta_2 RH_{\text{avg},i} + \beta_3 FF_{\text{avg},i} + \varepsilon_i \quad (2)$$

where β_0 is the intercept, β_1–β_3 are the predictor coefficients, and ε_i is an unexplained residual.

The second and third methods are XGBoost (Chen & Guestrin, 2016) and random forest (Breiman, 2001), both also fitted as 12 separate monthly models. The hyperparameters of both models were selected using random search with GroupKFold cross-validation grouped by year, so that a whole year is held out at a time, and the random seed was fixed so that the results can be reproduced.

XGBoost builds an ensemble of K regression trees. Each new tree is trained on the residual error of the previous ensemble, so the estimate for observation i is the sum of the outputs of all trees,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F} \quad (3)$$

where f_k is the k-th tree and $\mathcal{F}$ is the space of regression trees. The trees are built by minimizing an objective function that balances goodness of fit against model complexity,

$$\text{Obj} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \qquad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (4)$$

where l is the loss function, T is the number of leaves, w is the vector of leaf weights, and γ and λ are the regularization parameters. Because the estimate is a sum, several trees that all push a month upward can add up to a high value.

In contrast, random forest averages B regression trees. Each tree is grown on a bootstrap sample, and

each split may use only a random subset of the predictors, so the trees do not copy one another, and averaging reduces the variance of the result (Breiman, 2001),

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (5)$$

where T_b is the b-th tree. The same averaging pulls the estimates toward the center of the training distribution, so very large and rare values tend to be underestimated (Ejike et al., 2025; Kashefi et al., 2026).

E. Evaluation Criteria

Model performance was assessed using continuous metrics between observed monthly rainfall O_i and estimated monthly rainfall P_i over n verification months. RMSE measures the typical size of the error and gives more weight to large errors, whereas MAE reports the mean error without that weighting.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (6)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (7)$$

The Pearson correlation measures the strength of the linear relationship between the two series, the Nash–Sutcliffe efficiency (Nash & Sutcliffe, 1970) compares the model with the observed mean as a predictor, and the relative bias measures over- or underestimation of the total.

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(P_i - \overline{P})^2}} \quad (8)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (9)$$

$$\mathrm{RB} = 100 \times \frac{\sum_{i=1}^{n}(P_i - O_i)}{\sum_{i=1}^{n} O_i} \quad (10)$$

Two further measures were added. The normalized RMSE expresses the error as a percentage of the mean observation, so that stations and regions with different rainfall amounts can be compared, and the ±20% band is the percentage of months whose error does not exceed 20% of the observation,

$$\mathrm{nRMSE} = 100 \times \frac{\mathrm{RMSE}}{\overline{O}} \quad (11)$$

$$\mathrm{B}_{20} = \frac{100}{n}\sum_{i=1}^{n} I(|P_i - O_i| \le 0.2O_i) \quad (12)$$

where $I(\cdot)$ equals 1 when the condition holds and 0 otherwise. NSE > 0.5 is used as a usability threshold, following the guideline commonly applied to monthly hydrological simulations (Moriasi et al., 2007). The ±20% band is a strict criterion in dry months, where an error of a few millimeters already exceeds 20% of the observation.

Classification performance was assessed separately. Observed and estimated values were mapped into four monthly intensity classes adapted from the BMKG categories (BMKG, 2026b): low (< 100 mm), moderate (100–300 mm), high (300–400 mm), and very high (> 400 mm). The first two classes follow BMKG; BMKG places the boundary between its high and very high classes at 500 mm, whereas this study used 400 mm. For each class k, the probability of detection (POD) and the critical success index (CSI) were computed from the hits H_k (observed and estimated in class k), misses M_k (observed in k but estimated in another class), and false alarms F_k (estimated in k but observed in another class), and the four-class hit rate (HR) is the percentage of months placed in the correct class,

$$\mathrm{POD}_k = \frac{H_k}{H_k + M_k} \quad (13)$$

$$\mathrm{CSI}_k = \frac{H_k}{H_k + M_k + F_k} \quad (14)$$

$$\mathrm{HR} = \frac{100}{n}\sum_{i=1}^{n} I\big(c(P_i) = c(O_i)\big) \quad (15)$$

where $c(\cdot)$ denotes the class of a value. Both types of metrics are reported together because RMSE rewards safe estimates near the center, whereas class accuracy rewards decisiveness near class boundaries. Pooled metrics combine all calendar months, so the seasonal cycle alone raises R and NSE; station-months are also correlated in time and, for stations that share a cell, in space, so the effective sample size is smaller than n. No significance tests were applied, and differences of a few percent between methods are not interpreted as real differences.

III. RESULTS AND DISCUSSION

Bias correction raised the mean monthly correlation between the gauges and NASA POWER modestly (Table 3). On station-months outside the training sample, both tree ensembles had lower errors than MLR, and random forest had the lowest overall error (Table 5), but the relative errors remained large, 56–69% of the mean observed rainfall, and the ranking varied between stations (Table 7). The scores of the ensembles at the three Semarang stations are optimistic because their 2017–2019 months were also in the training sample. The subsections below first examine how

In Press

independent the verification is and then each result in turn.

A. Independence of the Verification Sample

Three features of the design determine how independent the verification is. First, the bias correction was fitted only on years before 2017, so no verification month entered the target through the correction. Second, 108 of the 355 verification station-months (Ahmad Yani, Klimatologi Jawa Tengah, and Maritim Tanjung Emas, January 2017 to December 2019) were also in the training sample. For these months, the ensembles were evaluated on months whose predictors and target they had already seen in training, which explains their near-perfect correlations of 0.94–0.98 at these stations (Table 7); MLR was evaluated only on the independent months. Third, the Semarang stations share one NASA POWER cell, and Gunungan and Parangjoho share another (Figure 1), so their predictor values are identical or nearly so, and their errors are not independent.

