Hybrid Machine Learning and Geostatistical Downscaling of ERA5-Land Data for High-Resolution Daily Precipitation Mapping over the Philippines

Kent Roger Truita
DOST - Advanced Science and Technology Institute
Quezon City, Philippines
kentroger.truita@asti.dost.gov.ph

Elmer Peramo
DOST - Advanced Science and Technology Institute
Quezon City, Philippines
elmer@asti.dost.gov.ph

Jeanette Badong-Carlos
DOST - Advanced Science and Technology Institute
Quezon City, Philippines
jeanette@asti.dost.gov.ph

Karl Ezra Pilario
University of the Philippines Diliman
Quezon City, Philippines
kspilario@up.edu.ph

***Abstract*—High-resolution precipitation maps are critical for hydrological modeling and hazard assessment in the Philippines, yet reanalysis datasets such as ERA5-Land are too coarse for many local applications. We develop a hybrid machine learning–geostatistical framework to downscale the daily ERA5-Land precipitation to a 1 km grid over the entire country. The models were trained and evaluated against observations from 374 rain gauges and automated weather stations for 2010–2020; among these, the histogram-based gradient boosting decision trees (HGBDT) performed the best, achieving an RMSE of 9.088 mm/day, an MAE of 4.204 mm/day, and an R^2 of 0.139, compared to the raw ERA5-Land (RMSE 13.284 mm/day, MAE 6.929 mm/day, R^2 -0.827). Applying the geostatistical correction to the model's residuals showed additional improvements, most notably during the dry season. This study shows that the downscaled precipitation map was able to capture the fine-scale orographic patterns that coarse reanalysis datasets usually miss. Although it still tends to underestimate heavier rainfall events (above 20 mm/day), it offers a more physically realistic and statistically improved high-resolution dataset that can support local hydrological work and hazard management.**

Keywords: Precipitation Downscaling, Machine Learning, ERA5-Land, Histogram-based Gradient Boosting, Philippines

I. Introduction

Precipitation is a crucial component of the hydrological cycle, playing a critical role in regulating water resources, agricultural activities, and flood dynamics [1]. The Philippines is one of the most disaster-prone countries in the world. An archipelagic nation located in the Western North Pacific, it exemplifies such complexity with its unique orographic patterns, as well as an interplay of tropical cyclones, monsoons, and mesoscale convective systems that lead to extreme spatiotemporal precipitation variability [2], [3]. Capturing this extreme spatiotemporal variability is essential for hydrological modeling, flood early warning, and climate-change impact assessment. However, quantifying an accurate spatiotemporal precipitation distribution remains a challenge.

Although the national rain-gauge network continues to expand over time, it remains relatively sparse in many mountainous and island regions, resulting in large uncertainties when gridded data are generally derived from station interpolation [4]. Furthermore, global reanalysis data such as ERA5-Land [5] allow for a physically consistent analysis of all atmospheric states on an hourly basis, but their coarse resolution means they can't capture small-scale terrain effects and coastal storm patterns that shape local rainfall. As a result, ERA5-Land tends to underestimate intense rainfall and fails to capture the detailed spatial variability needed for reliable local assessments, especially during events like tropical cyclone landfalls and monsoon surges [6].

Empirical statistical downscaling offers a practical way to connect local hydrometeorological conditions with large-scale atmospheric information. Dynamical downscaling, by contrast, uses regional climate models to resolve high-resolution atmospheric processes, but it requires substantial computing resources and inherits biases from the driving global models. Empirical downscaling instead uses statistical relationships between large-scale climate patterns and local observations to represent fine-scale features in complex environments [7], [8]. Conventional parametric regression methods have been the standard for this task, although they often fail to capture the complex, non-linear relationships inherent in the precipitation processes. [9]

Recent developments in hydroinformatics have shown that machine learning (ML) algorithms are more effective in downscaling climate variables than traditional statistical techniques. Research has demonstrated that non-linear algorithms like Random Forest (RF) and Gradient Boosting greatly outper-

form interpolation and linear regression methods in terms of reducing error metrics and enhancing rain event detection [10]. These AI-driven techniques are especially useful in areas with limited data since they can produce high-resolution projections that are essential for policy development and disaster mitigation [11].

This study presents a machine learning and geostatistical method to downscale ERA5-Land precipitation data from 9 km spatial resolution to a 1 km daily grid for the area of the Philippines. By leveraging the dense network of local automated weather stations and rain-gauges (PhilSensors) as the ground truth, this research applies physical atmospheric predictors, high-resolution topographical variables and geostatistical determinants within an optimized ML model. This method seeks to provide a reliable high-resolution precipitation dataset appropriate for hydrometeorological applications in the Philippines while correcting the biases present in reanalysis data.

