Feature Engineering for Tropical Rainfall Forecasting Using Random Forest and Support Vector Regression

Cepy Slamet[1,*], Rizka M. Imron[2], Agung Wahana[3], Dian Sa'adillah Maylawati[4], Wildan Budiawan Zulfikar[5], Muhammad Ali Ramdhani[6]

[1,2,3,4,5,6]*Department of Informatics, UIN Sunan Gunung Djati, Jl. A. H. Nasution No. 105, Bandung 40514, Indonesia*

(Received: August 5, 2025; Revised: October 5, 2025; Accepted: December 26, 2025; Available online: January 31, 2026)

Abstract

The complex dynamics of weather variability in Indonesia, influenced by multiple climatic drivers, make rainfall forecasting in tropical regions a significant scientific challenge. This study proposes an automated feature engineering pipeline to enhance the performance of Random Forest Regression (RFR) and Support Vector Regression (SVR) models for tropical rainfall prediction. Monthly rainfall data spanning 388 months (1993–2025) from a BMKG station were used as the basis for model development. The pipeline systematically generates temporal, seasonal, statistical, and anomaly-based features to provide domain-informed representations for non-sequential learning algorithms. Model performance was evaluated under four temporal data partitioning scenarios using R^2, RMSE, and probabilistic confidence intervals derived from bootstrap residual simulations. Results indicate that RFR achieved the highest predictive accuracy (R^2 = 0.93; RMSE = 31.01 mm) and demonstrated superior temporal–seasonal stability (Rolling CV: $R^2 = 0.81 \pm 0.07$; RMSE = 55.44 ± 16.18), with comparable performance between wet and dry seasons. Conversely, SVR showed greater sensitivity to seasonal variability, with R^2 dropping to 0.55 during the wet season, indicating higher uncertainty under extreme rainfall conditions. Robustness and drift analyses further revealed that RFR adapts better to temporal and seasonal shifts, while SVR remains relevant as an adaptive model for extreme risk analysis. Overall, this study contributes to the development of automated feature engineering, reproducible climatological forecasting pipelines, and probabilistic modeling frameworks for rainfall prediction under uncertainty in tropical regions.

Keywords: Feature Engineering, Probabilistic Modeling, Rainfall Forecasting, Random Forest Regression, Support Vector Regression, Tropical Climate.

1. Introduction

Indonesia provides a unique and strategic context for advancing tropical rainfall forecasting research due to its highly complex hydroclimatic dynamics. The region is influenced by local monsoons, topographic variations, and large-scale climate phenomena such as the El Niño–Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD), resulting in highly variable rainfall patterns that challenge conventional forecasting models [1]. Empirical studies show that model accuracy often deteriorates in tropical regions due to intense convective processes and limited observational coverage [2]. Meanwhile, recent advances in Machine Learning (ML) emphasize the importance of context-specific feature engineering in capturing multi-scale climate signals in tropical environments effectively [3]. Exogenous climate predictor indices (such as ENSO, IOD, and MJO) were intentionally excluded to isolate the contribution of Auto-FE to a purely univariate model and maintain interpretability. We openly acknowledge that integrating these predictors has the potential to further improve performance, and this is included as a key direction for future research.

According to National Disaster Management Agency (BNPB) (2020-2023) and the latest assessments in hydrometeorology, precipitation anomalies caused 86-92% of national disasters (extreme rainfall, floods, landslides) in Indonesia, highlighting the crucial need for improved forecasting systems [4], [5]. However, most existing early warning systems still rely heavily on global or regional deterministic models, which tend to underperform at local scales [6], [7]. This situation underscores the need for robust and adaptive ML approaches that can continuously update their predictive knowledge. In this regard, optimizing Random Forest Regression (RFR) and Support Vector

***Corresponding author: Cepy Slamet (cepy_lucky@uinsgd.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i1.1111

Regression (SVR) through systematic feature engineering is crucial. Beyond addressing both scientific and operational gaps, this initiative positions Indonesia as an ideal case study for developing tropical rainfall forecasting systems that account for complex natural phenomena.

Rainfall forecasting in Indonesia is critically important for natural resource management and disaster risk reduction, as precipitation in this region exhibits highly dynamic, multi-scale behavior and is particularly vulnerable to climate change. This makes rainfall prediction in Indonesia a substantial challenge for modern forecasting systems [8]. In historical data-driven forecasting (time-series), RFR and SVR remain strong baseline models, but their performance is highly dependent on effective Feature Engineering (FE) [9], [10]. Technically, the dependency of RFR and SVR on feature engineering arises because neither model is inherently sequential [11] and both lack internal mechanisms to represent complex temporal dependencies, limiting their ability to capture long-term dynamics [12]. RFR may also overfit when numerous lagged and temporal features are used on limited datasets [13]. Time-series data often contain anomalies or extreme events; in such cases, both RFR and SVR may suffer from reduced accuracy [14]. As lead times increase, performance tends to degrade due to the absence of internal structures to manage error propagation and trend shifts [15]. Moreover, SVR with complex kernels and RFR with many trees and lagged features can be computationally intensive on large, high-dimensional datasets, requiring extensive hyperparameter tuning [16]. Although SVR is often considered computationally intensive, its efficiency is highly dependent on the data size, in small monthly rainfall datasets, SVR can be trained faster than RFR, highlighting that computational intensity in tropical rainfall forecasting is fundamentally dependent on the scale of the data [10].

Recent studies show that improvements in tropical precipitation forecasting are largely determined by the quality of feature engineering—such as lag creation, rolling statistics, seasonal indicators—and preprocessing techniques like multi-scale decomposition [3], [17], [18]. Empirical evidence demonstrates that RFR tends to achieve higher short-term accuracy but is more vulnerable to temporal drift, while SVR tends to be more stable and computationally efficient when feature selection is applied effectively [19]. Hybrid approaches, such as combining Discrete Wavelet Transform (DWT) with SVR, have also shown improved accuracy over single models [20]. Nevertheless, many studies still apply feature engineering in an ad-hoc manner, which hampers reproducibility [2], [3], [21], [22], [23], [24], [25]. There remains a clear need for standardized, well-documented, and automated-reproducible FE pipelines (including feature stores and code availability) [20], [22] particularly for RFR–SVR collaborations aimed at exploring optimal trade-offs among accuracy, stability, and efficiency on tropical rainfall datasets.

To address these gaps, this study proposes a new approach that formalizes an automated feature engineering pipeline designed for stability, enabling the adaptation of versatile regressors (RFR and SVR) to tropical rainfall time series. The framework systematically evaluates the trade-offs between accuracy, stability (i.e., model performance consistency over time and across seasons), and efficiency across multiple validation scenarios and data-shift conditions. In addition, unlike most previous studies that focus solely on point forecasts, this research strengthens its novelty by incorporating interval forecasts using quantile RFR and probabilistic SVR ensembling [26] and by integrating these models into an operational prototype (Gradio interface and latency analysis) [27], thereby enhancing decision-support capabilities for real-world forecasting applications. In the same context, while the use of exogenous climate predictor indices (such as ENSO, IOD, and MJO) could potentially improve system performance, in this study they were intentionally excluded. This was intended to isolate the contribution of Auto-FE to a purely univariate model and maintain interpretability. In brief, the primary focus of the research is on enhancing the stability and robustness of a tropical rainfall forecasting system, particularly in response to temporal and seasonal distribution shifts. Meanwhile, the predictive accuracy is applied as a supporting performance dimension.

