Current **A**pplied **S**cience and **T**echnology e-ISSN: 2586-9396

Curr. Appl. Sci. Technol. ….., Vol... (No...), e0267316

Research article

Development of a Hybrid Model for Rainfall Forecasting in Northeastern Thailand: Integration of Seasonal Autoregressive Integrated Moving Average, Support Vector Regression, and Variational Mode Decomposition

Thanakon Sutthison*, Somporn Thepchim and Yaovaruk Thongphum

Program of Mathematics, Faculty of Science, Ubon Ratchathani Rajabhat University, Ubon Ratchathani, 34000, Thailand

Received: 22 April 2025, Revised: 11 August 2025, Accepted: 22 October 2025, Published: 5 February 2026

Abstract

This study presents a hybrid model integrating Seasonal Autoregressive Integrated Moving Average (SARIMA), Ensemble Variational Mode Decomposition (EVMD), and Support Vector Regression (SVR) to improve monthly rainfall forecasting in Northeastern Thailand. The approach addresses the challenges posed by the non-stationary and nonlinear nature of rainfall data. SARIMA is first applied to extract linear components, while EVMD is used to decompose residuals into Intrinsic Mode Functions (IMFs). Each IMF and the remaining residuals are forecasted using SVR. A dataset comprising 496 months of rainfall records (January 1983 to April 2024) from 12 meteorological stations under the Thai Meteorological Department was used. Model performance was evaluated using five statistical metrics: RMSE, PRMSE, RPD, MAE, and R^2. The hybrid SARIMA-EVMD-SVR model consistently outperformed SARIMA and SVR standalone models, achieving R^2 values above 0.84 and RPD values greater than 2.5 in most stations. The hybrid model improved forecasting accuracy by up to 39.26% over SVR and 36.11% over SARIMA. The results highlight the model's ability to effectively capture complex rainfall dynamics. Its adaptability offers potential for application in other time series forecasting tasks, contributing to enhanced decision-making in water resource planning and climate-related policy development.

Keywords: hybrid forecasting model; rainfall forecasting; SARIMA; EVMD; support vector regression

1. Introduction

Monthly rainfall forecasting plays a pivotal role in risk management, emergency planning, and policy formulation across agricultural, economic, and societal domains. It serves as a critical tool in mitigating losses from natural disasters such as floods and droughts (Akhtar et al., 2023; Pirone et al., 2023). Moreover, it is fundamental in the management of water

*Corresponding author: E-mail: thanakon.s@ubru.ac.th
https://doi.org/10.55003/cast.2026.267316

resources, particularly for agricultural irrigation, transportation systems, and urban drainage networks. Accurate rainfall forecasts facilitate water allocation planning, flood mitigation, and the optimization of water resource management strategies (Abebe & Endalie, 2023; He et al., 2024). The challenges associated with rainfall forecasting stem from the complexity of meteorological data, which encompasses multidimensional factors such as atmospheric circulation patterns, temperature, humidity, and wind speed. Additionally, the inherent nonlinear characteristics of rainfall variations pose significant obstacles to prediction accuracy (Zhou et al., 2021; Johny et al., 2022; Yin et al., 2024). The impact of climate change further exacerbates these challenges, resulting in altered rainfall patterns, increased frequency of extreme precipitation events, and heightened flooding risks. Consequently, precise rainfall forecasting has become indispensable for effective disaster preparedness and risk mitigation (He et al., 2022; Skarlatos et al., 2023). Beyond its role in water resource management, rainfall also plays a crucial part in hydropower production. The design, operation, and optimization of hydropower plants rely heavily on accurate rainfall and runoff data to evaluate resource potential and plan for sustainable energy generation. Reliable rainfall forecasts enhance energy production planning and support stable responses to fluctuating electricity demand (Skarlatos et al., 2023). Rainfall forecasting is a critical tool in various sectors, including water resource management, agriculture, and natural resource planning. Forecasting methodologies can be broadly categorized into three main approaches:

1. Numerical Weather Prediction (NWP): This technique employs physical principles to simulate atmospheric processes. Despite its scientific robustness, NWP is resource-intensive and requires substantial processing time, particularly in regions with dynamic weather patterns, which limits its applicability in certain areas (Abebe & Endalie, 2023; He et al., 2024).