II. Methodology

The overall workflow of this study is shown in Fig. 1. The methodology follows a step-by-step process, beginning with data collection and preprocessing and ending with the generation of high-resolution precipitation estimates.

Multiple datasets are combined to construct the downscaling framework, including ground-based observations, reanalysis products, and auxiliary geographic and climate information. The prepared dataset is then used to train and evaluate several machine learning models. Feature selection and hyperparameter tuning are carried out using cross-validation to ensure robust model performance.

Model results are compared using an independent test set, and the best-performing approach is applied to produce high-resolution precipitation maps. Further details of each step are provided in the corresponding subsections.

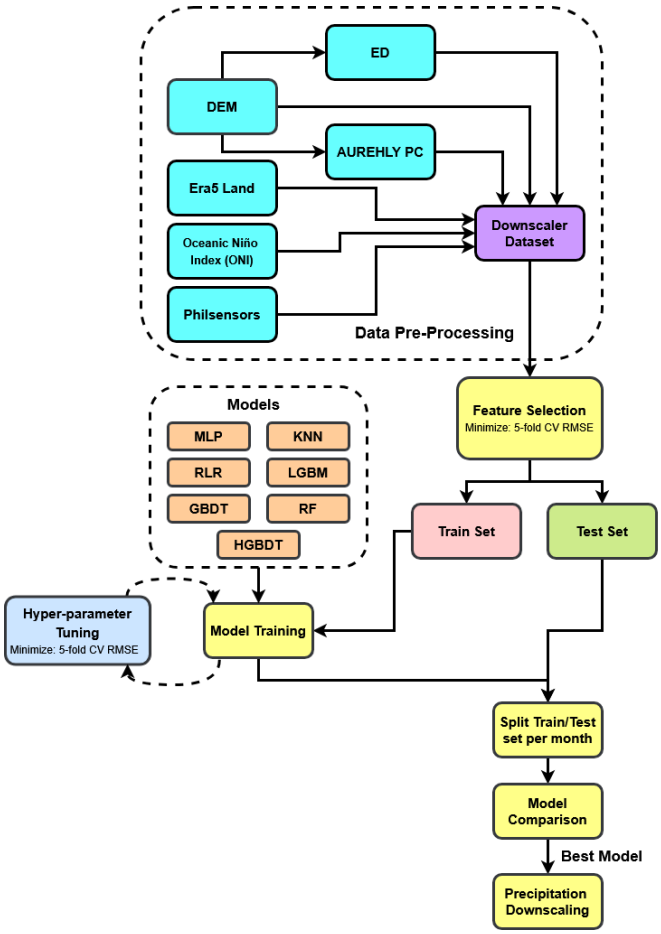


Fig. 1: Overall methodological framework for precipitation downscaling in the Philippines. The workflow integrates station data (PhilSensors), reanalysis (ERA5-Land), and topographic features (DEM) into a machine learning pipeline involving feature selection, hyperparameter tuning, and comparative model evaluation.

A. Study Area

The area of focus for this study is the Philippines, a Western North Pacific archipelagic nation. Home to more than 7600 islands, the region's geomorphology varies from mountain ranges like the Sierra Madre and Cordillera Central, which contribute to regional rainfall due to orographic lifting and rain shadow effects as shown in Fig. 2. The climate is influenced by the Asian Monsoon system (Southwest Monsoon and Northeast Monsoons) and the frequent occurrence of tropical cyclones. The region is a perfect testbed for high-resolution downscaling because of the extreme spatiotemporal rainfall variability produced by these factors interacting with the rough terrain [1].

B. Data Acquisition and Preprocessing

1) Ground-Based and Reanalysis Data: Daily precipitation data is obtained from the 374 Automatic Weather Stations (AWS) and rain gauges of the PhilSensors network from January 1, 2010, to December 31, 2020. The PhilSensors network spans to diverse topographic areas, and the altitudes of the stations range from almost 0 m to over 2,500 m.

The PhilSensors data were recorded at 15-minute intervals, and the precipitation measurements were aggregated to daily totals that served as the ground truth. Several quality checks were made to ensure data quality and prevent outliers, individual data points from the station data that shows unrealistic values such as negative precipitation or totals exceeding 1,000 mm/day, were identified as outliers and removed. The station network was then randomly divided into train and test for unbiased evaluation. 317 stations (84.8%) comprised the training set used for model training and hyperparameter optimization. 57 stations (15.2%) comprised the independent test set that was not used during model training.