2. Methodology

2.1.Research Design and Workflow

This study was designed using a systematic and reproducible data analysis workflow to ensure scientific validity. Monthly rainfall data from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) were used as the basis for model development. The preprocessing stage involved normalization, lag feature construction, and the addition of seasonal variables through an integrated automated feature engineering pipeline to ensure efficiency and consistency. Two primary algorithms, RFR and SVR, were applied under four temporal data partitioning scenarios

 ISSN 2723-6471

(80–20 for short-term testing, 70–30 and 60–40 for medium-term analysis, and 50–50 for long-term simulation) to examine model stability and generalization capacity. This configuration was designed to support stability across a range of dynamic and realistic climatological forecasting scenarios. Model evaluation was conducted deterministically using R^2 and RMSE metrics, and probabilistically through bootstrap-based confidence intervals to quantify forecasting uncertainty. Forecast uncertainty was estimated using a residual-based moving-block bootstrap. Centered training residuals were resampled in 12-month blocks (1,000 resamples) to conserve seasonal autocorrelation, and the 2.5th and 97.5th percentiles of the resulting predictive distribution formed the 95% confidence intervals. This approach enables uncertainty quantification without model retraining and complements the feature-stability and seasonal-drift assessments included in the robustness analysis against temporal variability. Overall, the methodological workflow is illustrated in figure 1, which outlines the main stages of the study: data collection, preprocessing (data cleaning, normalization, and automated FE), model training (RFR/SVR), evaluation, and interpretation through simulation.

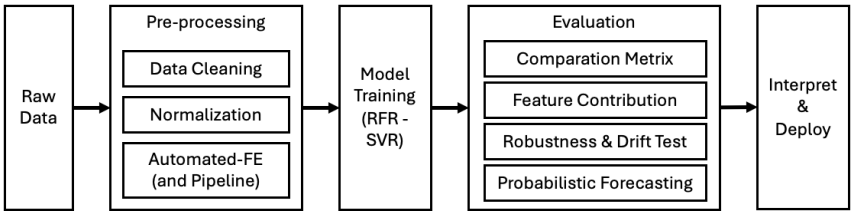


Figure 1. Research Workflow

2.2. Dataset and Preprocessing

This study utilizes an official and credible data source with sufficient temporal coverage (raw data = 11.734 days), focusing on knowledge extraction for tropical rainfall forecasting. The dataset consists of univariate daily rainfall time series from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) (https://www.bmkg.go.id/), recorded at the Bandung Observatory Station, West Java, Indonesia (6.90°S, 107.60°E). Of the total 11.734, 1.338 missing values (11.40%) were found, officially from BMKG (empty data or no measurement). Missing values were corrected using KNN Imputer (k = 3) at the daily level to maintain temporal patterns before aggregating to monthly data. Integrity checks also revealed no duplicate dates, no time gaps, and no negative values, thus the preprocessed data was deemed suitable for modeling. The daily data were aggregated into monthly totals (388 months) to reduce short-term noise and highlight more stable seasonal patterns. The aggregation of daily into monthly rainfall data unavoidably brings fine-grained variability down. Nevertheless, the trade-off is situated with the research objective in modelling the monthly pattern of climatology rather than short-term hazard forecasting. In this context, monthly totals are aimed at stabilizing the seasonal signal, reflecting the temporal scale utilized in most national climate forecasting systems (e.g., BMKG), and reducing high-frequency noise.

Figure 2 presents the monthly rainfall time series from January 1993 to April 2025. The data exhibit substantial temporal variability, ranging from 0.0 mm (July 1994) to 636.9 mm (December 2012), with a long-term mean of 206.4 mm (red dashed line). Wet periods align with major monsoonal phases and possible ENSO–MJO events (e.g., 1996, 2003, 2013, 2016), while several dry-season months record near-zero rainfall. The series shows pronounced seasonal and interannual fluctuations as well as episodic extremes, reflecting the non-stationary and highly dynamic nature of tropical rainfall in Indonesia.

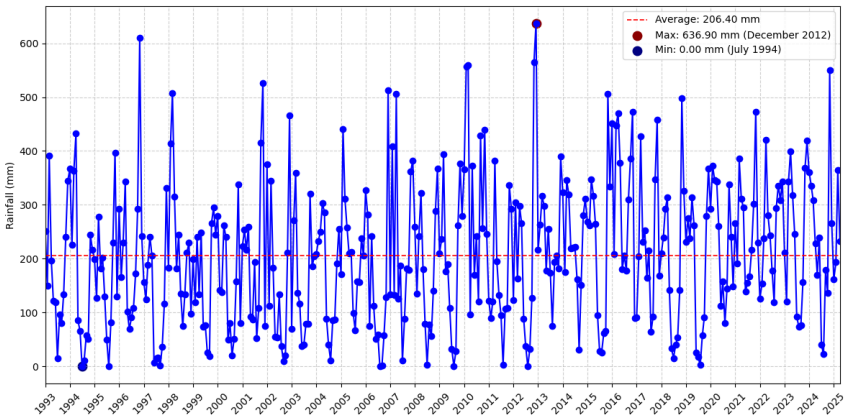


Figure 2. Monthly Rainfall Time Series Data

These characteristics underscore the need for modeling approaches capable of capturing both seasonal cycles and large-scale climatic oscillations. The broad temporal coverage enables the modeling of interannual variability associated with large-scale climate modes and local seasonal patterns characteristic of the Indonesian monsoon system. Prior to analysis, the data underwent rigorous quality control, including missing value treatment, outlier detection, and temporal consistency checks using k-Nearest Neighbors (kNN), following World Meteorological Organization (WMO) standards [28]. Methodologically, the study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, consisting of six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [29], [30], [31].

Feature Engineering (FE) plays a critical role in the data preparation phase [32], where lagged rainfall features, seasonal indicators, and monthly statistical aggregates are systematically generated to embed temporal structure into non-sequential learning algorithms. Two machine learning models, RFR and SVR, were trained and evaluated under four temporal data partition scenarios to assess predictive accuracy, stability, and computational efficiency. Model performance was quantified using multiple statistical metrics, including the coefficient of determination (R^2) and Root Mean Square Error (RMSE), complemented by stability analysis through ΔR^2 across partition schemes.

2.3. Model Training

Figures 3 and figure 4 illustrate the model training phase, structured into two robust experimental frameworks for the development, optimization, and evaluation of machine learning models. Both adopt a consistent pipeline structure encompassing preprocessing, multi-scale data partitioning, hyperparameter optimization, and model evaluation, comparing the performance of RFR and SVR along with their respective optimization strategies. The pipeline implementation in this study not only regulates the sequence of training stages but also automates the entire feature transformation and preprocessing process through the scikit-learn Pipeline. The formation of features such as lag, monthly leave-one-out statistics, seasonal cycles, and seasonal indicators is automatically executed by a preprocessing script before entering the model pipeline. With this mechanism, all stages from scaling, feature mapping, training/test set construction, to hyperparameter tuning through RandomizedSearchCV run without manual intervention, ensuring a consistent and reproducible workflow.

