iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

COMPARATIVE PERFORMANCE OF REGRESSION AND MACHINE-LEARNING MODELS FOR ESTIMATING MONTHLY RAINFALL EROSIVITY IN DATA-SCARCE SOUTH-EASTERN NIGERIA

Julius Okezue Onyeka*,
Lawrence Echefulechukwu Obi,
Stanley Chigozie Anyanwu,
Department of Civil Engineering, Faculty of Engineering, Imo State University, Owerri, Nigeria
*Corresponding author: nopsoftinc@yahoo.com

ABSTRACT
Estimating rainfall erosivity where sub-hourly rainfall records are unavailable is a major constraint on soil-loss modelling in West Africa. This study compared five models—multiple linear regression (MLR), degree-2 polynomial regression, Random Forest (RF), support vector regression (SVR) and an artificial neural network (ANN)—for estimating monthly rainfall erosivity at five stations in South-Eastern Nigeria. Erosivity was derived from hourly NASA POWER rainfall (2005–2025) as an hourly-intensity index (EIh). Predictors were monthly rainfall, maximum daily rainfall, number of rainy days, previous-month rainfall, station and calendar month. Models were trained on 988 station-months (2005–2021) and tested on a chronologically held-out set of 247 station-months (2021–2025). RF performed best (R^2 = 0.915; RMSE 36% lower than MLR), closely followed by polynomial regression (R^2 = 0.910), ANN (0.874), SVR (0.845) and MLR (0.792). A cubic polynomial overfitted (test R^2 = 0.873), and MLR produced physically impossible negative estimates. Permutation importance showed that maximum daily rainfall (0.76) and monthly rainfall (0.34) carried almost all predictive power, whereas station, month, rainy days and antecedent rainfall contributed little. The degree-2 polynomial provides near-RF accuracy in a closed-form equation suitable for regional RUSLE applications.

Keywords:
Rainfall erosivity, Random Forest, polynomial regression, support vector regression, artificial neural network, permutation importance, RUSLE

INTRODUCTION

The rainfall erosivity factor (R) of the Universal Soil Loss Equation and its revised version (RUSLE) quantifies the erosive power of rainfall and is computed from storm kinetic energy and maximum 30-min intensity [1];[2]. Because the required pluviograph records are scarce, many regions rely on empirical models that estimate erosivity from daily, monthly or annual rainfall [3];[4];[5]. These models are usually calibrated in temperate or Mediterranean climates, and their transfer to humid tropical regions is uncertain [6];[7].

In Nigeria, erosivity has mostly been estimated with the Modified Fournier Index or daily-rainfall relationships [8];[9];[10]. These approaches assume fixed, usually linear or power-law, relationships between rainfall amount and erosivity, although erosivity depends strongly and non-linearly on intensity. Machine-learning (ML) methods can capture such non-linear relationships without specifying their form and have been widely applied in hydrological prediction [11];[12]. Their relative merits for erosivity estimation in West Africa, and their trade-off with interpretable regression equations, have not been assessed.

This study therefore (i) developed and compared MLR, polynomial regression, RF, SVR and ANN models for estimating monthly erosivity at five stations in South-Eastern Nigeria; (ii) identified the rainfall variables that control monthly erosivity; and (iii) assessed whether an interpretable closed-form model can match the accuracy of black-box ML models.

MATERIALS AND METHODS

Study Area and Data

The study covered five stations in South-Eastern Nigeria (Abakaliki, Awka, Enugu, Owerri and Umuahia; Table 1), a humid tropical region with a monsoon rainfall regime controlled by the Inter-Tropical Discontinuity. Hourly rainfall for the five stations was retrieved from the NASA Prediction of Worldwide Energy Resources (POWER) archive for 1 January 2005 to 13 August 2025 (Africa/Lagos time). The dataset comprised 903,600 station-hours with fields for timestamp (YYYYMMDDHH), rainfall (mm h^{-1}) and station. Timestamps were parsed and sorted by station and time, 360 negative values were set to zero as physically implausible, and no duplicate station-hours were found. Hourly records were aggregated to daily totals and daily maximum hourly intensity, then to monthly

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

totals (with rainy-day counts and maximum daily rainfall) and annual totals. The spatio-temporal characteristics of rainfall and erosivity at these stations are reported in a companion paper [13].