The atmospheric predictors was derive from the ERA5-Land reanalysis data [5]. Although these are physically uniform and produced on an hourly frequency, the spatial resolution of the ERA5-Land data (∼9 km) is still too coarse for downscaled, location-specific studies. For ease of use, the following eight

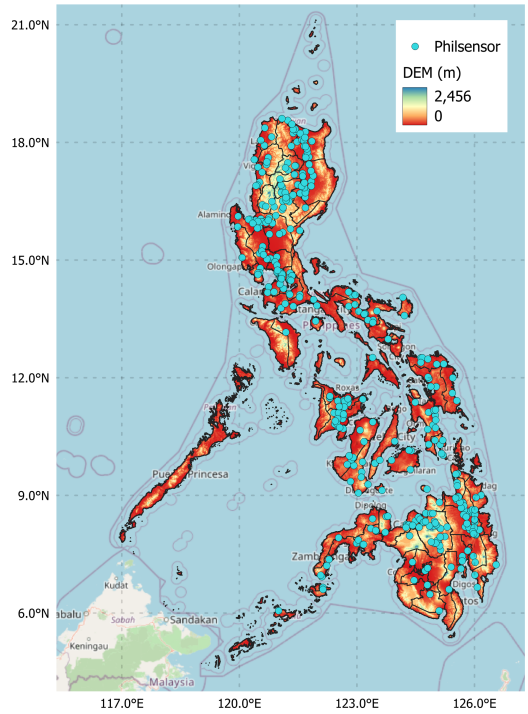


Fig. 2: Digital elevation model of the Philippines with Philsensors station locations overlaid.

dynamic variables of interest were chosen: 2m temperature, 2m dewpoint, surface pressure, total precipitation, u/v wind components, 2m solar radiation and evaporation.

2) Topographic and Geostatistical Features: To resolve sub-grid heterogeneity, we integrated high-resolution static features:

- **Topography:** Derived from the Copernicus GLO-30 DEM, resampled to 1 km. Features include elevation, slope, and aspect.
- **Proximity:** Euclidean distances to rivers, coastlines, and lakes, plus land-sea percentage, to capture coastal convergence effects.
- **Climate Index:** The Oceanic Niño Index (ONI) was included to account for ENSO-driven climate variability.

3) AUREHLY Principal Components: To model spatial autocorrelation, we used the AUREHLY (Analysis of Using the Relief for Hydrometeorology) method. Similar to a previous study [12], AUREHLY was computed by getting the elevation difference between the point and the grids. We selected 32 nearest neighbors along latitude and 46 along longitude for each station, resulting in 1,550 points (1550 = (2 × 16 + 1) × (2 × 23 + 1) - 1). Principal Component Analysis (PCA) was applied then to the matrix of elevation difference retaining the top 15 components (PC1–PC15). These components act as additional information about the terrains surrounding each point.

C. Feature Selection Strategy

A Sequential Feature Selection method was used to identify the most optimal subset of features. A model was first trained on the atmospheric features from ERA5-Land, which served as a baseline. Topographic features and PCs were candidate variables that were added iteratively. A variable was retained only if it showed a reduction in the RMSE on the 5-fold cross-validation. This implies that variables leading to a lower RMSE are those that contribute to improving the model's predictive performance [12].

D. Machine Learning Models

Different types of regression-based machine learning models were used to represent the complex and non-linear relationships between atmospheric conditions and local precipitation. Within the downscaling framework, these models serve as the deterministic component and are used to generate initial precipitation estimates. To account for spatially structured errors, Ordinary Kriging was then applied to the model residuals, resulting in a combined approach commonly known as Regression Kriging (RK). The set of evaluated ML models includes both interpretable baseline models and more advanced approaches such as ensemble methods and neural networks, which are described in the following sections.

1) Ridge Linear Regression (RLR): we used Ridge Linear Regression (RLR) as one of the model we explored. The standard Ordinary Least Squares (OLS) regression sometimes tends to yield unstable estimates when some predictors are highly correlated and they are also prone to overfitting based on small fluctuations in the training data. RLR can account for this limitation by adding an L2 regularization penalty term to the objective function. This penalty shrinks the regression coefficients towards zero, thereby reducing model variance and improving generalization [10], [13]. The prediction for a new input vector x is given by:

$$y_{\text{pred}} = \mathbf{x}^\top \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{1}$$

where $\mathbf{X}$ is the matrix of predictors, $\mathbf{y}$ is the vector of observed rainfall, and λ is the complexity parameter that controls the strength of the regularization.