2. Radar-Based Forecasting: This approach is effective for short-term rainfall prediction and delivers rapid results. However, its accuracy heavily depends on the quality of radar data and is constrained by limited spatial coverage (He et al., 2022; Wang et al., 2022).

3. Historical Data-Based Forecasting: This method leverages historical meteorological data and time series to predict rainfall. It is widely used due to its flexibility, lower computational requirements, and cost-effectiveness, making it suitable for organizations with limited resources. Popular models in this category include stochastic models such as SARIMA and machine learning techniques like artificial neural networks (ANN) and support vector machines (SVM) (He et al., 2022; Wang et al., 2022). This study focuses on developing advanced rainfall time series forecasting models that aim to enhance predictive accuracy while minimizing resource consumption and computational demands. The ultimate objective is to provide an efficient and scalable forecasting framework that supports diverse applications across critical sectors.

A comprehensive review of the literature highlights persistent challenges in rainfall forecasting due to the volatile and complex nature of rainfall data, which encompasses both linear and nonlinear components. Traditional models, such as the Seasonal Autoregressive Integrated Moving Average (SARIMA), are effective in capturing linear patterns but lack the capability to address nonlinear dynamics. Conversely, artificial intelligence techniques such as Support Vector Regression (SVR) and Artificial Neural Networks (ANN) excel in modeling nonlinear relationships but are less effective in handling linear components (Chen et al., 2021; Alqahtani et al., 2023; Parviz & Ghorbanpour, 2024). To overcome these limitations, researchers have proposed hybrid models that integrate the strengths of traditional statistical approaches and artificial intelligence methods. These models have demonstrated superior forecasting accuracy by effectively combining linear

and nonlinear modeling capabilities. As a result, hybrid approaches have gained significant traction in rainfall forecasting research, consistently outperforming standalone models and offering a promising pathway for advancing predictive accuracy (Chen et al., 2021; He et al., 2022; Johny et al., 2022; Yin et al., 2024).

Forecasting rainfall time series, which often exhibits high complexity and volatility, poses significant challenges due to its non-stationary nature and the presence of noise. Non-stationary data is characterized by statistical properties, such as mean and variance, that change over time. For instance, rainfall during wet and dry seasons often demonstrates distinct differences, reflecting long-term climatic variations (Chen et al., 2021; Wu et al., 2024b). At the same time, noise, or external interference, arising from factors such as weather variability or measurement errors, obscures critical information and further complicates analysis and forecasting (Parsaie et al., 2024). These challenges necessitate data processing techniques capable of reducing noise and isolating long-term trends from non-stationary data. Methods such as decomposition techniques and the development of hybrid models, which combine data decomposition with parameter optimization, are particularly effective in addressing these complexities and enhancing predictive accuracy. Effectively managing non-stationary characteristics and noise is fundamental for accurate rainfall forecasting and understanding climatic variations. Advanced methodologies that integrate decomposition and model refinement hold promises for improving the reliability of rainfall predictions and addressing the inherent variability in climate-related data (Chen et al., 2021; He et al., 2022; Parsaie et al., 2024; Wu et al., 2024b). To address these issues, researchers have developed advanced data decomposition techniques, including Empirical Mode Decomposition (EMD), Wavelet Transform (WT), and Ensemble Empirical Mode Decomposition (EEMD), which are aimed at noise reduction and data stabilization. These methods, particularly the extraction of Intrinsic Mode Functions (IMFs), have facilitated the implementation of hybrid models capable of handling large datasets and improving predictive accuracy (Ali et al., 2020; Johny et al., 2022; Parsaie et al., 2024).