Table 1: Meteorological stations, land cover and dominant soils

Station	Lat. (°N)	Long. (°E)	Dominant land cover	Dominant soil
Owerri	5.4763	7.0259	Urban expansion, lowland rainforest, agriculture	Sandy loam to sandy clay loam; ferrallitic
Umuahia	5.5320	7.4860	Mixed agriculture, oil palm, fallow, forest remnants	Coastal plain sands; ferrallitic, high clay
Awka	6.2126	7.0712	Urbanising savanna, cropland, built-up	Sandy loam; derived savanna soils
Abakaliki	6.3249	8.1137	Intensive agriculture, grassland, plantations	Vertisols and hydromorphic soils
Enugu	6.5244	7.5086	Built-up, derived savanna, cropland	Acidic soils on Enugu Shale and Ajali formations

Target Variable

Daily erosivity was computed as

$$EIh_d = P_d \times I_{max,1h} \quad (1)$$

where P_d is daily rainfall (mm) and $I_{max,1h}$ is the maximum hourly intensity of that day (mm h^{-1}). Because 30-min intensities cannot be resolved from hourly data, the maximum hourly intensity was used as a conservative and reproducible substitute for I_{30}, and no kinetic-energy term was applied. Monthly and annual indices were obtained by summing EIh_d over the relevant period. All rainfall events above 0 mm were retained so that monthly calibration remained possible. The target was the monthly sum of daily EIh at each station. Monthly EIh is strongly right-skewed: a small number of extreme storm months exceed 1,000,000 mm² h^{-1}, while dry-season months are near zero.

Predictors and Data Split

Six predictors were used for each station-month (Table 2). Station and month were one-hot encoded with Abakaliki and January as baselines, giving 19 input features. The data were split chronologically to respect time order: the first 80% of station-months (n = 988; February 2005–July 2021, pooled across stations) formed the training set and the last 20% (n = 247; July 2021–August 2025) the test set.

Table 2: Predictor variables

Variable	Symbol	Definition	Unit
Monthly rainfall	R	Sum of daily rainfall in the month	mm
Maximum daily rainfall	M	Largest daily total in the month	mm
Rainy days	D	Number of days with rainfall > 0 mm	days
Antecedent rainfall	L	Previous month's rainfall at the same station	mm
Station	–	One-hot, 4 dummies (Abakaliki baseline)	–
Month	–	One-hot, 11 dummies (January baseline)	–

Models

Multiple linear regression. MLR assumes an additive linear relationship:

$$EIh = \beta_0 + \beta_1 R + \beta_2 M + \beta_3 D + \beta_4 L + \Sigma(\text{station, month effects}) \quad (2)$$

Polynomial regression. The four continuous predictors were expanded into linear, squared and pairwise-interaction terms, retaining the categorical controls (29 features at degree 2), and fitted by ordinary least squares:

$$EIh = \beta_0 + \Sigma\beta_i X_i + \Sigma\beta_{ii} X_i^2 + \Sigma_{i<j}\beta_{ij} X_i X_j + \Sigma(\text{station, month effects}) \quad (3)$$

Degrees 1, 2 and 3 were compared on the same split to examine the bias–variance trade-off.

Random Forest. RF averages the predictions of N regression trees, each grown on a bootstrap sample with random feature subsets [14]:

$$\hat{y} = (1/N)\, \Sigma_{i=1\ldots N}\, f_i(x) \quad (4)$$

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Support vector regression. SVR with a radial basis function (RBF) kernel fits a function that keeps training points within a margin ε while penalising deviations through the cost parameter C. Predictors and target were standardised using training-set statistics. C, γ and ε were tuned by grid search with 5-fold time-series cross-validation on the training set.

Artificial neural network. A multi-layer perceptron was trained on standardised inputs with early stopping (patience = 20 epochs). Architectures (32), (64, 32), (64, 64) and (128, 64, 32) and L2 penalties of 10^{-4}, 10^{-3} and 10^{-2} were tested. Final configurations are summarised in Table 3.