2) K-Nearest Neighbors (KNN): KNN is a non-parametric instance-based machine learning algorithm that does not assume a fixed functional form for the rainfall-topography relationship. Instead, it operates on the principle of feature similarity. For a target pixel x, the prediction is derived from the values of its neighbors in the training set; this method was shown to be effective for downscaling in recent hydrological studies [10]:

$$y_{\text{pred}} = \frac{1}{k} \sum_{i \in \mathcal{N}_x} y_i \tag{2}$$

where $\mathcal{N}_x$ is the set of the k nearest neighbors to x in the multi-dimensional feature space. This approach inherently preserves the way precipitation values vary together across space, keeping the original spatial structure of the data intact [14].

3) Multilayer Perceptron (MLP): To capture the non-linear relationships of the data, we utilized a Multilayer Perceptron (MLP). The feedforward neural network usually consists of an input layer, one or more hidden layers with non-linear activation functions, and a linear output layer for the rainfall prediction. For a single hidden layer, the model output can be written as [7]:

$$y_{\text{pred}} = \mathbf{W}_2\, g\big(\mathbf{W}_1^\top \mathbf{x} + \mathbf{b}_1\big) + b_2, \tag{3}$$

where $\mathbf{x}$ is the input feature vector; $\mathbf{W}_1$ and $\mathbf{b}_1$ denote the weight matrix and bias vector of the hidden layer; $g(\cdot)$ is the non-linear activation function (e.g., hyperbolic tangent or ReLU); and $\mathbf{W}_2$ and b_2 are the output-layer weight vector and bias term, respectively. The network parameters are optimized using gradient-based learning to minimize the loss.

4) Random Forest (RF): To address the complex non-linearity of tropical precipitation, we also used Random Forest. RF is an ensemble method that constructs a multitude of decor-related decision trees using bootstrap aggregating (bagging) and random feature selection at each split [15]. Each tree T_b is trained on a random subset of the data, and the final prediction is the average of the outputs from the ensemble of B trees:

$$y_{\text{pred}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{4}$$

where $\mathbf{x}$ denotes the input feature vector (e.g., elevation, coordinates). This ensembling strategy reduces model variance, making the RF less prone to overfitting than individual decision trees [16].

5) Gradient Boosting Decision Trees (GBDT): Unlike Random Forest, which builds trees independently, Gradient Boosting constructs trees sequentially. Each new tree $h_m(x)$ is trained to approximate the pseudo-residuals of the current ensemble. The final prediction is expressed as the sum of an initial model F_0 and the weighted contributions of M sequentially trained trees [17]:

$$y_{\text{pred}} = F_0(x) + \sum_{m=1}^{M} \nu\, h_m(x) \tag{5}$$

where ν is the learning rate. By iteratively targeting the remaining errors, GBDT can reduce model bias relative to bagged tree ensembles, and previous work has shown improved performance for typhoon-related precipitation extremes [18].

6) Light Gradient Boosting Machine (LGBM): LGBM is an optimized version of the gradient boosting method mentioned above. LGBM uses the same additive prediction model as standard GBDT, but instead of growing trees level-wise, it grows trees leaf-wise (best-first). It grows the leaf that minimizes loss instead of growing individual leaves layer by layer, LGBM can achieve faster loss minimization for a given model size. Combined with histogram-based split finding and other system-level optimizations, this makes LGBM highly efficient for large datasets and well suited to precipitation reconstruction tasks [19].

7) Histogram-Based Gradient Boosting (HGBDT): To handle the large spatial grids more efficiently, we used the Histogram-Based Gradient Boosting method (HGBDT). It follows the same general idea as standard gradient boosting, but it speeds things up by grouping continuous variables, such as elevation or humidity, into histogram bins. By working with these bins instead of every individual value, the algorithm reduces the cost of searching for the best splits. The complexity drops from being tied to the full dataset to being tied only to the number of bins and features. In practice, this keeps most of the accuracy of traditional GBDT models while cutting down training time and memory requirements [19], [20].

E. Hyperparameter Optimization

To better compare the best configuration of each model, hyperparameter tuning was done using the Optuna optimization framework. For each machine learning model, appropriate search ranges were specified for the key hyperparameters, as summarized in Table I. The optimization objective was to minimize the root mean squared error (RMSE), evaluated over 50 trials using five-fold cross-validation on the training dataset. This cross-validation helps ensure that the selected hyperparameter configurations generalize well to previously unseen data. The best optimized hyperparameter values used for each model are also shown in Table I.

TABLE I: Search ranges and optimal hyperparameter values for each model as found by Optuna.