These features also explain an apparent inconsistency between Tables 3 and 7: at several stations, the models correlate more strongly with the gauge than the corrected product that they were trained to reproduce. At the Semarang stations, the cause is the overlap with the training sample. At stations without overlap, such as Pabelan (corrected product 0.80; models 0.86–0.90) and Gunungan (0.77; 0.80–0.83), a likely cause is that the model estimates are smoother than the corrected product: they follow the seasonal cycle carried by the predictors, whereas the corrected product also carries month-to-month noise from the daily reanalysis rainfall, and a correlation pooled over 23–36 months is dominated by the seasonal cycle. The same effect explains why the pooled correlation of MLR on the verification months (0.70) is higher than its within-month R^2 values suggest (mean 0.275, Table 4): the equations are fitted within one calendar month, where the seasonal cycle is absent. Skill relative to a station climatology was not computed, so part of the reported R and NSE may reflect the seasonal cycle alone.

B. Effect of Bias Correction on Monthly Correlation

The coefficients in Table 2 show how the correction reshapes the distribution. In the lowest bin, the slope is far below one: zero at Gunungan and Parangjoho and only 0.1681 at Pabelan. This means that the raw data produce many light-rain values on days that are dry at the gauge, and the correction suppresses them; at Gunungan and Parangjoho, all raw values below 5 mm/day are set to zero, which lowers dry-season totals wherever the raw rainfall falls mostly as light rain. From the second bin upward, the situation reverses. The slope rises above one and reaches 3.9414 at Gunungan and 3.6857 at Pabelan in the 35–40 mm/day bin, so large NASA POWER values must be multiplied several times to match the gauge depth at the same exceedance probability. This is the numerical form of the underestimation of heavy rain (Tan et al., 2023), and it is the weakness that the correction is meant to fix.

The bin-wise fit also has weaknesses. Because each bin is fitted separately, the transfer function is neither continuous nor monotonic at every bin edge. At Gunungan, x = 20 mm/day gives 24.3 mm/day from the 15–20 mm/day equation but 22.8 mm/day from the 20–25 mm/day equation, and x = 40 mm/day gives 70.9 mm/day from the 35–40 mm/day equation but 63.3 mm/day from the 40–45 mm/day equation. At the lower edge of the 5–10 mm/day bin, the equations return small negative values (−0.76 mm at Gunungan and −0.63 mm at Parangjoho), which are set to zero. Steep upper-bin slopes, such as 7.2222 at Gunungan for 40–45 mm/day, are likely fitted to few points and can amplify single heavy days. The correction keeps the rank order within each bin but not necessarily across bin edges, and a monotone, continuous transfer function, such as interpolated empirical quantile mapping (Gudmundsson et al., 2012), would be preferable.

Table 3 presents the effect of the correction on the monthly correlation between the gauge and satellite data in the verification months. Averaged through Fisher's z transformation, the correlation increased from 0.81 to 0.84 (arithmetic means 0.762 and 0.810). Correlation increased at five stations, and the largest gains occurred at the stations with the lowest initial correlation. Maritim Tegal increased by 0.178, from 0.429 to 0.607; Parangjoho by 0.119; Gunungan by 0.092; and Pabelan by 0.065. At the other five stations, whose initial correlation was mostly already high, the change ranged only from −0.016 to +0.011. With 23–59 months per station, changes of a few hundredths lie within sampling variability, and no significance test was applied.

IN PRESS

Table 3. Monthly Correlation between Gauge and NASA POWER Rainfall before and after Bias Correction in the Verification Months

Station	*n*	R before	R after	ΔR
Tunggul Wulung	24	0.893	0.904	+0.011
Banjarnegara	23	0.723	0.707	−0.016
Klimatologi Jawa Tengah	59	0.946	0.943	−0.002
Ahmad Yani	59	0.935	0.929	−0.006
Maritim Tanjung Emas	59	0.863	0.860	−0.003
Maritim Tegal	23	0.429	0.607	+0.178
Gunungan	36	0.676	0.768	+0.092
Pabelan	36	0.734	0.799	+0.065
Parangjoho	36	0.656	0.775	+0.119
Arithmetic mean		0.762	0.810	+0.049
Mean (Fisher *z*)		**0.810**	**0.838**	**+0.028**

Note: R is the Pearson correlation between monthly gauge totals and monthly NASA POWER totals at the same station over the verification months of Table 1 (*n* months). The correction was fitted on years before 2017, so these months are independent of the correction. The Fisher-*z* mean back-transforms the average of artanh(R).

Why does a distribution correction change a monthly correlation at all? The bin equations act on daily values. Suppressing many light-rain days and amplifying heavy days changes the relative size of monthly totals: a month with several heavy days gains much more than a month with many light days. The monthly series can co-vary differently with the gauge even though the daily ranks within each bin are kept. Where the raw series had many spurious light-rain days and suppressed heavy days, as at Maritim Tegal and the upper Bengawan Solo gauges, this re-weighting brought the monthly totals closer to the gauge pattern. Correlation is, however, not the quantity that the correction is designed to fix, and its effects on bias, RMSE, wet-day frequency, and heavy-rain statistics were not evaluated here (Section III.J). The pattern agrees with the results of distribution-based corrections in the Banjir Kanal Timur watershed, Semarang (Lufira et al., 2026), and the Welang watershed (Suhartanto et al., 2025).

C. Monthly Multiple Linear Regression Equations

Twelve MLR equations, one for each month, were fitted to the pooled training sample in the form of Equation (2), each on 474 station-months. The coefficients and coefficients of determination are listed in Table 4. For example, the October equation is $RR = -792.57 + 11.20T_{avg} + 9.37RH_{avg} - 46.44FF_{avg}$, with $R^2 = 0.443$. Table 4 offers three clues to why MLR lags behind the two ensembles.