Table 3: Final model configurations

Model	Configuration	Training R^2
MLR	19 features, OLS	0.832
Polynomial (degree 2)	29 features, OLS	0.945
Random Forest	[number of trees, max depth]	[ ]
SVR (RBF)	C = 100, γ = 0.01, ε = 0.01; 663 of 988 support vectors (67%)	0.973
ANN (MLP)	64 → 32 → 1, α = 0.001, 178 iterations (early-stopped)	0.984

Evaluation

Models were evaluated on the test set using the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE):

$$MAE = (1/n)\,\Sigma|O_i - P_i|, \quad RMSE = \sqrt{[(1/n)\,\Sigma(O_i - P_i)^2]} \tag{5}$$

where O_i and P_i are observed and predicted monthly EIh. Variable importance in the RF model was measured by permutation importance (the drop in test R^2 when one predictor is randomly shuffled), and the response to each predictor was examined with partial dependence plots.

RESULTS

Model Comparison

All non-linear models outperformed MLR on the test set (Table 4; Figure 1). RF achieved the highest accuracy (R^2 = 0.915; MAE = 31,832; RMSE = 73,116), followed closely by degree-2 polynomial regression (R^2 = 0.910; RMSE = 75,432). ANN (R^2 = 0.874) and SVR (R^2 = 0.845) ranked third and fourth, and MLR was weakest (R^2 = 0.792; RMSE = 114,646). Relative to MLR, RF reduced RMSE by 36% and MAE by 62%. The polynomial model closed most of the gap between MLR and RF, coming within 0.005 of RF's R^2 and 3% of its RMSE.

Table 4: Test-set performance of the five models (n = 247 station-months)

Model	R^2	MAE	RMSE	Rank
Random Forest	0.915	31,832	73,116	1
Polynomial regression (degree 2)	0.910	36,446	75,432	2
ANN (MLP)	0.874	42,212	89,232	3
SVR (RBF)	0.845	38,463	98,792	4
MLR	0.792	82,996	114,646	5

MAE and RMSE in mm² h^{-1}.

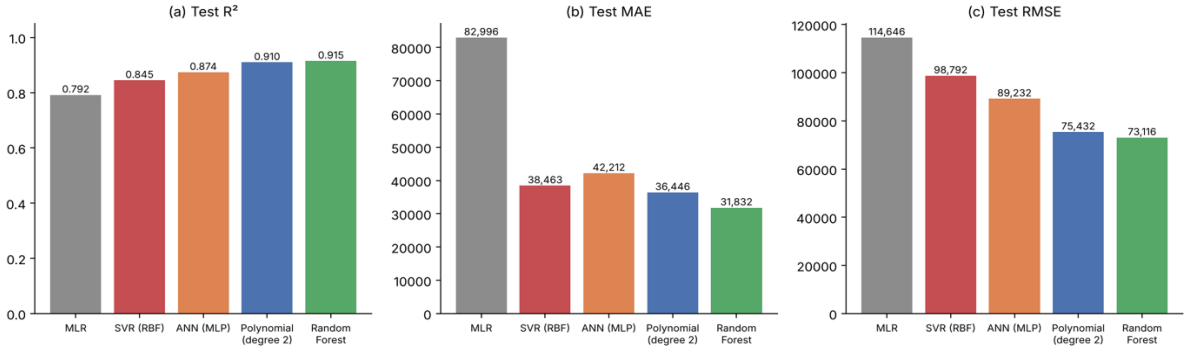


Figure 1: Test-set (a) R^2, (b) MAE and (c) RMSE of the five models

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

SVR and ANN showed the largest gaps between training and test R^2 (0.973 → 0.845 and 0.984 → 0.874), indicating some memorisation of training-period patterns despite tuning and early stopping. The high share of support vectors (67%) points to a complex, non-smooth response surface, consistent with the storm-driven, heavy-tailed nature of monthly erosivity.

Effect of Polynomial Degree

Increasing the polynomial degree from 1 to 2 improved both training R^2 (0.832 → 0.945) and test R^2 (0.792 → 0.910), showing that quadratic and interaction terms capture genuine non-linear structure (Table 5; Figure 2). Degree 3 raised training R^2 only slightly (0.956) but reduced test R^2 to 0.873 and increased RMSE to 89,385, a clear sign of overfitting. Degree 2 was therefore adopted. Its predictions lay close to the 1:1 line over most of the range, with the largest deviations in extreme months above 1,000,000 $mm^2\ h^{-1}$ (Figure 3).