Model	Hyper-parameters	Search Range	Best Value
RLR	Alpha (Regularization)	[0.01, 100] (Log)	7.64
	Variogram Model	{spherical, exponential, gaussian, linear}	Gaussian
KNN	No. of neighbors	[3, 20]	20
	Weighting	{uniform, distance}	Uniform
	Variogram Model	{spherical, exponential, gaussian, linear}	Gaussian
ANN	No. of layers	[1, 2]	1
	Layer size	{50, 100, 150}	100
	Activation	{relu, tanh}	ReLU
	Alpha (Regularization)	[0.01, 100] (Log)	0.0006
	Initial learning rate	[0.0001, 0.01] (Log)	0.0006
	Variogram Model	{spherical, exponential, gaussian, linear}	Gaussian
RF	No. of estimators	[50, 300]	282
	Max depth	[5, 30]	26
	Min samples split	[2, 20]	10
	Min samples leaf	[1, 10]	10
	Variogram Model	{spherical, exponential, gaussian, linear}	Linear
GBDT	No. of estimators	[50, 300]	235
	Max depth	[3, 15]	8
	Learning rate	[0.01, 0.3] (Log)	0.018
	Subsample	[0.6, 1.0]	0.888
	Variogram Model	{spherical, exponential, gaussian, linear}	Spherical
LGBM	No. of estimators	[50, 300]	300
	Max depth	[3, 15]	8
	Learning rate	[0.01, 0.3] (Log)	0.017
	Num leaves	[20, 150]	125
	Subsample	[0.6, 1.0]	0.774
	Variogram Model	{spherical, exponential, gaussian, linear}	Spherical
HGBDT	Max iterations	[50, 300]	228
	Max depth	[3, 15]	11
	Learning rate	[0.01, 0.3] (Log)	0.028
	Variogram Model	{spherical, exponential, gaussian, linear}	Linear

F. Model Evaluation

To better reflect the strong seasonal pattern of rainfall in the Philippines, such as the shifts between the dry months

and the monsoon period, the training and testing data were further divided into twelve subsets that match the calendar months. Separate models were then trained and evaluated for each month so that the learning algorithms could adapt to the atmospheric conditions that tend to dominate during that specific part of the year.

Model performance was assessed on the held-out test stations using continuous metrics such as the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The Kling–Gupta Efficiency (KGE) was also computed to evaluate hydrological consistency:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \qquad (6)$$

where r, β, and γ correspond to correlation, bias ratio, and variability ratio respectively.

To understand how the models behaved across different rainfall intensities, precipitation was grouped into No Rain, Light, Moderate, and Heavy categories. The Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI) were then calculated for each category. These metrics provided a clearer view of the models' ability to capture events across the seasonal cycle [10].

G. High-Resolution Mapping

The model that performed the best was then used to create daily rainfall maps at a 1 km scale. To prepare the inputs, the ERA5 predictor variables were upscaled through nearest-neighbor interpolation. The AUREHLY principal components at each pixel were standardized using the mean and standard deviation derived from the original fifteen components. The ML model was used to estimated precipitation for every pixel, drawing on the broad physical patterns in the reanalysis data while preserving the finer spatial structure provided by terrain and station measurements.

III. Results and Discussion

A. Comparative Assessment of Machine Learning Models

The performance of each models was evaluated on four metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and Kling-Gupta Efficiency (KGE) based on the test set (57 stations). Table II summarizes the findings on average over the course of the 12-month study period. Among the evaluated ML models, the Histogram-based Gradient Boosting Decision Tree (HGBDT) showed the most promising performance, achieving the lowest RMSE and MAE of 9.088 and 4.204 respectively. It achieved the highest R^2 among the models, with a value of 0.139. While an R^2 of 0.139 may appear low in absolute terms, it shows a significant improvement over the raw ERA5-Land baseline, which shows a negative R^2 of -0.827 and a much higher RMSE and MAE of 13.284 and 6.929. This confirms that the downscaling framework successfully captures the local-scale information that coarse reanalysis data cannot.

The top-performing models were tree-based ensemble models. HGBDT produced the lowest test RMSE of 9.088 with RF and LGBM only slightly higher at 9.105 and 9.127 respectively. Interestingly, RLR achieved a competitive performance compared from the tree-based ensemble models with a test RMSE of 9.27, not that far off from the more complex GBDT as well. The instance based KNN model, however, had a much higher Test RMSE of 9.313 than the regression and tree-based methods. In contrast, the ANN failed to capture the precipitation dynamics effectively, producing a test RMSE value of 14.291, which is the highest error rate among the machine learning models, indicating it performed worse than a simple climatological average.

TABLE II: Performance metrics of different precipitation downscaling models averaged across 12 months. Best values for the testing set are highlighted in bold.