Among these methods, Variational Mode Decomposition (VMD) is widely regarded for its superior ability to process non-stationary signals compared to EMD, WT, and EEMD (He & Huang, 2023). However, the pre-determination of the number of modes (K) in VMD poses a significant challenge. Specifically, overestimating K can amplify noise due to excessive decomposition, thereby leading to the loss of essential structures in the original data. Conversely, underestimating K results in mode redundancy, which consequently obscures data trends and introduces analytical errors. Moreover, decomposing all modes simultaneously increases computational complexity, particularly because it involves solving variational problems, such as calculating the center frequencies and bandwidths of each mode. As a result, this process becomes increasingly resource-intensive, especially as the number of modes (K) rises (Zuo et al., 2020; He & Huang, 2023; Wang et al., 2024a). To address these challenges, previous studies have proposed iterative optimization methods that initialize and update mode components, center frequencies, and Lagrange multipliers to obtain optimal parameter values (Zuo et al., 2020). This study adopts such an approach to streamline the decomposition process and improve the accuracy of signal component separation.

However, the application of hybrid decomposition-based algorithms to rainfall time series remains underexplored, highlighting the need for further research to evaluate their potential in this field. Recent studies on rainfall prediction have demonstrated continuous progress, particularly through the use of hybrid decomposition-based algorithms that integrate signal decomposition techniques with learning models to improve forecasting accuracy and reliability. For example, Guo et al. (2023) utilized an EMD-VMD-LSTM model

to address discontinuities in time series data, whereas Zhang and Wu (2023) developed the CEEMD-ELM-FFOA model to improve forecasting efficiency. Similarly, Jiang (2023) integrated CEEMD with LSTM for monthly rainfall prediction. In addition, Pinheiro and Ouarda (2023) employed EANN, leveraging climate indices to forecast rainfall in Brazil. Jamei et al. (2023) proposed the TVF-EMD-SVD-EDBi-LSTM model to handle increasingly complex datasets. Shao et al. (2024) introduced the IDSTA-TCN model, combining Enhanced Bayesian-Optimized Variational Mode Decomposition (EBO-VMD) and a spatial-temporal attention mechanism to forecast rainfall in Qijiang and Tunxi. Furthermore, Wu et al. (2024a) advanced VMD by integrating it with a Gated Convolutional Network, enhanced with Gelu activation and an attention mechanism to reduce variability in data. Also, Hou et al. (2024) used CEEMDAN-VMD-BiLSTM to make forecasting better on complicated datasets, and Zheng et al. (2024) used SMFSD-AVOA-LSSVM to predict monthly rainfall. Rezaiy and Shabri (2024) used W-EEMD-ARIMA and EEMD-SVM to forecast the SPI index. Additionally, Parviz and Ghorbanpour (2024) implemented EMD-MODWT-SARIMA-SVR to address complex drought-related data. In Thailand, Waqas et al. (2024) used the Biorthogonal Wavelet Transform (BWT) along with LSTM-RNN and RBFNN to predict daily rainfall, showing that it works well in that area. Although techniques like VMD, EMD, EEMD, and CEEMDAN are effective in reducing noise and revealing hidden data structures, they are not without limitations. For instance, WT requires careful selection of wavelet functions, EMD and EEMD are prone to mode mixing, CEEMDAN demands high computational resources, and VMD requires accurate pre-specification of the number of modes. These studies not only demonstrate significant progress but also reveal existing limitations, offering valuable insights for the development of new techniques to improve the performance and reliability of rainfall forecasting in the future. Consequently, addressing these limitations could facilitate more robust and scalable methodologies, which are critical for tackling increasingly complex environmental challenges.