Table 5: Effect of polynomial degree on training and test performance

Degree	Features	R^2 (train)	R^2 (test)	MAE (test)	RMSE (test)
1 (linear)	19	0.832	0.792	82,996	114,646
2 (quadratic)	29	0.945	0.910	36,446	75,432
3 (cubic)	49	0.956	0.873	44,312	89,385

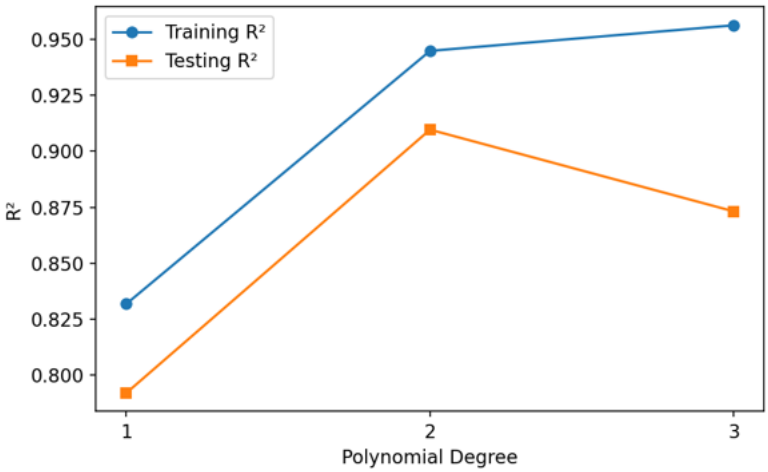


Figure 2: Training and test R^2 against polynomial degree

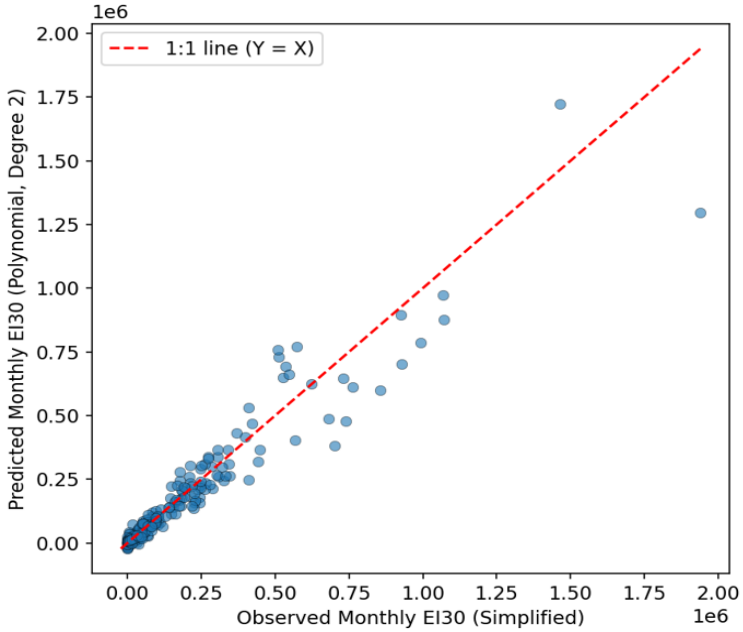


Figure 3: Predicted versus observed monthly EIh for the degree-2 polynomial model (test set)

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

The fitted coefficients of the continuous and interaction terms are listed in Table 6. Because the expanded terms are strongly collinear, individual coefficients should not be interpreted in isolation; the equation is intended for prediction within the calibrated range.

Table 6: Fitted coefficients of the degree-2 polynomial model (continuous terms)

Term	Coefficient	Term	Coefficient
Intercept	15,710.64	L^2	0.00107
R	36.42	R × M	0.00570
M	149.46	R × D	−1.5505
D	−1,863.46	R × L	−0.00414
L	−11.15	M × D	−4.0029
R^2	0.00626	M × L	0.00763
M^2	0.1115	D × L	0.4666
D^2	48.27		

R = monthly rainfall; M = maximum daily rainfall; D = rainy days; L = previous-month rainfall. Station and month coefficients omitted.