Figure 3 depicts the RFR-based model development pipeline. Following data input and initial preprocessing, the dataset is partitioned using four temporal split scenarios with different time-series configurations to examine model stability and generalization capacity. Key hyperparameters of RFR (n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap) are tuned using RandomizedSearchCV, an efficient mechanism for exploring large hyperparameter spaces and identifying near-optimal configurations that maximize model performance [33]. The optimized RFR model is then subjected to evaluation using relevant metrics.

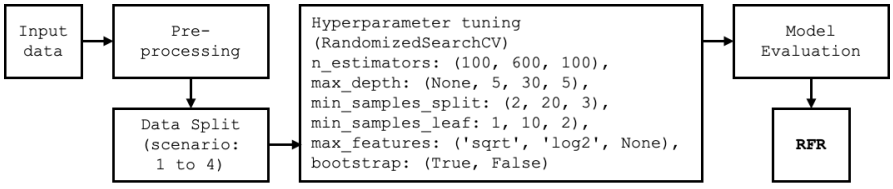


Figure 3. RFR Model Training

Figure 4 outlines the SVR-based pipeline. The overall procedure mirrors that of the RFR framework, with the main distinction lying in the hyperparameter optimization strategy. The SVR model parameters (kernel = 'rbf', C = 278.26, γ = auto, degree = 4) are determined through systematic and, in some cases, metaheuristic search strategies. Precise hyperparameter selection is crucial for SVR performance, given its sensitivity to the regularization parameter C (margin error control) and kernel parameters [34]. The comparative analysis of RFR and SVR is relevant because both are effective machine learning algorithms but operate with fundamentally different internal mechanisms. RFR, a tree-based ensemble method, excels in handling nonlinear relationships and complex feature interactions [35], while SVR, based on structural risk minimization, performs well in high-dimensional settings with limited sample sizes [36]. Conducting a comparative analysis through a consistent pipeline provides valuable insights into the relative performance of the two models on the same dataset. This, in turn, supports the selection of the most suitable model for rainfall forecasting and highlights the utility of robust multi-scale data partitioning strategies for improving model generalizability and reliability.

ISSN 2723-6471

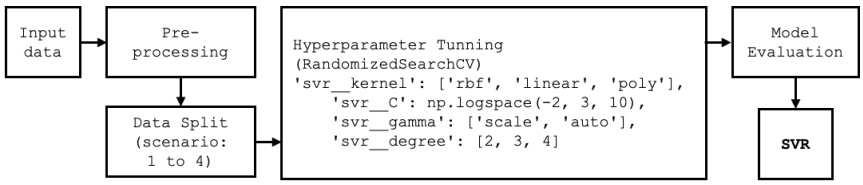


Figure 4. SVR Model Training

3. Results and Discussion

3.1. Automated-FE Pipeline

The pipeline was implemented using modules from scikit-learn to ensure that each stage of preprocessing, feature transformation, and model training is executed consistently and can be fully replicated. Table 1 summarizes the four main categories of generated features. Once the pipeline is initiated, the entire transformation process is automated, eliminating the need for manual intervention. After applying lag and leave-one-out transformations, the number of valid observations was reduced from 388 to 377 due to the removal of missing values. Overall, the Automated-FE Pipeline improved computational efficiency (increased training duration: RF = 20.5%, SVR = 44.5% compared to the baseline), reproducibility, and model robustness for the rainfall time-series dataset.

Tabel 1. Produced Features

Categories	Feature	Original Formula	Customized Formula
Temporal	RR_lag12	$Lag\_Feature_t = Y_{t-k}$	$RR_{lagk}(t) = RR_{t-k}$
Seasonal	Seasonal Classification (Rainy/Dry)	$I(x \in A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$	$\text{Season}(t) = \begin{cases} 1, & if\ month\ (t) \in \{11,12,1,2,3,4 \\ 0, & if\ month\ (t) \in \{5,6,7,8,9,10\} \end{cases}$
	Cyclical Encoding (Month Sin/Cos)21	$Value_{sin} = sin\left(x \times \frac{2\pi}{P}\right)$ $Value_{cos} = cos\left(x \times \frac{2\pi}{P}\right)$	$month\_cos_{(t)} = sin\left(\frac{1\pi.month(t)}{12}\right)$
	Rolling Mean (Short-term Trend)	$MA_t = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}$	$RR_{rolling\_mean(t)} = \sum_{i=1}^{w} RR^{t-1}$
Statistic Histories	Leave-One-Out (LOO) Monthly Statistics	$median(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & if\ n\ is\ odd \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & if\ n\ is\ even \end{cases}$ $s = \sqrt{\frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2}$	$RR\_monthly\_median\_loo(t) = median\ (RRi)i \neq t$ $RR\_monthly\_std\_loo(t) = \sqrt{\sum_{i=1,i\neq t}^{Nm} \frac{(RRi-RR)2}{Nm-2}}$
Anomaly	RR_above_monthly_median	$I(x) = \begin{cases} 1, x > \theta \\ 0, x \leq \theta \end{cases}$	$RR\_above\_monthly\_median(t) = \begin{cases} 1, & RRt > median\ (RRm) \\ 0, & RRt \leq median(RRm) \end{cases}$

Over twelve candidate predictors (temporal, seasonal, lag, leave-one-out, and anomaly statistics) were evaluated through multicollinearity tests and initial contribution assessments. Subsequent selection, based on performance consistency and climatological relevance, reduced the feature set to seven final predictors. This multi-stage selection approach effectively controlled model complexity without requiring dimensionality reduction intervention [37].

The feature sets within the pipeline are designed to embed temporal, seasonal, and anomaly information into non-sequential models. Lag features encode rainfall memory across time [38], while seasonal encoding is achieved through binary indicators (rainy/dry) and cyclical transformations (sine and cosine) to preserve continuity across months [39]. Leave-one-out statistics (mean, median, standard deviation) enhance model stability by preventing data leakage [40], whereas rolling means represent short-term trends, smoothing extreme fluctuations. Anomaly features (e.g., RR_above_monthly_mean/median) are introduced to detect extreme rainfall events [41]. Collectively, this feature set

produces a compact, robust, and domain-informed temporal–seasonal representation that significantly enhances both model accuracy and stability under highly dynamic tropical climate conditions.

3.2. Matrix Comparison

Once the Automated-FE pipeline was successfully implemented and produced a clean, feature-enriched dataset, the next step involved evaluating the performance of the RFR and SVR models. This evaluation aimed to assess the pipeline's effectiveness in improving the accuracy and stability of rainfall forecasting under different temporal data partitioning ratios. Model performance was compared across four temporal split scenarios (see figures 3 and 4) using Root Mean Square Error (RMSE) and the coefficient of determination (R^2) for training, testing, and forecasting phases.