First, the explanatory power is low. The R^2 ranges from 0.096 in March to 0.443 in October (adjusted R^2 0.090–0.439), with a mean of 0.275. In other words, the three surface predictors explain on average only about 28% of the within-month variance in monthly rainfall.

Second, the explanatory power changes with the season and is weakest when it is most needed. In the wet months of December to March, the mean R^2 is only 0.180, whereas in the dry and transition months of June to November it rises to 0.349. However, the wet months contribute most of the rainfall. This is consistent with MLR missing the peaks of the rainy season, as seen later in Figure 3.

Third, the direction of the predictor effects is consistent, but their magnitudes are unstable. The relative humidity coefficient is positive in all months, and the air temperature coefficient is positive in 11 of 12 months, whereas the wind speed coefficient is negative in 11 of 12 months. These directions agree with the general understanding that humid, warm air favors rainfall. However, the February coefficients jump sharply, with an intercept of −11,551.75 and a temperature coefficient of 125.93. With 474 observations per equation, this is unlikely to be a small-sample effect; it more likely reflects collinearity between temperature and humidity in the pooled sample, which inflates the coefficients without improving the fit. Variance inflation factors were not computed in this study.

In Press

Table 4. Monthly Multiple Linear Regression Equations Fitted on the Pooled Training Sample

Month	R^2	Adj. R^2	Intercept (β_0)	β_1 (T_{avg})	β_2 (RH_{avg})	β_3 (FF_{avg})
Jan	0.115	0.109	−5,393.43	55.52	49.78	10.30
Feb	0.364	0.360	−11,551.75	125.93	100.44	−4.89
Mar	0.096	0.090	−1,306.09	0.56	18.10	−3.61
Apr	0.208	0.203	−2,900.89	32.29	27.08	−32.48
May	0.280	0.275	−3,315.05	36.78	29.13	−4.17
Jun	0.330	0.326	−1,654.39	23.94	13.93	−23.58
Jul	0.316	0.312	−1,098.84	19.46	8.82	−21.77
Aug	0.284	0.279	−292.05	6.45	2.36	−12.82
Sep	0.306	0.302	−706.67	15.31	5.96	−32.58
Oct	0.443	0.439	−792.57	11.20	9.37	−46.44
Nov	0.417	0.413	−666.04	−1.15	12.15	−39.40
Dec	0.143	0.138	−4,624.42	68.23	39.21	−74.11

Note: The equations follow $RR = \beta_0 + \beta_1 T_{avg} + \beta_2 RH_{avg} + \beta_3 FF_{avg}$, with RR in mm/month, T_{avg} in °C, RH_{avg} in %, and FF_{avg} in m/s. Each equation was fitted on $n = 474$ station-months; Adj. R^2 is the adjusted coefficient of determination.

D. Comparison on Independent Months

Table 5 compares the three methods on the same sample: the 178 station-months at the six stations whose verification months were all outside the training sample. The pooled RMSE was 126.2 mm/month for MLR, 111.3 mm/month for XGBoost, and 102.9 mm/month for random forest, that is, 69%, 61%, and 56% of the mean observed rainfall of 182.1 mm/month. Random forest reduced the RMSE of MLR by 23.4 mm/month (18.5%), and XGBoost reduced it by 15.0 mm/month (11.9%). The median station NSE was 0.41, 0.56, and 0.57, and the NSE exceeded 0.5 at two, four, and four of the six stations, respectively. Random forest gave the lowest RMSE at four stations and XGBoost at two.

Table 5. Comparison of the Three Methods on Independent Station-Months at Six Stations

Method	n	Pooled RMSE (mm/month)	RMSE / mean obs. (%)	Mean station R	Median station NSE	Stations with NSE > 0.5	Stations with lowest RMSE
MLR	178	126.2	69.3	0.67	0.41	2	0
XGBoost	178	111.3	61.1	0.75	0.56	4	2
Random forest	178	102.9	56.5	0.76	0.57	4	4

Note: Stations: Banjarnegara, Gunungan, Maritim Tegal, Pabelan, Parangjoho, and Tunggul Wulung, whose verification months were all outside the training sample. Pooled RMSE = $(\Sigma n_s \mathrm{RMSE}_s^2 / \Sigma n_s)^{1/2}$ over the station values in Table 7; the mean observed rainfall of this sample is 182.1 mm/month. No confidence intervals were computed.

The advantage of the ensembles over MLR is real but much smaller than the all-month comparison suggests (Section III.E), and the gap between XGBoost and random forest (8.4 mm/month) rests on six stations with 23–36 months each. Without a block bootstrap or repeated runs with different seeds, this gap cannot be called significant. The relative errors of 56–69% also show that none of the methods yet gives accurate monthly totals at these stations.

E. Overall Metrics on All Verification Months

Figure 2 and Table 6 show the metrics on all verification months: 247 months for MLR and 355 for the ensembles. On this sample, random forest and XGBoost reached RMSE values of 79.8 and 87.1 mm/month and NSE values of 0.78 and 0.74, against 125.3 mm/month and 0.42 for MLR. These ensemble values are optimistic. They include the 108 Semarang months that were also training months, at which stations the ensembles reached NSE values of 0.87–0.95 over all 59 months (Table 7), and they come from a different sample than the MLR values. They should not be read as the accuracy expected in independent months; Table 5 gives the fair comparison.