Model Behaviour for a Single Station-Month

For Abakaliki in August 2021 (monthly rainfall 3,299.92 mm, maximum daily rainfall 362.00 mm, 31 rainy days, previous-month rainfall 1,814.55 mm; observed EIh = 46,633), MLR returned a negative estimate of −18,148, which is physically impossible (Table 7). The polynomial model under-predicted by about one-third but remained positive, RF and ANN came within 15% of the observed value, and SVR over-predicted by 25%. Across this example, the RF per-tree predictions had a standard deviation of 8,039 around the ensemble mean.

Table 7: Predictions for Abakaliki, August 2021 (observed EIh = 46,633)

Model	Predicted EIh	Error (%)
MLR	−18,148	−139
Polynomial (degree 2)	31,355	−33
Random Forest	53,152	+14
SVR (RBF)	58,227	+25
ANN (MLP)	49,786	+7

Controls on Monthly Erosivity

Permutation importance showed that maximum daily rainfall was by far the most influential predictor (0.763 ± 0.061), followed by monthly rainfall (0.341 ± 0.022; Table 8; Figure 4). Month (0.0025), antecedent rainfall (0.0019) and rainy days (0.0015) contributed almost nothing once rainfall magnitude was known, and station had a slightly negative importance (−0.0007), meaning that its spatial information was already captured by the rainfall variables. Partial dependence plots showed EIh rising steeply and non-linearly with maximum daily rainfall and monthly rainfall, and responding only weakly to rainy days and antecedent rainfall (Figure 5).

Table 8: Permutation importance of predictors in the Random Forest model

Predictor	Mean importance	Std
Maximum daily rainfall	0.7626	0.0612
Monthly rainfall	0.3414	0.0219
Month	0.0025	0.0010
Antecedent rainfall	0.0019	0.0026
Rainy days	0.0015	0.0010
Station	−0.0007	0.0032