Model	RMSE		MAE		R^2		KGE	
	Train	Test	Train	Test	Train	Test	Train	Test
HGBDT	9.443	**9.088**	4.161	**4.204**	0.274	**0.139**	0.235	0.081
RF	8.371	9.105	3.544	4.244	0.427	0.135	0.349	0.086
LGBM	8.847	9.127	3.907	4.208	0.366	0.131	0.318	**0.104**
GBDT	8.280	9.217	3.805	4.258	0.448	0.114	0.383	0.093
RLR	10.163	9.266	4.486	4.370	0.159	0.105	0.100	0.003
KNN	9.500	9.313	4.157	4.245	0.266	0.094	0.236	0.088
ANN	12.734	14.291	7.040	8.033	-0.382	-1.248	-0.504	-1.062
ERA5	14.091	13.284	7.210	6.929	-0.608	-0.827	-0.445	-0.697

B. Impact of Geostatistical Residual Correction

We examined whether geostatistical post-processing makes the ML predictions more accurate by exploring the effect of Ordinary Kriging on model residuals. A clear seasonal pattern emerges, as shown in Fig. 3. During the dry season and the transition months (January to May and November to December), Kriging consistently reduces the RMSE. The largest improvements occur in December with a reduction of about 0.035 and in January with a reduction of about 0.032. These gains appear because rainfall during these periods is relatively uniform, which allows the spatial structure of the errors to be interpolated effectively.

The opposite happens during the Southwest Monsoon from June to October. In this period, the correction increases the error, at times by up to 0.025. Rainfall becomes highly convective and varies sharply across short distances, which violates the assumptions needed for Kriging to work well. As a result, the method oversmooths features that the HGBDT model already captured. Overall, residual correction is helpful only during stable rainfall conditions, while the uncorrected model performs better during the convective wet season.

C. Categorical Skill and Detection Capability

To assess the operational value of the HGBDT model, we examined its performance across four rainfall intensity categories. Fig. 4 shows the monthly patterns of POD, FAR, and CSI.

The model performs best in detecting days without rainfall. During the dry season (Months 1-5 and 12), the POD hovers around 0.71 and the FAR remains below 0.10 consistently

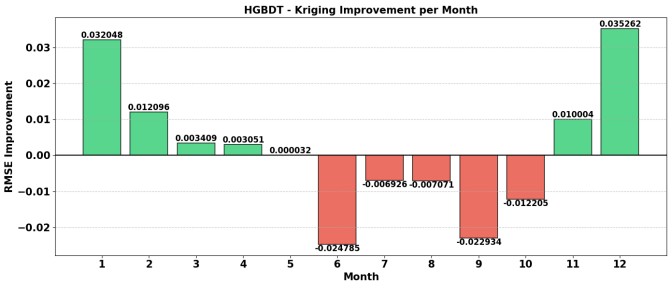


Fig. 3: Monthly RMSE improvement attributed to the Kriging residual correction. Positive values (Green) indicate improvement; negative values (Red) indicate deterioration.

resulting in CSI values upwards of 0.70; these are stable values for dry seasons.

For Light and Moderate rain, the model maintains a moderate POD (around 0.50–0.60), but this is paired with very high false alarm rates, especially for Light Rain (FAR $>$ 0.80). This shows a general wet bias common in gridded products across the Philippines [2]. This means that such algorithms overestimate frequency in low precipitation events to avoid global residuals during interpolation or regression [4].

The model struggles most with heavy rainfall. The POD for Heavy Rain (>20 mm) drops below 0.15 during the wet season (Months 6–9), and CSI values across all rain categories remain low (<0.20). Although the model captures spatial rainfall patterns reasonably well, it has difficulty estimating the intensity of localized rain.