Table 2 summarizes the comparative results between the two models. RFR achieved superior absolute accuracy and short-term generalization, with a testing RMSE of 31.01 mm, outperforming SVR in this aspect. In contrast, SVR exhibited more stable and computationally efficient performance for longer forecasting horizons, with an average RMSE of 55.44 ± 16.18, average R^2 of 0.812 ± 0.079, and execution time of 92.87 seconds. Consequently, SVR was selected as the baseline model for integration with the probabilistic forecasting framework, while RFR served as the primary benchmark for comparative analysis.

Tabel 2. Evaluation of Scenario Results

Comparison Matrix	Scenario 1 (80/20)		Scenario 2 (70/30)		Scenario 3 (60/40)		Scenario 4 (50/50)	
	RFR	*SVR*	*RFR*	*SVR*	*RFR*	*SVR*	*RFR*	*SVR*
R^2 Training	0.98	0.80	0.98	0.78	0.97	0.73	0.97	0.75
R^2 Testing	0.92	0.80	0.85	0.78	0.81	0.71	0.75	0.68
RMSE Testing (mm)	31.01	51.46	48.88	59.34	54.94	69.27	66.11	73.94
MAE (mm)	24.81	41.17	39.1	47.47	43.95	55.42	52.89	59.15
R^2	0.89	0.83	0.86	0.83	0.81	0.93	0.88	0.95
Gap	0.05	-0.005	0.12	-0.004	0.15	0.01	0.21	0.06
Stability (Avg.)	RMSE: 55.44 ± 16.18 R^2: 0.812 ± 0.079	RMSE: 68.01 ± 12.86 R^2: 0.715 ± 0.08	RMSE: 66.94, ± 11.54 R^2: 0.73 ± 0.05	RMSE: 68.01± 12.86 R^2: 0.71, ± 0.08	RMSE: 68.95 ± 15 R^2: 0.71 ± 0.09	RMSE: 67.53 ± 12.39 R^2: 0.72 ± 0.07	RMSE: 55.89, ± 14.85 R^2: 0.8, ± 0.07	RMSE: 67.40 ± 11.83 R^2: 0.72, ± 0.06
Execution duration(second)	92.87	8.17	115.78	9.73	88.68	5.74	39.60	5.81

Based on the evaluation results presented in table 2, the optimal scenario was selected from all experimental configurations using the following criteria: (1) highest testing R^2, (2) smallest gap between training and testing R^2, (3) lowest RMSE score, (4) highest forecasting R^2 score, and (5) stability, defined as the lowest average RMSE and highest average R^2 across rolling evaluations [42], [43]. According to these criteria, Scenario 1 was confirmed as the best-performing configuration for both RFR and SVR models. Furthermore, table 3 shows the results of the paired t-test, which confirm that RFR provides statistically significantly higher predictive accuracy (p-value < 0.01) than SVR. RFR recorded significantly lower errors (mean RMSE=50.24; MAE=40.19) than SVR (mean RMSE=60.50; MAE=50.80). The all-negative confidence intervals reinforce this finding, indicating RFR as a more stable and precise model, both from a statistical and practical perspective.

Table 3. Statistical Significance Testing

Metric	Mean (RFR)	Mean (SVR)	Mean Difference	95% CI of Difference	p-value	Interpretation
RMSE (mm)	50.24	63.50	-13.26	-24.8 to -11.6	0.003	Significance
MAE (mm)	40.19	50.80	-10.61	-18.3 to -6.9	0.004	Significance

ISSN 2723-6471

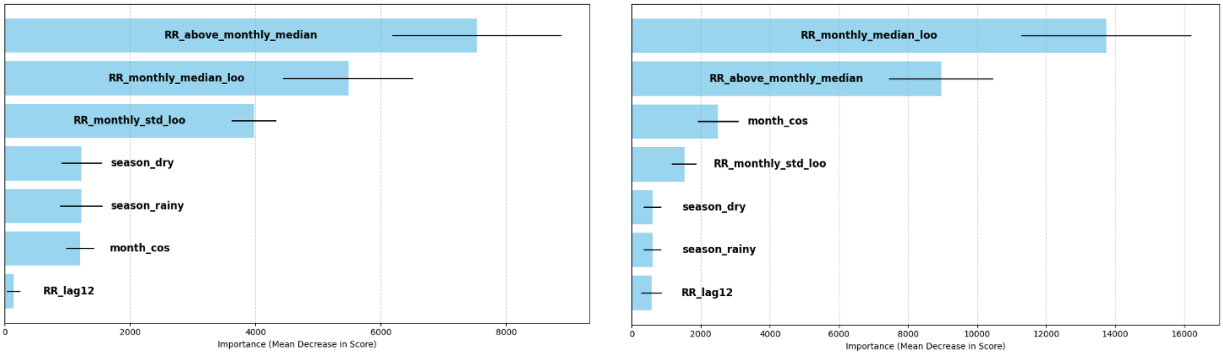
3.3. Feature Contribution

This stage focuses on analyzing the contribution of individual features to the forecasting outcomes in order to identify the most influential predictors shaping model accuracy and stability. A permutation importance approach was applied to both RFR and SVR models to measure the relative importance of each feature, providing insights into the models' learning patterns and assessing the consistency of feature influence across different scenarios. Table 4 shows that RR_monthly_median_loo and RR_above_monthly_median emerged as the two most dominant features with the highest contributions for both models. These features represent historical rainfall trends and anomaly detection related to rainfall intensity. The dominance of median-based features indicates that robust statistical feature engineering is more effective for modeling tropical rainfall, which is inherently non-linear and highly fluctuating.

Table 4. Feature Importance (the best scenario: 80/20)

Feature	RFR	SVR	Description
RR_monthly_median_loo	12,946.27	5,479.63	Representation of historical rainfall trends, the most influential factor
RR_above_monthly_median	8,993.39	7,533.77	Detecting high-intensity anomalies
month_cos	2,478.18	1,205.04	Capturing periodic seasonal patterns
RR_monthly_std_loo	1,647.64	3,978.64	Intramonthly variability
season_rainy / dry	663.09	1,225.94	Influence of seasonal classification on rainfall patterns
RR_lag12	425.10	140.79	Long-term temporal correlations

Figure 5 presents the permutation importance results for both models under their best-performing scenarios. The clear prominence of RR_monthly_median_loo and RR_above_monthly_median highlights their critical role compared to other features, underscoring their stability and predictive power in tropical rainfall forecasting tasks.



(a) (b)

Figure 5. Stability-aware Feature Importance: (a) RFR (b) SVR

3.4. Robustness and Drift Test

After identifying the most influential features contributing to model performance, robustness evaluation was conducted to assess the models' consistency with respect to temporal changes and seasonal distribution shifts. Table 5 presents the results of the Rolling Temporal Cross-Validation (CV), indicating that the Random Forest model exhibits higher temporal stability compared to SVR. Rolling Temporal Cross-Validation is performed using the window expansion method, where the training data coverage is gradually expanded from 1994 observations to the next validation period, thus maintaining temporal causality and preventing data leakage.