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

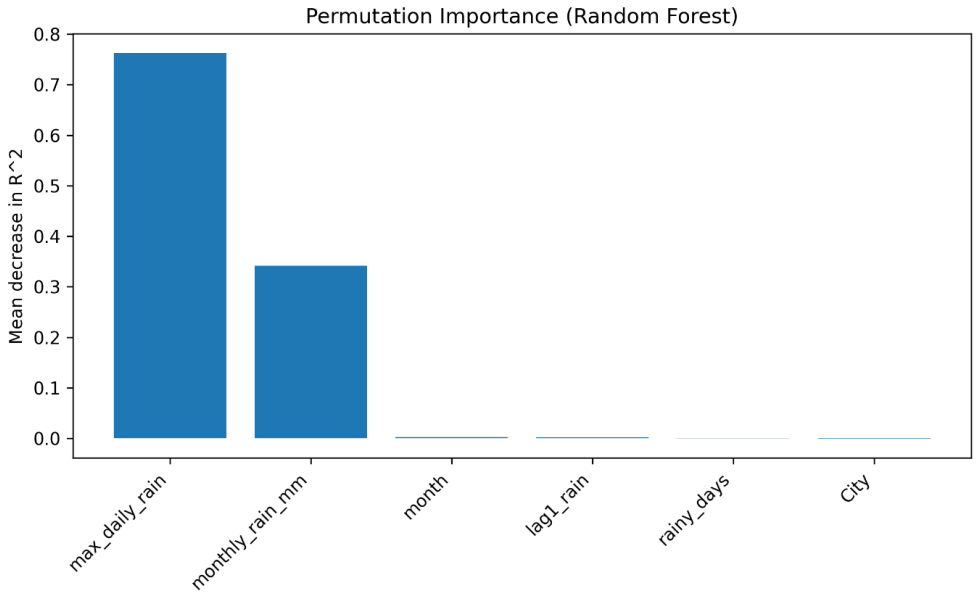


Figure 4: Permutation importance of predictors in the Random Forest model

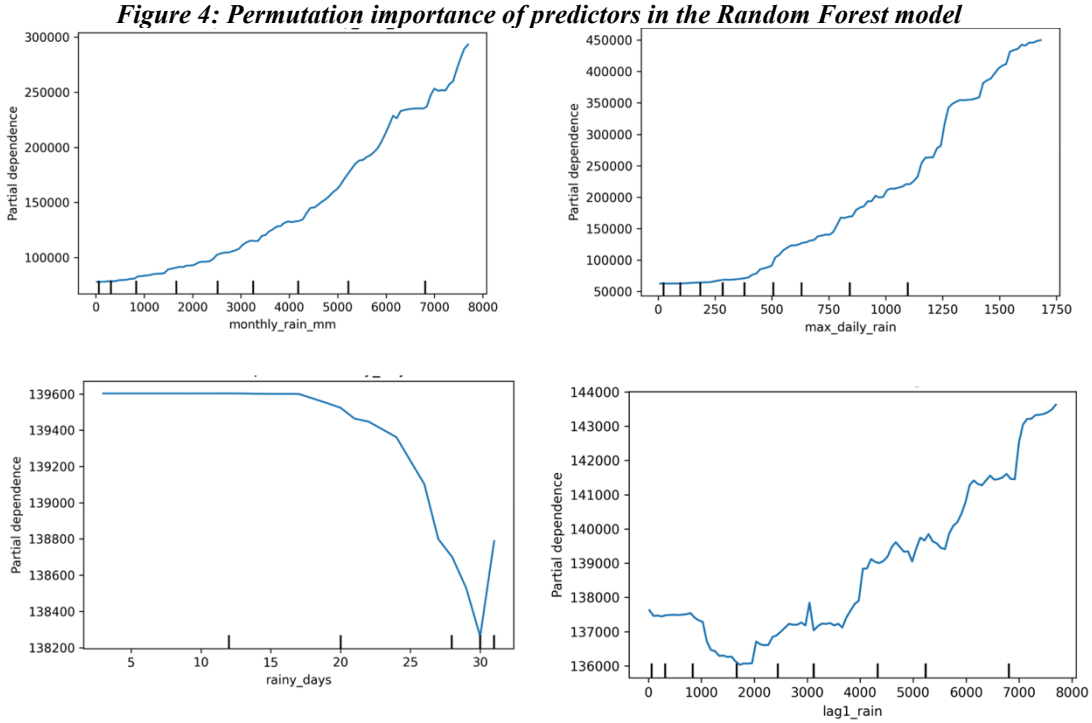


Figure 5: Partial dependence of monthly EIh on (a) monthly rainfall, (b) maximum daily rainfall, (c) rainy days and (d) antecedent rainfall

DISCUSSION

Why Tree-Based and Quadratic Models Perform Best

Erosivity is inherently non-linear in rainfall: erosive energy rises faster than rainfall amount, and a few intense days can dominate a month's erosivity. MLR cannot represent this, which explains both its lower accuracy and its negative estimates. Quadratic and interaction terms capture much of the curvature, especially the interaction of rainy days with the two rainfall-magnitude predictors, and RF captures the same structure through recursive partitioning. SVR and ANN did not improve on these simpler models. This is common for small-to-moderate tabular datasets: with fewer than 1,000 training months, kernel and neural methods have too little data to realise their flexibility and tend to fit training-specific noise.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

Physical Meaning of the Dominant Predictors

The dominance of maximum daily rainfall agrees with the physical basis of erosivity: intense storms carry more kinetic energy and generate more runoff, so the largest event in a month contributes disproportionately to its total. Monthly rainfall adds information on the overall water input. The negligible contribution of rainy days confirms that the frequency of rain matters far less than its intensity—many low-intensity rain days produce less erosion than a few intense storms. The small role of station and month implies that rainfall statistics alone capture most spatial and seasonal variation in erosivity across the region, which supports a single regional model rather than separate station equations.

Practical Use and Limitations

Where accuracy is the only goal, RF is marginally superior. Where transparency, integration into RUSLE-type empirical frameworks or easy use in spreadsheets matters, the degree-2 polynomial is preferable, since it delivers nearly identical accuracy as an explicit equation. Several limitations apply. First, the target is an hourly-intensity index rather than a kinetic-energy-based EI_{30}. Second, because the target is computed from daily rainfall, predictors such as maximum daily rainfall share information with it, which partly explains the high R^2 values; the models are best understood as converting monthly-scale rainfall statistics into an event-based erosivity measure. Third, all models struggled with the few extreme storm months, which appear to be an irreducible source of error at monthly resolution. Fourth, the models are calibrated on satellite-derived rainfall and require validation with gauge data before operational use.