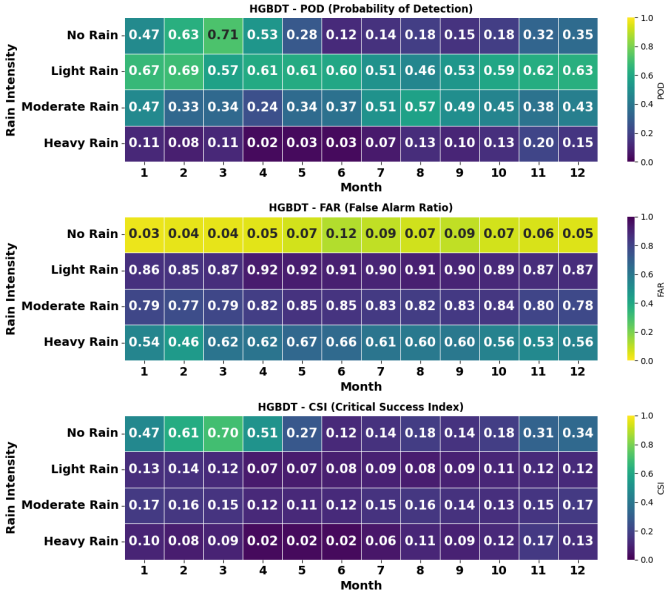
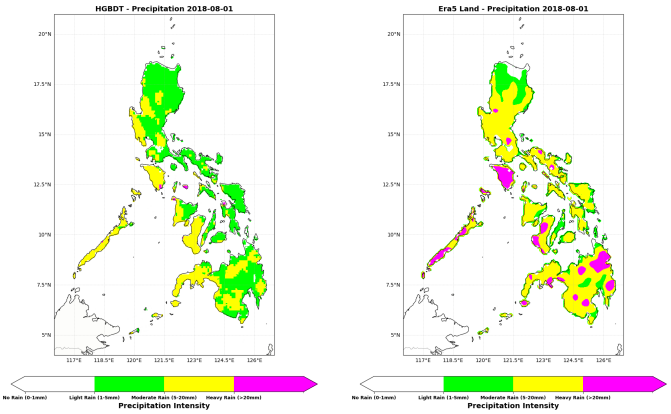


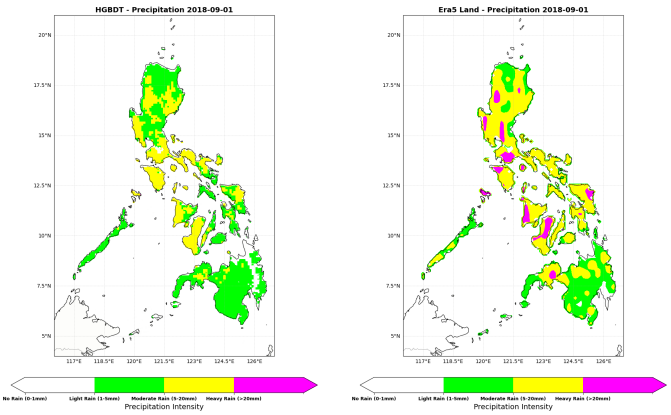
Fig. 4: Categorical performance metrics (POD, FAR, CSI) of the HGBDT model across four rainfall intensity classes (No Rain, Light, Moderate, Heavy) for each month.

D. Spatial Reconstruction of Daily Rainfall

To check how well the downscaling framework performs, we compared the 1 km precipitation maps it produced with the original ERA5-Land data, as shown in Fig. 5. The comparison shows that the model brings back much of the small-scale detail that is missing in the coarse reanalysis data. The HGBDT output better defines orographic features, such as the rain shadows in Northern Luzon on August 1, which appear much smoother and more uniform in the baseline maps. It also captures the patchy structure of tropical convection by breaking up broad areas of moderate rainfall into smaller pockets of stronger precipitation, as seen on September 1. These improvements suggest that the model makes good use of static topographic information to reduce spatial smoothing and produce a more realistic picture of the highly variable rainfall patterns found across the Philippines.



(a) Comparison for 2018-08-01: The HGBDT model (left) resolves sharp orographic boundaries and rain shadows that are blurred in the ERA5 baseline (right).



(b) Comparison for 2018-09-01: The HGBDT model (left) reconstructs localized high-intensity pockets (magenta) within broader rain bands, correcting the smoothing effect seen in ERA5 (right).

Fig. 5: Visual comparison of spatial downscaling results. The HGBDT model consistently recovers fine-scale spatial structures and intensity gradients that are lost in the coarse ERA5-Land reanalysis data.

E. Discussion

The strong results from the HGBDT model has shown how useful nonlinear machine-learning approaches can be for downscaling climate data in tropical regions with complex terrain. The model was able to capture how topography and atmospheric variables jointly shape local rainfall, which helps reduce several biases present in the ERA5-Land dataset. The spatial patterns it produces also look more physically realistic, with rainfall placed correctly on windward slopes and coastal structures appearing much sharper than in the coarser reanalysis maps.

There is still a balancing act when it comes to lowering the RMSE while also representing rainfall extremes. The model often predicts too many light rain events, which increases the false alarm rate, and it tends to miss heavier rainfall, which lowers the probability of detection. This behavior reflects the difficulty of inferring short-lived convective storms from large-scale atmospheric variables. The seasonal limits of the Kriging correction also suggest that adding physical covariates to the interpolation, such as through Regression Kriging, might give more stable results during the convective season.