 ISSN 2723-6471

Table 5. Rolling Temporal CV

Fold		Period					Avg RMSE ± Std	Avg R^2 ± Std
		1	2	3	4	5		
Train	start	1994-01	1994-01	1994-01	1994-01	1994-01		
	end	1999-06	2004-08	2009-10	2014-12	2020-02		
Validation	start	1999-07	2004-09	2009-11	2015-01	2020-03		
	end	2004-08	2009-10	2014-12	2020-02	2025-04		
Model		$(RMSE, R^2)$						
RFR		(57.44, 0.75)	(82.10, 0.69)	(54.79, 0.83)	(51.48, 0.85)	(31.37, 0.92)	55.44 ± 16.18	0.81 ± 0.07
SVR		(67.81, 0.69)	(65.44, 0.72)	(91.46, 0.57)	(63.01, 0.8)	(52.33, 0.77)	68.01 ± 12.86	0.71 ± 0.08

RFR achieved an average RMSE of 55.44 ± 16.18 mm and R^2 = 0.81 ± 0.07, indicating high predictive accuracy with low variability across folds. In contrast, SVR produced RMSE = 68.01 ± 12.86 mm and R^2 = 0.71 ± 0.08, with larger fluctuations across different forecasting horizons. This suggests that RFR is more consistent in capturing long-term rainfall patterns, whereas SVR is relatively more sensitive to temporal shifts.

Seasonal drift was estimated using a hybrid scheme, combining BMKG's climatological definitions (rainy: Nov–Apr; dry: May–Oct) with a data-driven threshold when months exceeding the long-term median rainfall (206.4 mm) are classified as rainy [44], [45]. This approach captures both operational seasonality and interannual variability to increase robustness and reproducibility. The Out-of-Distribution (OOD) Drift Test, summarized in table 6, further supports these findings. Under the best-performing scenario (80–20), RFR demonstrated high seasonal robustness, with negligible performance differences between wet and dry seasons (ΔR^2 = 0.002; R^2_rainy = 0.96; R^2_dry = 0.96). Conversely, SVR exhibited reduced performance during the rainy season (ΔR^2 = 0.005; R^2_rainy = 0.69; R^2_dry = 0.74) due to its sensitivity to extreme variability. Overall, these results confirm that RFR (Scenario 1) maintains higher stability against both temporal and seasonal drift, while SVR is more adaptive but less consistent—highlighting the importance of robust statistical feature engineering for stabilizing tropical rainfall forecasting models.

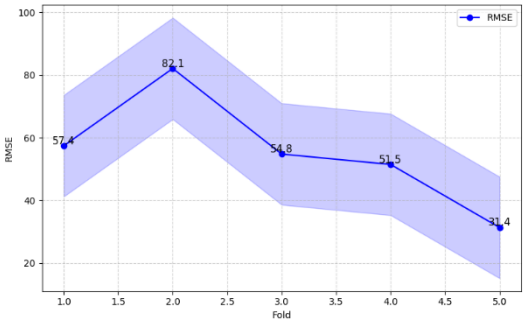
Table 6. Out-of-Distribution Drift Evaluation and the Best Scenario

Model	Scenario	Rainy		Dry		ΔR^2	Description
		RMSE	R^2	RMSE	R^2		
RFR	80 – 20	22.21	0.96	18.07	0.96	0.002	Nearly identical; the model remains highly stable across seasons
	70 – 30	42.35	0.87	37.79	0.85	0.01	Performs remains strong with minimal seasonal difference
	60 – 40	51.70	0.8	51.1	0.73	0.07	Minor drift is observed, but it remains withing acceptable stability limit
	50 – 50	52.04	0.8	51.1	0.73	0.06	Consistent with previous scenarios; RFR shows greater resilience to drift
SVR	80 – 20	65.53	0.69	49.68	0.74	0.05	Performance decreases during the rainy season, indicating sensitivity to high variability
	70 – 30	67.51	0.67	52.32	0.71	0.04	Still stable, but the same trend persists: better performance in the dry season
	60 – 40	78.42	0.55	55.13	0.68	0.13	Widening performance gaps indicate emerging seasonal drift
	50 – 50	78.08	0.56	55.88	0.67	0.11	Drift persists, with SVR showing reduced stability under high rainfall conditions

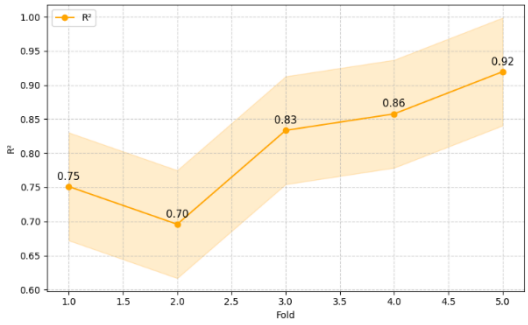
Figure 6 illustrates these patterns more clearly. Panels (a) and (b) show that RFR maintains a consistent downward RMSE trend and upward R^2 trajectory across folds, with minimal variability, indicating strong temporal stability and robustness to distributional changes. In contrast, SVR (panels c–d) shows larger performance fluctuations, particularly in fold 3, indicating greater sensitivity to temporal dynamics. Furthermore, in the seasonal evaluation (panels e–f), RFR

ISSN 2723-6471

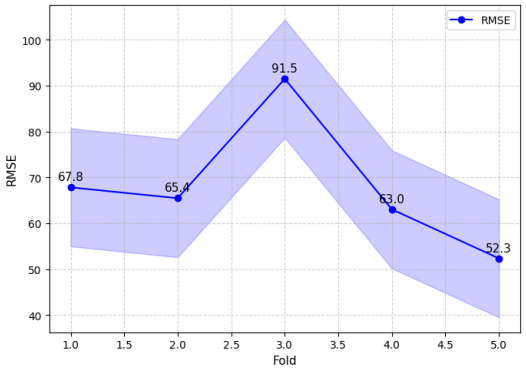
exhibits no significant differences in RMSE and R^2 between wet and dry seasons, confirming its resilience to seasonal variability. SVR, on the other hand, shows a more pronounced accuracy decline during the wet season, reflecting its sensitivity to extreme rainfall variability.



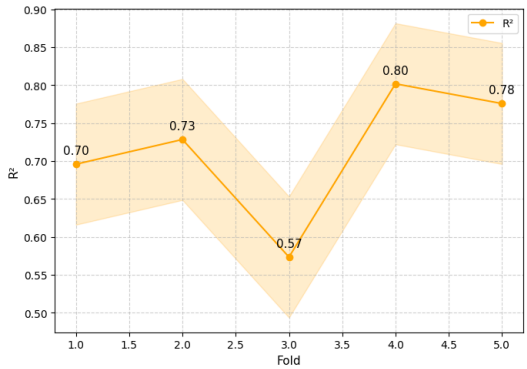
(a) Rolling CV – RMSE per Fold (RFR)



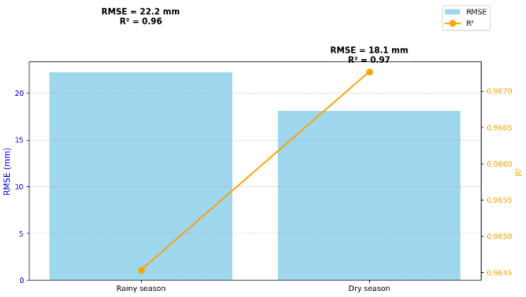
(b) Rolling CV – R^2 per Fold (RFR)