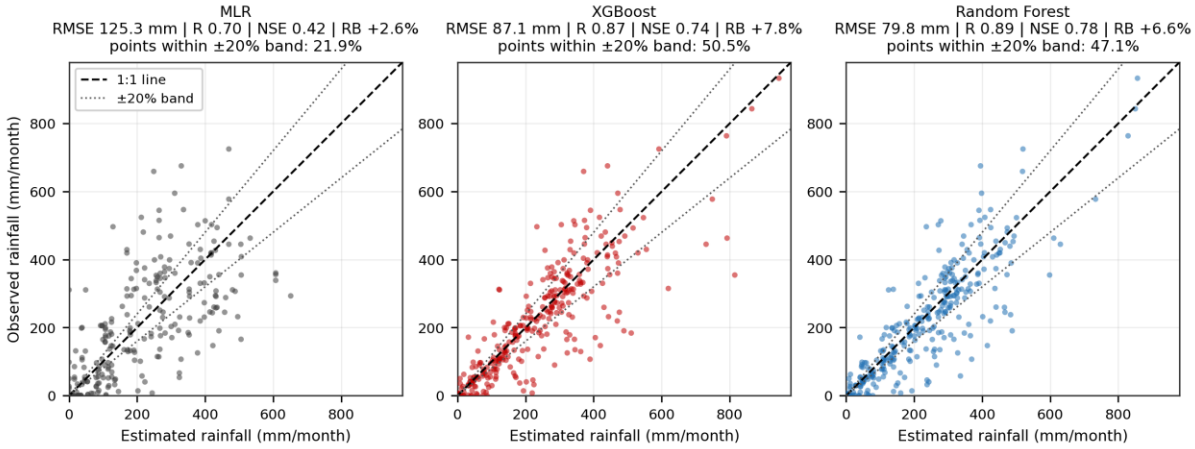


Figure 2. Estimated versus Observed Monthly Rainfall for MLR, XGBoost, and Random Forest on All Verification Months. Estimates are on the horizontal axis and observations on the vertical axis (Piñeiro et al., 2008). The dashed line is the 1:1 line, and the dotted lines bound the ±20% band. The MLR panel contains the 247 independent months; the XGBoost and random forest panels contain 355 months, including 108 training months at the three Semarang stations, so their panel statistics are optimistic (Section III.E).

Table 6. Overall Metrics on All Verification Months, Including Training Months for the Ensembles

Method	*n* (incl. training months)	RMSE (mm/month)	MAE (mm/month)	R	NSE	RB (%)	Months within ±20% band (%)
MLR	247 (0)	125.3	92.5	0.70	0.42	+2.6	21.9
XGBoost	355 (108)	87.1	52.2	0.87	0.74	+7.8	50.5
Random forest	355 (108)	79.8	50.5	0.89	0.78	+6.6	47.1

Note: MLR was evaluated on the 247 independent months and the ensembles on all 355 months, including 108 months at the three Semarang stations that were also in the training sample. The ensemble values are optimistic and are not directly comparable with the MLR values (Section III.E); Table 5 gives the like-for-like comparison.

The scatter in Figure 2 makes these numbers more concrete. In the MLR panel, the points are widely spread on both sides of the 1:1 line, and only 21.9% fall within the ±20% band. A group of points with estimates of 200–600 mm/month and much smaller observations is also visible. In both ensemble panels, the points cluster more tightly around the 1:1 line in the low-to-moderate range, and about half of them fall within the band. All three methods spread out above about 400 mm/month, which indicates that the remaining error is mainly a magnitude error in the heaviest months. The relative bias of all three methods is positive (+2.6%, +7.8%, and +6.6%), so all three tend to overestimate total rainfall slightly. The near-zero MLR value means that over- and underestimates cancel in the total, not that individual months are accurate.

F. Performance at Individual Stations

Table 7 lists the station results. At the six stations with independent months, random forest gave the lowest RMSE at four (Gunungan, Pabelan, Parangjoho, and Tunggul Wulung) and XGBoost at two (Banjarnegara and Maritim Tegal), whereas MLR was lowest at none. The NSE exceeded 0.5 at two, four, and four of these stations for MLR, XGBoost, and random forest, respectively. The three Semarang stations show the highest scores, with random forest NSE values of 0.90–0.95, but for the ensembles these scores include training months and are not comparable with the MLR values, which cover only March 2024 to January 2026.

Table 7. Performance of the Three Methods at Each Rain Gauge Station

Station	RMSE (mm/month)			R			NSE		
	MLR	XGB	RF	MLR	XGB	RF	MLR	XGB	RF
Ahmad Yani†	111.4	38.5	39.1	0.74	0.98	0.98	0.43	0.95	0.95
Banjarnegara	149.5	127.4	130.4	0.63	0.73	0.71	0.35	0.53	0.51
Gunungan	97.1	86.9	85.7	0.80	0.82	0.83	0.52	0.62	0.63
Klimatologi Jawa Tengah†	130.8	57.6	45.3	0.72	0.95	0.98	0.37	0.91	0.94
Maritim Tanjung Emas†	125.2	59.3	52.8	0.66	0.94	0.96	0.25	0.87	0.90
Maritim Tegal	126.2	77.2	118.9	0.71	0.77	0.63	−0.13	0.58	−0.01
Pabelan	84.1	119.7	80.2	0.88	0.86	0.90	0.75	0.49	0.77
Parangjoho	103.8	85.3	79.6	0.77	0.83	0.86	0.47	0.64	0.69
Tunggul Wulung	201.5	163.8	135.8	0.20	0.50	0.63	−0.39	0.08	0.37

Note: XGB is XGBoost and RF is random forest. †For XGBoost and random forest, the values cover 59 months, of which 36 (January 2017–December 2019) were also training months; the MLR values cover only the 23 independent months (March 2024–January 2026). At the other six stations, all three methods were evaluated on the same independent months.