Based on the identified research gap, the hybrid model SARIMA-EVMD-SVR was developed to address the challenges of monthly rainfall data characterized by non-stationarity and non-linearity. This model utilizes historical data from twelve meteorological observation stations in the Northeastern region of Thailand, under the supervision of the Thai Meteorological Department (TMD). This region was selected as the study area for several compelling reasons. First, the Northeastern region heavily relies on seasonal rainfall for agriculture, which serves as the primary source of income for the majority of its population. Second, the region is particularly vulnerable to the impacts of climate variability, such as droughts and floods, which severely affect agricultural productivity and water resource management. Therefore, accurate and reliable rainfall forecasting is critically important for supporting early warning systems, efficient agricultural planning, and data-driven policymaking in this region. Furthermore, the availability of high-quality, long-term monthly rainfall data from the TMD enhances both the feasibility and credibility of the analytical framework within this geographical context. It is specifically designed to enhance both the accuracy and reliability of rainfall forecasting. The modeling process begins with the Seasonal Autoregressive Integrated Moving Average (SARIMA) to extract linear components, such as trends and seasonal patterns. The residuals, representing non-linear components, are then processed using Variational Mode Decomposition (VMD) to decompose the data into Intrinsic Mode Functions (IMFs). This decomposition reduces noise and improves the structural representation of the data. Subsequently, each IMF is forecasted using Support Vector Regression (SVR), a robust method well-suited for capturing complex and non-linear relationships. The performance of the proposed model is evaluated by comparing it against traditional models including SARIMA and SVR using a set of statistical evaluation metrics. These include root mean square error (RMSE),

relative root mean square error (RRMSE), residual predictive deviation (RPD), mean absolute error (MAE), accuracy improvement (AI), and coefficient of determination (r^2). These metrics ensure a comprehensive assessment of the model's accuracy and reliability across various dimensions of forecast performance. The SARIMA-EVMD-SVR model not only enhances the precision of rainfall forecasting but also exhibits adaptability for application to other complex time series datasets, such as electricity demand forecasting and climate change projections. Furthermore, it offers significant potential for supporting evidence-based policymaking and sustainable resource management over the long term.

2. Materials and Methods

2.1 Study area and dataset

This study utilizes historical monthly rainfall data from 12 meteorological stations in the Northeastern region of Thailand, under the supervision of the Thai Meteorological Department (TMD). The dataset spans from January 1983 to April 2024, covering a total of 496 months (unit: mm). Figure 1 shows the locations of TMD meteorological stations in the Northeastern region of Thailand. Before performing descriptive statistical analysis, we conducted a systematic data validation process to check for missing values and outliers, ensuring data accuracy (Waqas et al., 2024). Table 1 presents the descriptive statistics of monthly rainfall data across 12 TMD stations.