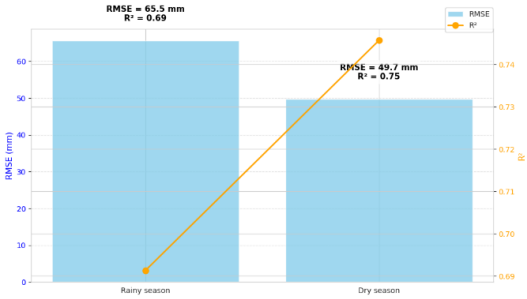
(c) Rolling CV – RMSE per Fold (SVR)



(d) Rolling CV – R^2 per Fold (SVR)



(e) Seasonal Drift Evaluation RFR



(f) Seasonal Drift Evaluation SVR

Figure 6. Rolling Cross-Validation and Drift Robustness Evaluation

3.5. Probabilistic Forecasting

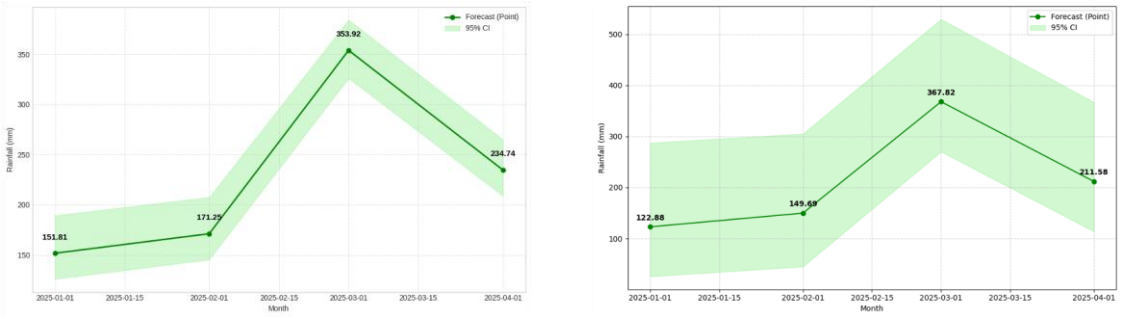
This phase adopts a bootstrap residual simulation approach (as many as 1000 times) to generate 95% confidence intervals (CIs) for each forecast [46], providing not only point estimates but also realistic uncertainty bounds that reflect potential fluctuations associated with extreme rainfall events. Table 7 shows that both RFR and SVR models project peak rainfall in March 2025, with forecast values ranging between 350–380 mm. However, the width of the SVR's CI (367.82–532.09 mm) is substantially larger than that of RFR (323.73–386.73 mm), indicating that RFR produces more stable and precise estimates. In contrast, SVR exhibits wider uncertainty ranges (average width 228.96 mm) compared to RFR (62.99 mm), reflecting its higher sensitivity to anomalies and extreme variability in the data.

Table 7. Confidence Interval (CI)

Months	RFR (mm)	Interval RFR (95% CI)	Gap CI:RFR	SVR (mm)	Interval SVR (95% CI)	Gap CI:SVR
Jan 2025	127.88	[100.76, 157.36]	56.60	122.88	[24.81, 264.87]	240.06

Feb 2025	157.83	[130.99, 198.36]	67.37	149.69	[51.86, 311.20]	259.34
Mar 2025	350.89	[323.73, 386.73]	63.00	367.54	[367.82, 532.09]	164.27
Apr 2025	240.77	[213.62, 276.61]	62.99	211.69	[113.69, 365.86]	252.17

Figure 7 illustrates these patterns visually. The green shaded areas represent the 95% confidence intervals. Panel (a) shows that RFR maintains narrow and consistent intervals across months, demonstrating the model's robustness to outliers and forecasting stability. Conversely, panel (b) indicates that SVR exhibits larger interval fluctuations, with a significant widening around March–April, pointing to a higher level of forecast uncertainty.



(a) Probabilistic Forecasting Comparison (RFR) (b) Probabilistic Forecasting Comparison (SVR)

Figure 7. Probabilistic Forecasting Comparison

3.6. Deployment Result

The results of the pipeline integration within the Gradio interface are displayed through an output display that estimates the forecasting scores, along with confidence intervals, and visualizes the trend for a selected month range. Figure 8 presents the deployment results of the two models for tropical rainfall forecasting over a two-year horizon (May 2025 – April 2027). The graph displays key forecast values and their 95% uncertainty ranges as shaded areas, allowing users to understand the model's probabilistic representation. All interface functions have been tested for output consistency and pipeline stability, ensuring a smooth integration. The RFR model (green plot) exhibits a smoother forecast curve with narrow uncertainty intervals, underscoring its predictive stability. In contrast, the SVR model (purple plot) shows higher fluctuations around the mean forecast values, reflecting its sensitivity to extreme variations. Overall, both models effectively capture the dynamics of tropical climatology; however, they differ fundamentally in terms of forecast stability and sensitivity to variability, with RFR being more stable and SVR more reactive to extremes.

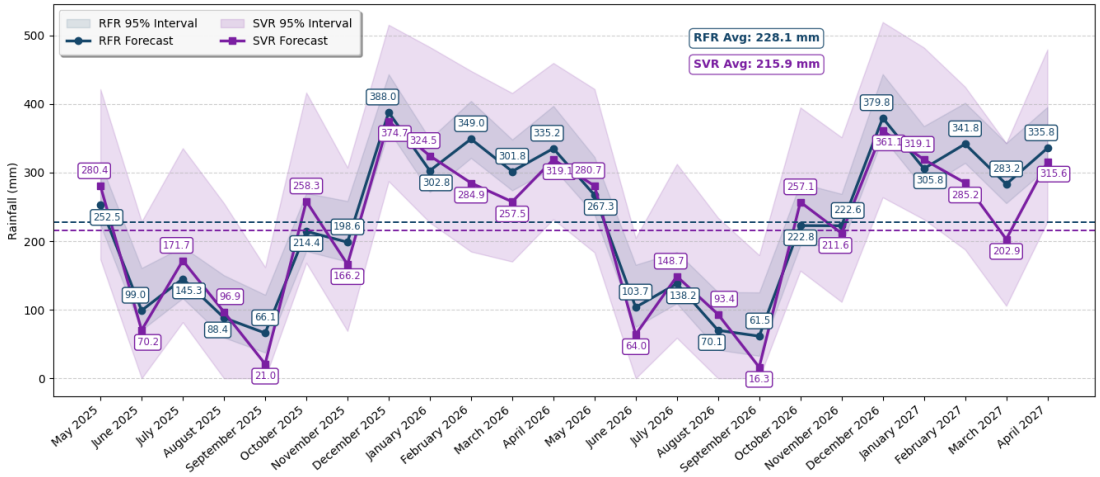


Figure 8. RFR and SVR Forecasting Results

4. Conclusion

This study advances tropical rainfall forecasting by developing a reproducible Auto-FE pipeline and demonstrating that robust statistical features can improve the performance and stability of non-sequential machine learning models.