IV. Conclusion

We produced a daily 1 km × 1 km precipitation dataset for the Philippines by downscaling ERA5-Land with several machine-learning models. This study showed that the HGBDT model was the most effective at reducing error and better capturing subdaily characteristics of precipitation compared to the reanalysis. Residual correction via Ordinary Kriging was successful during dry seasons but not during the predominantly convective typhoon season. The model also effectively captured dry days, although it did overpredict the number of light rains and failed to capture the majority of heavy rain events. Future studies should assess other deep learning models and models that can account for data imbalance, especially for extremes. Ultimately, the final product provides a clearer and more locally observed representation of precipitation, which serves as a reliable dataset for hydrological and agricultural efforts across the Philippines.

Acknowledgment

This work is funded and monitored by the Department of Science and Technology (DOST) – Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) with Project No. 1212543

References

[1] J. G. Gacu, S. A. Kantoush, and B. Q. Nguyen, "Full-cycle evaluation of multi-source precipitation products for hydrological applications in the Magat River Basin, Philippines," *Remote Sens.*, vol. 17, no. 3375, 2025.

[2] E. A. S. Tana, L. M. P. Olaguera, *et al.*, "Assessing the performance of GSMaP and IMERG in representing the diurnal cycle of precipitation in the Philippines during the southwest monsoon season," *Atmos. Res.*, vol. 317, p. 107983, 2025.

[3] C. G. Mesias and G. Bagtasa, "AI-based tropical cyclone rainfall forecasting in the Philippines using machine learning," *Meteorol. Appl.*, vol. 32, no. e70083, 2025.

[4] J. C. A. C. Peralta, G. T. T. Narisma, and F. A. T. Cruz, "Validation of high-resolution gridded rainfall datasets for climate applications in the Philippines," *J. Hydrometeorol.*, vol. 21, pp. 1571–1587, 2020.

[5] J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, *et al.*, "ERA5-Land: A state-of-the-art global reanalysis dataset for land applications," *Earth Syst. Sci. Data*, vol. 13, pp. 4349–4383, 2021.

[6] G. Bagtasa and B. A. Racoma, "Does the Sierra Madre Mountain Range in Luzon act as a barrier to typhoons?," *Philipp. J. Sci.*, vol. 152, no. S1, pp. 63–77, 2023.

[7] D. Boateng and S. G. Mutz, "pyESDv1.0.1: An open-source Python framework for empirical-statistical downscaling of climate information," *Geosci. Model Dev.*, vol. 16, pp. 6801–6822, 2023.

[8] H. Tang, J. Chen, *et al.*, "High-resolution spatiotemporal mapping from ERA5 reanalysis: integrating multi-source data with machine learning for the city of Shenzhen," *Build. Environ.*, vol. 287, 2025.

[9] C. Chen, Q. Chen, *et al.*, "Comparison of different methods for spatial downscaling of GPM IMERG V06B satellite precipitation product over a typical arid to semi-arid area," *Front. Earth Sci.*, vol. 8, 2020.

[10] G. V. Nguyen, X.-H. Le, *et al.*, "Machine learning approaches for reconstructing gridded precipitation based on multiple source products," *J. Hydrol. Reg. Stud.*, vol. 48, 2023.

[11] M. Waqas and U. W. Humphries, "Artificial intelligence-driven precipitation downscaling and projections over Thailand using CMIP6 climate models," *Big Earth Data*, vol. 9, no. 3, 2025.

[12] N. Giorgos, P. T. Nastos, and Y. Kapsomenakis, "Creating high-resolution precipitation and extreme precipitation indices datasets by downscaling and improving on the ERA5 reanalysis data over Greece," *Eng*, vol. 5, pp. 1885–1904, 2024.

[13] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[14] D. Yates, S. Gangopadhyay, B. Rajagopalan, and K. Strzepek, "A technique for generating regional climate scenarios using a nearest neighbor algorithm," *Water Resour. Res.*, vol. 39, no. 7, p. 1199, 2003.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[16] O. Baez-Villanueva, M. Zambrano-Bigiarini, H. Beck, *et al.*, "RF-MEP: A novel random forest method for merging gridded precipitation products and ground-based measurements," *Remote Sens. Environ.*, vol. 239, p. 111606, 2020.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] Z. Shen and B. Yong, "Downscaling the GPM-based satellite precipitation retrievals using gradient boosting decision tree approach over Mainland China," *J. Hydrol.*, vol. 602, p. 126803, 2021.

[19] G. Ke, Q. Meng, T. Finley, *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, 2017.

[20] G. Ntagkounakis, P. Nastos, and Y. Kapsomenakis, "Statistical downscaling of ERA5 reanalysis precipitation over the complex terrain of Greece," *Environ. Sci. Proc.*, vol. 26, no. 1, p. 81, 2023.