CONCLUSION

Five models were compared for estimating monthly rainfall erosivity at five stations in South-Eastern Nigeria using monthly rainfall statistics. Random Forest ($R^2 = 0.915$) and degree-2 polynomial regression ($R^2 = 0.910$) performed best, followed by ANN (0.874), SVR (0.845) and MLR (0.792). Quadratic terms captured most of the non-linearity that RF exploits, while a cubic model overfitted. Maximum daily rainfall and monthly rainfall dominated prediction; station, month, rainy days and antecedent rainfall added little. The degree-2 polynomial is recommended where an interpretable equation is needed and RF where maximum accuracy is required. Future work should validate the models against gauge-based EI_{30}, test them with monthly rainfall alone and add explicit treatment of extreme storm months.

DECLARATIONS

Funding: The authors received no specific funding for this work.

Conflict of interest: The authors declare no competing interests.

Data availability: The hourly rainfall data are freely available from the NASA POWER archive (https://power.larc.nasa.gov). Derived daily, monthly and annual erosivity datasets are available from the corresponding author on reasonable request.

REFERENCES

[1] Wischmeier, W.H., and Smith, D.D. (1978). Predicting Rainfall Erosion Losses: A Guide to Conservation Planning. Agriculture Handbook No. 537. USDA, Washington, DC.

[2] Renard, K.G., Foster, G.R., Weesies, G.A., McCool, D.K., and Yoder, D.C. (1997). Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE). Agriculture Handbook No. 703. USDA, Washington, DC.

[3] Lee, J.-H., and Heo, J.-H. (2011). Evaluation of estimation methods for rainfall erosivity based on annual precipitation in Korea. Journal of Hydrology, 409, 30–48.

[4] Mello, C.R., Viola, M.R., Beskow, S., and Norton, L.D. (2013). Multivariate models for annual rainfall erosivity in Brazil. Geoderma, 202, 88–102. https://doi.org/10.1016/j.geoderma.2013.03.009

[5] Xie, Y., Yin, S., Liu, B., Nearing, M.A., and Zhao, Y. (2016). Models for estimating daily rainfall erosivity in China. Journal of Hydrology, 535, 547–558. https://doi.org/10.1016/j.jhydrol.2016.02.020

[6] Oliveira, P.T.S., Wendland, E., and Nearing, M.A. (2012). Rainfall erosivity in Brazil: A review. Catena, 100, 139–147. https://doi.org/10.1016/j.catena.2012.08.006

[7] Nearing, M.A., Yin, S., Borrelli, P., and Polyakov, V.O. (2017). Rainfall erosivity: An historical review. Catena, 157, 357–362. https://doi.org/10.1016/j.catena.2017.06.004

[8] Salako, F.K. (2010). Development of isoerodent maps for Nigeria from daily rainfall amount. Geoderma, 156, 372–378. https://doi.org/10.1016/j.geoderma.2010.03.006

[9] Okorafor, O.O., Akinbile, C.O., Adeyemo, A., and Egwuonwu, C.C. (2017). Determination of rainfall erosivity index (R) for Imo State, Nigeria. American Journal of Engineering Research, 6(2), 1–13.

[10] Ekpo, A.E., Orakwe, L.C., and Nwachukwu, C.P. (2021). Determination of rainfall erosivity in Akwa Ibom State using Modified Fournier Index. Journal of Engineering and Applied Sciences, 19(1), 427–435.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

[11] Solomatine, D.P., and Siek, M.B. (2004). Application of machine learning techniques for rainfall-runoff modelling. [Complete the source details from the original.]
[12] Valipour, M., Shahsavani, D., and Choubin, B. (2021). Data-driven modelling approaches for hydrological forecasting and prediction. [Complete the source details from the original.]
[13] Anyanwu, S.C., et al. (submitted). Multi-scale spatio-temporal variability and trends of rainfall erosivity in South-Eastern Nigeria from hourly satellite-derived rainfall.
[14] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

 © Author(s) | CC BY 4.0 International License.