Through a systematic evaluation of RFR and SVR across multiple temporal scenarios, the results underscore the value of temporal-seasonal feature design and robustness testing in capturing the dynamic characteristics of tropical rainfall. The incorporation of a probabilistic framework provides quantitative uncertainty estimates that complement point forecasts and enhance interpretability for scientific assessments. Collectively, the methodological contributions, including automated feature engineering, temporal robustness analysis, and uncertainty quantification, offer a foundation for methodological improvements in data-driven rainfall prediction. Future research should integrate multi-source predictors (e.g., satellite rainfall, climate indices, reanalysis products), extend the framework to spatial and multi-location forecasts, and explore hybrid architectures or deep learning to further enhance model generalization and potential applications in tropical climate studies.

5. Declarations

5.1. Author Contributions

Conceptualization: C.S., R.M.I., M.R.A., and A.W.; Methodology: C.S., M.R.A.; Software: R.M.I.; Investigation: A.W.; Resources: R.M.I.; Data Curation: R.M.I.; Validation: D.S.M. and W.B.Z; Writing Original Draft Preparation: C.S., R.M.I., and D.S.M.; Writing Review and Editing: C.S., M.R.A., and D.S.M.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

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

References

[1] F. Simanjuntak, I. Jamaluddin, T.-H. Lin, H. A. W. Siahaan, and Y.-N. Chen, "Rainfall forecast using machine learning with high spatiotemporal satellite imagery every 10 minutes," *Remote Sens. (Basel)*, vol. 14, no. 23, pp. 1–19, Nov. 2022, doi: 10.3390/rs14235950.

[2] A. J. Hill and R. S. Schumacher, "Forecasting excessive rainfall with random forests and a deterministic convection-allowing model," *Weather Forecast.*, vol. 36, no. 5, pp. 1693–1711, Oct. 2021, doi: 10.1175/WAF-D-21-0026.1.

[3] X. Zhao *et al.*, "A comprehensive review of methods for hydrological forecasting based on deep learning," *Water (Switzerland)*, vol. 16, no. 10, pp. 1–29, May 2024, doi: 10.3390/w16101407.

[4] E. Yanfatriani *et al.*, "Extreme rainfall trends and hydrometeorological disasters in tropical regions: Implications for climate resilience," *Emerg. Sci. J.*, vol. 8, no. 5, pp. 1860–1874, Oct. 2024, doi: 10.28991/ESJ-2024-08-05-012.

[5] F. I. W. Rohmat, A. J. Löhr, F. Pratama, N. S. Burnama, and A. A. Kuntoro, "Quantifying time-dependent flood resilience index in a densely populated urban environment in Manado, Indonesia," *Int. J. Disaster Risk Reduct.*, vol. 116, pp. 1–18, Jan. 2025, doi: 10.1016/j.ijdrr.2024.105112.

[6] R. Ramadhan *et al.*, "Trends in rainfall and hydrometeorological disasters in new capital city of Indonesia from long-term satellite-based precipitation products," *Remote Sens. Appl.*, vol. 28, pp. 1–14, Nov. 2022, doi: 10.1016/j.rsase.2022.100827.

[7] E. Hermawan *et al.*, "Large-scale meteorological drivers of the extreme precipitation event and devastating floods of early-February 2021 in Semarang, Central Java, Indonesia," *Atmosphere (Basel)*, vol. 13, no. 7, pp. 1–20, Jul. 2022, doi: 10.3390/atmos13071092.

[8] H. Shen, B. A. Tolson, and J. Mai, "Time to update the split-sample approach in hydrological model calibration," *Water Resour. Res.*, vol. 58, no. 3, pp. 1–26, Mar. 2022, doi: 10.1029/2021WR031523.

[9] P. O. Bojang, T. C. Yang, Q. B. Pham, and P. S. Yu, "Linking singular spectrum analysis and machine learning for monthly rainfall forecasting," *Appl. Sci. (Switzerland)*, vol. 10, no. 9, pp. 1–20, May 2020, doi: 10.3390/app10093224.

[10] O. Ejike, D. Ndzi, and M. Z. Shakir, "Comparative study of machine learning-based rainfall prediction in tropical and temperate climates," *Climate*, vol. 13, no. 8, pp. 1–27, Aug. 2025, doi: 10.3390/cli13080167.

[11] L. T. Pham, L. Luo, and A. Finley, "Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds," *Hydrol. Earth Syst. Sci.*, vol. 25, no. 6, pp. 2997–3015, Jun. 2021, doi: 10.5194/hess-25-2997-2021.

[12] B. Goehry, H. Yan, Y. Goude, P. Massart, and J. M. Poggi, "Random forests for time series," *REVSTAT—Stat. J.*, vol. 21, no. 2, pp. 283–302, Jun. 2023, doi: 10.57805/revstat.v21i2.400.

[13] P. Gajewski, B. Čule, and N. Rankovic, "Unveiling the power of ARIMA, support vector and random forest regressors for the future of the Dutch employment market," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 3, pp. 1365–1403, Jul. 2023, doi: 10.3390/jtaer18030069.

[14] S. Hia, H. Kuswanto, and D. D. Prasetyo, "Robustness of support vector regression and random forest models: A simulation study," *Data Sci. Emerg. Technol.*, pp. 465–479, Sep. 2023, doi: 10.1007/978-981-99-0741-0_33.

[15] S. Bhoopathi, K. Sumanth, L. Akanksha, and L. Pal, "Evaluating the performance of ANN, SVR, RF, and XGBoost in the prediction of maximum temperature and heat wave days over Rajasthan, India," *J. Hydrol. Eng.*, vol. 29, no. 6, pp. 1–16, Jun. 2024, doi: 10.1061/JHYEFF.HEENG-6243.

[16] M. Abbasi, A. Farokhnia, M. Bahreinimotlagh, and R. Roozbahani, "A hybrid of random forest and deep auto-encoder with support vector regression methods for accuracy improvement and uncertainty reduction of long-term streamflow prediction," *J. Hydrol. (Amst.)*, vol. 597, pp. 1–16, Jun. 2021, doi: 10.1016/j.jhydrol.2020.125717.

[17] P. Cihan, "Comparative performance analysis of deep learning, classical, and hybrid time series models in ecological footprint forecasting," *Appl. Sci. (Switzerland)*, vol. 14, no. 4, pp. 1–18, Feb. 2024, doi: 10.3390/app14041479.

[18] M. A. Saleh, H. M. Rasel, and B. Ray, "A comprehensive review towards resilient rainfall forecasting models using artificial intelligence techniques," *Green Technol. Sustain.*, vol. 2, no. 3, pp. 1–21, Sep. 2024, doi: 10.1016/j.grets.2024.100104.

[19] F. R. Aderyani, S. J. Mousavi, and F. Jafari, "Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN," *J. Hydrol. (Amst.)*, vol. 614, pp. 1–17, Nov. 2022, doi: 10.1016/j.jhydrol.2022.128463.

[20] D. Li *et al.*, "Prediction of rainfall time series using the hybrid DWT-SVR-Prophet model," *Water (Switzerland)*, vol. 15, no. 10, pp. 1–18, May 2023, doi: 10.3390/w15101935.