Two stations deserve special mention. At Maritim Tegal, XGBoost clearly beats random forest, with an RMSE of 77.2 versus 118.9 mm/month, and random forest has an NSE of −0.01, no better than the observed mean; it is also the station with the lowest initial correlation in Table 3. At Tunggul Wulung, all three methods perform poorly (NSE −0.39 to 0.37), although the corrected product correlates strongly with the gauge there (0.904; Table 3). At Pabelan, XGBoost has a higher RMSE than MLR (119.7 versus 84.1 mm/month) because of two large overestimates (Section III.G).

Maritim Tegal and Tunggul Wulung have only 23–24 months, so their NSE and R values are unstable: one or two poorly estimated wet months can change NSE substantially. The weaker results at inland and mountain stations such as Banjarnegara, where the ensemble NSE stalls at 0.51–0.53, may reflect orographic rainfall at scales much smaller than the 0.5° × 0.625° predictor cells, but this explanation was not tested; comparing station and cell elevations would be a first check.

G. Time Series at Gunungan, Pabelan, and Parangjoho

Figure 3 shows the observed and estimated time series at three upper Bengawan Solo stations from January 2017 to December 2019, all of which are independent of training. Three features recur at all three stations. First, all three methods followed the seasonal pattern. Estimates rose with the onset of the rainy season and stayed low throughout the dry season, with no spurious rain in the dry months of June to September. This seasonal agreement accounts for most of the correlations of 0.77–0.90 at these stations (Table 7).

Second, MLR behaved differently from the ensembles in the wet season. In January 2017, MLR estimated about 355 mm/month at Gunungan and 420 mm/month at Parangjoho, whereas the observations were only about 190 and 260 mm/month, and both ensembles were closer, at about 250–280 mm/month. At the start of the 2018/2019 wet season, MLR rose earlier than the observations: in December 2018, it estimated about 330 mm/month at both stations against observations below 100 mm/month, while the ensembles gave about 180–200 mm/month. This agrees with the low wet-season R^2 of MLR in Table 4.

Third, the remaining weakness lies in the size of the peaks. At Pabelan, XGBoost overshot markedly in February 2017 and February 2018, at approximately 790 and 815 mm/month, whereas the observations were approximately 465 and 355 mm/month. Consequently, its RMSE rose to 119.7 mm/month and its NSE dropped to 0.49, even though its correlation remained 0.86. Three explanations are possible: the additive structure of boosting can let several trees push the same month upward (Section II.D), the hyperparameters selected for these months may favor such behavior, or the corrected target itself was very high in these months because the steep upper-bin slopes in Table 2 amplify heavy raw days. Figure 3 does not show the corrected NASA POWER series, so these causes cannot be separated here.

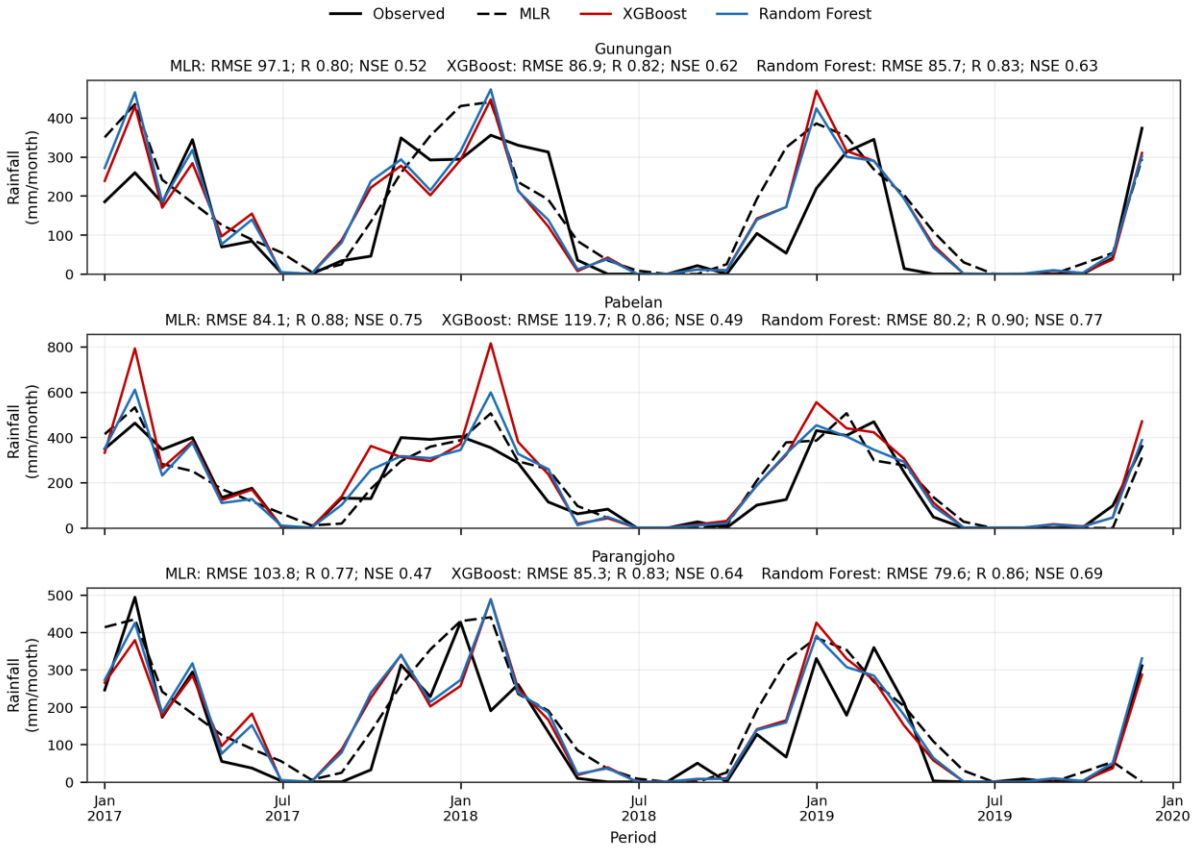


Figure 3. Observed and Estimated Monthly Rainfall at Gunungan, Pabelan, and Parangjoho, January 2017 to December 2019. Observations are shown as a solid black line, MLR as a dashed black line, XGBoost in red, and random forest in blue; all months shown are independent of training.

At Gunungan and Parangjoho, all three methods underestimated some peaks and overestimated others. This suggests that the surface predictors carry no information on the year-to-year variation in the timing and strength of the rainy season, which in Java is linked to large-scale drivers such as ENSO (Aldrian & Susanto, 2003); no climate index was included as a predictor. Random forest gave the lowest RMSE at all three stations, with values of 85.7, 80.2, and 79.6 mm/month and NSE values of 0.63, 0.77, and 0.69, respectively.

H. Monthly Rainfall Intensity Classes

Many field decisions are communicated as intensity classes rather than single numbers, so the estimates and observations were also mapped into the four classes defined in Section II.E. The four-class hit rate reached 53.8% for MLR (247 months), 75.5% for XGBoost, and 75.2% for random forest (355 months), with the POD and CSI of each class given in Table 8. As in Section III.E, the ensemble values include the 108 training months, and the difference between XGBoost and random forest corresponds to about one month, so it does not distinguish the two methods.

Table 8. Assessment of Monthly Rainfall Intensity Classes

Intensity class (mm/month)	*n*		POD			CSI		
	MLR	XGB and RF	MLR	XGB	RF	MLR	XGB	RF
Low (< 100)	87	135	0.736	0.844	0.837	0.640	0.826	0.825
Moderate (100–300)	87	125	0.552	0.776	0.784	0.338	0.591	0.587
High (300–400)	41	53	0.171	0.566	0.566	0.109	0.361	0.361
Very high (> 400)	32	42	0.438	0.643	0.619	0.255	0.474	0.464

Note: *n* is the number of observed months in each class: 247 independent months for MLR and 355 months for XGBoost and random forest, including 108 training months. POD and CSI follow Equations (13) and (14). The four-class hit rates (Equation 15) are 53.8%, 75.5%, and

75.2% for MLR, XGBoost, and random forest, respectively. Classes are adapted from BMKG (2026b), with the upper boundary at 400 mm instead of 500 mm.

The low class was handled best by both ensembles, with PODs of 0.844 and 0.837 and CSI of 0.826 and 0.825, respectively, because dry months carry a clear weather signal in the form of low humidity. In contrast, the 300–400 mm/month class was handled poorly by all three methods, especially MLR, whose POD was only 0.171. This is more of a class-boundary problem than an algorithm problem because the class is only 100 mm wide, whereas the typical model error is of the same size. Neither ensemble was consistently better: random forest had the higher POD in the moderate class (0.784 versus 0.776) and XGBoost in the very high class (0.643 versus 0.619). The advantage of the ensembles over MLR in hit rate (about 21 percentage points) is larger than on the continuous metrics, but it is computed on different samples and should be confirmed on independent months.

I. Comparison with Previous Studies

Table 9 places these results beside other published applications for context only. The studies differ in climate, period, target, spatial support, and verification design, so their numbers are not directly comparable. The monthly MLR of Cahyono et al. (2026) used BMKG station temperature, humidity, and wind speed in western Java and reported RMSE values of 162–295 mm/month; the stations, target, and verification periods differ, so the comparison shows only that both studies find a weak linear signal in these predictors. The CHIRPS-based ensemble in the Bengawan Solo watershed reached an RMSE of 69.2 mm (Jumadi et al., 2025), and a random forest correction of GPM IMERG in Semarang reached an NSE of 0.61 (Lufira et al., 2026); both used rainfall-based inputs, unlike the present models. Sabu et al. (2025) found ERA5 to be the best of eight global products in the Wae Mese watershed, but in a different climate and without local correction. The finding from Iran that XGBoost slightly outperformed random forest (Kashefi et al., 2026) was not reproduced here, but the gap between the two algorithms in this study is small and untested, so neither result contradicts the other.

Table 9. Published Applications Shown for Context

Study and region	Approach	Time step	Reported performance
Cahyono et al. (2026), western Java	Monthly MLR on BMKG station *T*, *RH*, and *FF*	Monthly	RMSE 162–295 mm/month; R 0.35–0.48
Jumadi et al. (2025), Bengawan Solo watershed	ML and DL ensembles on CHIRPS	Monthly	Stacked ensemble MAE 53.7 mm; RMSE 69.2 mm; XGBoost best single model
Lufira et al. (2026), Banjir Kanal Timur watershed, Semarang	Five bias-correction methods for GPM IMERG	Monthly	Random forest best; NSE 0.61; R 0.78
Sabu et al. (2025), Wae Mese watershed, Labuan Bajo	Assessment of eight global rainfall products	Monthly and 15-day	ERA5 best; NSE 0.63; RMSE 73.4 mm; RB 7.68%; R 0.75
This study, nine Central Java stations	MLR, XGBoost, and random forest on bias-corrected NASA POWER	Monthly	Independent months (six stations): pooled RMSE 126.2 / 111.3 / 102.9 mm/month; median station NSE 0.41 / 0.56 / 0.57

Note: ML is machine learning and DL is deep learning. The studies differ in climate, period, target, spatial support, and verification design, so the values are shown for context only and are not directly comparable.

With the target and verification months held fixed and nearly the same predictors (the ensembles also used the station coordinates), the choice of method changed the pooled RMSE on independent months by up to 18.5%. This study adds two findings. First, a bin-wise probability-duration correction applied to NASA POWER raises its monthly correlation with the gauges mainly where the raw correlation is weakest, although its effect on magnitude errors remains to be quantified. Second, on a gauge-corrected target, tree ensembles estimate monthly totals at gauged stations more accurately than monthly MLR, but with relative errors of 56–61%, they are not yet accurate enough for operational gap filling without further validation. The framework is not suitable for design flood estimation or warnings of extreme rainfall.