[21] M. Čistý, M. Danko, S. Kohnová, B. Považanová, and A. Trizna, "Machine learning enhanced by feature engineering for estimating snow water equivalent," *Water (Switzerland)*, vol. 16, no. 16, pp. 1–17, Aug. 2024, doi: 10.3390/w16162285.

[22] W. Fang *et al.*, "An evaluation of random forest based input variable selection methods for one month ahead streamflow forecasting," *Sci. Rep.*, vol. 14, no. 1, pp. 1–12, Dec. 2024, doi: 10.1038/s41598-024-81502-y.

[23] J. H. Stagge, D. E. Rosenberg, A. M. Castronova, A. Ostfeld, and A. S. Jones, "Journal of Water Resources Planning and Management's reproducibility review program: Accomplishments, lessons, and next steps," *J. Water Resour. Plan. Manag.*, vol. 150, no. 8, pp. 1–4, Aug. 2024, doi: 10.1061/JWRMD5.WRENG-6559.

[24] C. R. Corona and T. S. Hogue, "Machine learning in stream and river water temperature modeling: A review and metrics for evaluation," *Hydrol. Earth Syst. Sci.*, vol. 29, no. 12, pp. 2521–2549, Jun. 2025, doi: 10.5194/hess-29-2521-2025.

[25] R. A. Collenteur *et al.*, "Data-driven modelling of hydraulic-head time series: Results and lessons learned from the 2022 groundwater time series modelling challenge," *Hydrol. Earth Syst. Sci.*, vol. 28, no. 23, pp. 5193–5208, Dec. 2024, doi: 10.5194/hess-28-5193-2024.

[26] Y. Cai *et al.*, "Predicting tropical cyclone extreme rainfall in Guangxi, China: An interpretable machine learning framework addressing class imbalance and feature optimization," *Meteorol. Appl.*, vol. 32, no. 3, pp. 1–23, May 2025, doi: 10.1002/met.70052.

[27] B. J. Kim, M. Kim, J. Yoo, and B. Kim, "Rapid simulation for real-time flood depth prediction using support vector machine," *Sci. Rep.*, vol. 15, pp. 1–14, Dec. 2025, doi: 10.1038/s41598-025-17090-2.

[28] A. Kumar *et al.*, "World Meteorological Organization (WMO)-accredited infrastructure to support operational climate prediction," *Bull. Amer. Meteorol. Soc.*, vol. 105, no. 11, pp. E2126–E2143, Nov. 2024, doi: 10.1175/BAMS-D-23-0284.1.

[29] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data Knowl. Eng.*, vol. 139, pp. 1–14, May 2022, doi: 10.1016/j.datak.2022.102013.

[30] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, and J. R. Coronado-Hernández, "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia—Saber 11 test," *Procedia Comput. Sci.*, vol. 198, pp. 512–517, Jan. 2022, doi: 10.1016/j.procs.2021.12.278.

[31] U. Kannengiesser and J. S. Gero, "Modelling the design of models: An example using CRISP-DM," *Proc. Des. Soc.*, pp. 2705–2714, Jul. 2023, doi: 10.1017/pds.2023.271.

[32] E. Valla, S. Nõmm, K. Medijainen, P. Taba, and A. Toomela, "Tremor-related feature engineering for machine learning-based Parkinson's disease diagnostics," *Biomed. Signal Process. Control*, vol. 75, pp. 1–12, May 2022, doi: 10.1016/j.bspc.2022.103551.

[33] M. S. Ahammed *et al.*, "Hypertuning random forest for enhancing cyber security in industrial power electronics," *Int. J. Comput. Inf. Technol.*, pp. 2529–2533, Dec. 2024, doi: 10.1109/ICCIT64611.2024.11022494.

[34] M. Loki, A. Mindila, and W. Cheruiyot, "Hyper-parameter optimization of kernel functions on multi-class text categorization: A comparative evaluation," *WIREs Data Min. Knowl. Discov.*, vol. 15, no. 1, pp. 1–15, Mar. 2025, doi: 10.1002/widm.1572.

[35] S. Peng, W. Wang, Y. Chen, X. Zhong, and Q. Hu, "Regression-based hyperparameter learning for support vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18799–18813, Dec. 2024, doi: 10.1109/TNNLS.2023.3321685.

[36] Y.-G. Wang, J. Wu, Z.-H. Hu, and G. J. McLachlan, "A new algorithm for support vector regression with automatic selection of hyperparameters," *Pattern Recognit.*, vol. 133, pp. 1–14, Jan. 2023, doi: 10.1016/j.patcog.2022.108989.

[37] X. Liu, W. Chen, W. Du, P. Li, and X. Wang, "Application of artificial intelligence and machine learning in lung transplantation: A comprehensive review," *Front. Digit. Health*, pp. 1–18, May 2025, doi: 10.3389/fdgth.2025.1583490.

[38] Y. Okadome and Y. Nakamura, "Feature extraction method using lag operation for sub-grouped multidimensional time series data," *IEEE Access*, vol. 12, pp. 98945–98959, Jul. 2024, doi: 10.1109/ACCESS.2024.3429529.

[39] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philos. Trans. R. Soc. A*, vol. 379, no. 2194, pp. 1–14, Apr. 2021, doi: 10.1098/rsta.2020.0209.

[40] K. Yi *et al.*, "A survey on deep learning-based time series analysis with frequency transformation," *ACM SIGKDD Explor. Newsl.*, pp. 6206–6215, Aug. 2025, doi: 10.1145/3711896.3736571.

[41] D. Dimoudis, T. Vafeiadis, A. Nizamis, D. Ioannidis, and D. Tzovaras, "Utilizing an adaptive window rolling median methodology for time series anomaly detection," *Procedia Comput. Sci.*, pp. 584–593, Dec. 2022, doi: 10.1016/j.procs.2022.12.254.

[42] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023, doi: 10.1109/TPAMI.2022.3195549.

[43] M. Akrout, A. Feriani, F. Bellili, A. Mezghani, and E. Hossain, "Domain generalization in machine learning models for wireless communications: Concepts, state-of-the-art, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 3014–3037, Oct. 2023, doi: 10.1109/COMST.2023.3326399.

[44] R. Ramadhan, M. Marzuki, W. Suryanto, S. Sholihun, H. Yusnaini, and R. Muharsyah, "Rainfall variability in Indonesia new capital associated with the Madden–Julian Oscillation and its contribution to flood events," *Quaternary Sci. Adv.*, vol. 13, pp. 1–12, Jan. 2024, doi: 10.1016/j.qsa.2024.100163.

ISSN 2723-6471

[45] M. E. Mateling, C. Pettersen, M. S. Kulie, K. S. Mattingly, S. A. Henderson, and T. S. L'Ecuyer, "The influence of atmospheric rivers on cold-season precipitation in the Upper Great Lakes region," *J. Geophys. Res. Atmos.*, vol. 126, no. 13, pp. 1–18, Jul. 2021, doi: 10.1029/2021JD034754.

[46] M. Gu, Y. Lin, V. C. Lee, and D. Y. Qiu, "Probabilistic forecast of nonlinear dynamical systems with uncertainty quantification," *Physica D*, vol. 457, pp. 1–14, Jan. 2024, doi: 10.1016/j.physd.2023.133938.