J. Limitations

The results should be read with the following limitations. (1) Only nine stations were used, three of them in one NASA POWER cell in Semarang, and the verification periods are short and unequal (23–59 months). (2) The verification is temporal only; no station was withheld from training, so performance at ungauged locations is unknown. (3) 108 of the 355 verification months were also training months for the ensembles, so the like-for-like comparison rests on six stations and 178 months, and station-months are not independent in time or, for

shared cells, in space. (4) The bias correction uses gauge data, so the target is not independent of the gauges at the calibration stage; its bin-wise transfer functions are neither continuous nor monotonic at every bin edge, and its effects on bias, wet-day frequency, and extremes were not quantified. (5) The raw and corrected NASA POWER rainfall and a station climatology were not evaluated as baselines in the verification months, so the added value of the models over the corrected product is not established; at Tunggul Wulung, the corrected product correlates with the gauge far better than any model (0.904 versus 0.20–0.63), and at Banjarnegara about as well (0.707 versus 0.63–0.73). (6) No confidence intervals, significance tests, repeated seeds, variable-importance analysis, or confusion matrices were computed. (7) The predictors are three reanalysis variables at a 0.5° × 0.625° resolution, without station elevation or large-scale climate indices, and NASA POWER wind speed is less reliable than its temperature and humidity (Rodrigues & Braga, 2021).

IV. CONCLUSION

Rain gauge records in Central Java are sparse and interrupted, and this study asked whether monthly rainfall at gauged stations can be estimated from NASA POWER surface meteorology when the target is NASA POWER rainfall corrected against the local gauge distribution. Three methods, MLR, XGBoost, and random forest, were compared on the same corrected target at nine stations.

The bin-wise probability-duration correction, fitted on years before 2017, raised the mean monthly correlation with the gauges in the verification months from 0.81 to 0.84, mainly at stations with weak raw agreement, such as Maritim Tegal. The twelve monthly MLR equations explained on average only 28% of the within-month variance, and least in the wet months.

On 178 station-months that were not used in training, random forest and XGBoost reduced the pooled RMSE of MLR by 18.5% and 11.9%, to 102.9 and 111.3 mm/month, but their errors remained 56–61% of the mean observed rainfall. Scores computed on all verification months were optimistic for the ensembles because 108 Semarang months were also training months. The difference between the two ensembles was small, varied between stations, and was not tested for significance, so neither can yet be recommended over the other; MLR remains useful as a transparent benchmark.

These conclusions are limited to temporal verification at nine stations (Section III.J). Future work should (i) retrain the models without the verification months and verify them at withheld stations, (ii) evaluate the raw and corrected NASA POWER rainfall and a station climatology as baselines, (iii) replace the bin-wise equations with a monotone, continuous quantile mapping and report its effect on bias and extremes, (iv) add block-bootstrap confidence intervals, repeated seeds, and variable-importance analysis, and (v) test quantile mapping of the model outputs and site-specific predictors such as elevation and climate indices before the framework is applied to ungauged locations.

AUTHOR'S CONTRIBUTION

Conceptualization: CC;
Data curation: CC;
Formal analysis: CC;
Methodology: CC, DSK;
Validation: CC, DSK, and DEH;
Visualization: CC and DEH;
Writing: CC, DSK, and DEH.

AVAILABILITY DATA AND MATERIALS

The NASA POWER data are openly available through the POWER API (NASA Langley Research Center, 2026), and the BMKG station records through BMKG Data Online (BMKG, 2026a). The records of the three upper Bengawan Solo rain gauges are held by BBWS Bengawan Solo.

REFERENCES

Aldrian, E., & Susanto, R. D. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology, 23*(12), 1435–1452. https://doi.org/10.1002/joc.950

Badan Meteorologi, Klimatologi, dan Geofisika. (2026a). *Data Online BMKG* [Data set]. Retrieved, from https://dataonline.bmkg.go.id/

Badan Meteorologi, Klimatologi, dan Geofisika. (2026b). *Prediksi curah hujan kumulatif bulanan deterministik* [Deterministic monthly cumulative rainfall prediction]. Retrieved October 27, 2026, from https://www.bmkg.go.id/iklim/prediksi-hujan-bulanan/curah-hujan-deterministik

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cahyono, C., Baria, A. B., Pierre, A. J., & Irawan, A. P. (2026). Statistical estimation of monthly rainfall in Jakarta, South Tangerang, and Bandung: A multiple linear regression baseline using BMKG station data. *Engineering,*

IN PRESS

Technology & Applied Science Research, 16(3), 36979–36985. https://doi.org/10.48084/etasr.19328

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate, 28*(17), 6938–6959. https://doi.org/10.1175/JCLI-D-14-00754.1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Ejike, O., Ndzi, D., & Shakir, M. Z. (2025). Comparative study of machine learning-based rainfall prediction in tropical and temperate climates. *Climate, 13*(8), Article 167. https://doi.org/10.3390/cli13080167

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Technical note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods. *Hydrology and Earth System Sciences, 16*(9), 3383–3390. https://doi.org/10.5194/hess-16-3383-2012

Hidayah, E., Halik, G., & Nusantara, W. (2024). Performance of GPM-IMERG satellite precipitation for rainfall-runoff modeling in Indonesia. *Water Practice and Technology, 19*(10), 3909–3928. https://doi.org/10.2166/wpt.2024.240

Jumadi, J., Danardono, D., Roziaty, E., Ulinuha, A., Supari, S., Choy, L. K., Sattar, F., & Nawaz, M. (2025). AI-driven ensemble learning for spatio-temporal rainfall prediction in the Bengawan Solo River watershed, Indonesia. *Sustainability, 17*(20), Article 9281. https://doi.org/10.3390/su17209281

Kashefi, M., Karami, H., Niksefat, M., & Ghazvinian, H. (2026). Prediction of monthly precipitation and maximum 24 h precipitation using random forest, decision tree and XGBoost models. *Modeling Earth Systems and Environment, 12*(1), Article 70. https://doi.org/10.1007/s40808-025-02714-3

Lavers, D. A., Simmons, A., Vamborg, F., & Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society, 148*(748), 3152–3165. https://doi.org/10.1002/qj.4351

Lufira, R. D., Suhartanto, E., Andawayanti, U., Asmaranto, R., & Utami, R. T. (2026). A comparative assessment of advanced bias correction methods for GPM IMERG in Banjir Kanal Timur watershed. *Engineering, Technology & Applied Science Research, 16*(1), 32097–32103. https://doi.org/10.48084/etasr.15484

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE, 50*(3), 885–900. https://doi.org/10.13031/2013.23153

NASA Langley Research Center. (2026). *Prediction of Worldwide Energy Resources (POWER) daily point data*, from https://power.larc.nasa.gov/

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology, 10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ougahi, J. H., & Mahmood, S. A. (2022). Evaluation of satellite-based and reanalysis precipitation datasets by hydrologic simulation in the Chenab river basin. *Journal of Water and Climate Change, 13*(3), 1563–1582. https://doi.org/10.2166/wcc.2022.410

Papacharalampous, G., Tyralis, H., Doulamis, A., & Doulamis, N. (2023). Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. *Hydrology, 10*(2), Article 50. https://doi.org/10.3390/hydrology10020050

Piñeiro, G., Perelman, S., Guerschman, J. P., & Paruelo, J. M. (2008). How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling, 216*(3–4), 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006

Ramadhan, R., Marzuki, M., Yusnaini, H., Muharsyah, R., Suryanto, W., Sholihun, S., Vonnisa, M., Battaglia, A., & Hashiguchi, H. (2022). Capability of GPM IMERG products for extreme precipitation analysis over the Indonesian Maritime Continent. *Remote Sensing, 14*(2), Article 412. https://doi.org/10.3390/rs14020412

Ramadhan, R., Marzuki, M., Yusnaini, H., Muharsyah, R., Tangang, F., Vonnisa, M., & Harmadi, H. (2023). A preliminary assessment of the GSMaP version 08 products over Indonesian Maritime Continent against gauge data. *Remote Sensing, 15*(4), Article 1115. https://doi.org/10.3390/rs15041115

Reichle, R. H., Liu, Q., Koster, R. D., Draper, C. S., Mahanama, S. P. P., & Partyka, G. S. (2017). Land surface precipitation in MERRA-2. *Journal of Climate, 30*(5), 1643–1664. https://doi.org/10.1175/JCLI-D-16-0570.1

Rodrigues, G. C., & Braga, R. P. (2021). Evaluation of NASA POWER reanalysis products to estimate daily weather variables in a hot summer Mediterranean climate. *Agronomy, 11*(6), Article 1207. https://doi.org/10.3390/agronomy11061207

Sabu, M. K. H., Yudianto, D., & Wijaya, O. T. (2025). Evaluasi curah hujan berbasis data global pada DAS Wae Mese, Labuan Bajo [Evaluation of global data-based rainfall in the Wae Mese watershed, Labuan Bajo]. *Jurnal Sumber Daya Air, 21*(1), 15–30. https://doi.org/10.32679/jsda.v21i1.906

Saleh, A., Tan, M. L., Yaseen, Z. M., & Zhang, F. (2024). Integrated machine learning models for enhancing tropical rainfall prediction using NASA POWER meteorological data. *Journal of Water and Climate Change, 15*(12), 6022–6042. https://doi.org/10.2166/wcc.2024.719

Sedghnejad, N., Nozari, H., & Marofi, S. (2024). Comparative analysis of classification techniques and input-output patterns for monthly rainfall prediction. *Water Science, 38*(1), 192–208. https://doi.org/10.1080/23570008.2024.2323879

Suhartanto, E., Andawayanti, U., Hidayat, M. N., Lufira, R. D., & Utami, R. T. (2025). A comparative evaluation of bias correction techniques for improving GPM-IMERG precipitation data in the Welang watershed, Indonesia. *Engineering, Technology & Applied Science Research, 15*(6), 29935–29941. https://doi.org/10.48084/etasr.13762

Suroso, S., Santoso, P. B., Birkinshaw, S., Kilsby, C., Bardossy, A., & Aldrian, E. (2023). Assessment of TRMM rainfall data for flood modeling in three contrasting catchments in Java, Indonesia. *Journal of Hydroinformatics, 25*(3), 797–814. https://doi.org/10.2166/hydro.2023.132

Tan, M. L., Armanuos, A. M., Ahmadianfar, I., Demir, V., Heddam, S., Al-Areeq, A. M., Abba, S. I., Halder, B., Kilinc, H. C., & Yaseen, Z. M. (2023). Evaluation of NASA POWER and ERA5-Land for estimating tropical precipitation and temperature extremes. *Journal of Hydrology, 624*, Article 129940. https://doi.org/10.1016/j.jhydrol.2023.129940

Themeßl, M. J., Gobiet, A., & Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology, 31*(10), 1530–1544. https://doi.org/10.1002/joc.2168

Tuasikal, H., Juwono, P. T., Sholichin, M., & Suhartanto, E. (2026). Evaluating linear scaling and random forest bias correction of GPM-IMERG V7 rainfall in Moluccas Islands watersheds. *Engineering, Technology & Applied Science Research, 16*(2), 34706–34711. https://doi.org/10.48084/etasr.17008

In Press