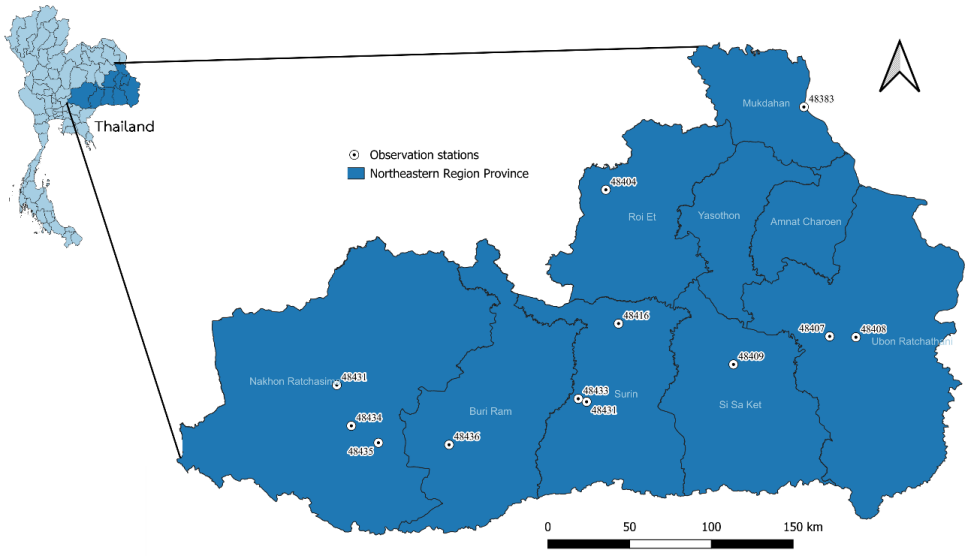


Figure 1. Distribution of observation stations over the northeastern region of Thailand

2.2 Data preparation

The monthly rainfall data from 12 meteorological stations in northeastern Thailand, under TMD supervision (Table 1 and Figure 1), exhibit non-linear characteristics. This study applies decomposition techniques and single models integrated with hybrid models to improve forecasting accuracy. The models include SARIMA, SVR, and hybrid models, such

Table 1. General ınformation on the study area and monthly rainfall data from January 1983 to April 2024

Observation Station Name	Abbreviation	Station Code	Median	Mean	SD	Min	Max	CV
Mukdahan	MK	48383	67.2	120.16	139.8	0	803.3	116.35
Roi Et	RE	48404	73.35	112.79	120.38	0	661.1	106.73
Ubon Ratchathani (Center)	UBC	48407	92.7	137.56	147.51	0	653.9	107.23
Ubon Ratchathani (Agro)	UBA	48408	79.55	134.08	146.52	0	715.8	109.28
Sisaket	SK	48409	79.7	120.16	132.02	0	827.5	109.87
Surin (Tha Tum)	SURT	48416	76.2	109.94	111.92	0	549.6	101.8
Surin	SUR	48431	95.4	119.21	119.2	0	676.3	99.99
Surin (Agro)	SURA	48433	86.4	118.38	120.45	0	651.5	101.75
Nakhon Ratchasima	NKR	48431	63.05	92.12	93.74	0	546.1	101.76
Nakhon Ratchasima (Pak Chong)	NKRP	48435	76.75	95.55	88.21	0	435	92.31
Nakhon Ratchasima (Chok Chai)	NKRC	48434	66.75	88.99	86.63	0	428.2	97.34
Buriram (Nang Rong)	BURN	48436	80.35	102.5	97.97	0	419.2	95.58

as SARIMA-ESVMD-SVR. Proper data preprocessing is crucial for SVR and hybrid models. This study employs the Partial Autocorrelation Function (PACF) to select input variables for the SVR model (see Section 2.6). Then, the Min-Max Scaler is applied to normalize the data to a 0–1 range before training (Sutthison, 2024). After preprocessing, the dataset is divided into two groups. Group 1 (70% or 347 monthly observations) was used for model training, while Group 2 (30% or 149 samples) was used for validation without randomization (Parsaie et al., 2024). This approach ensures that test results reflect the actual temporal sequence. The transformation is explained as follows:

$$\tilde{x}_i^c = \frac{x_i^c - x_{min}^c}{x_{max}^c - x_{min}^c}, \quad (1)$$

Where: $\tilde{x}_i^c$ is the normalize value of the i - th data point for c - th feature.

x_i^c is the original value of the i - th data point for the c - th feature.

x_{min}^c is the minimum value of the c -th feature across all data points.

x_{max}^c is maximum value of the c -th feature across all data points.

2.3 Variational mode decomposition (VMD)

The Variational Mode Decomposition (VMD) algorithm, proposed by Dragomiretskiy and Zosso (2014), is used to decompose complex signals into band-limited intrinsic mode functions (IMFs). These IMFs, derived from the original time series $f(t)$, are amplitude-modulated and frequency-modulated (AM-FM) components and can be represented by the following formula (Zuo et al., 2020).

$$u_k(t) = A_k(t)\cos(\phi_k(t(x))) \tag{2}$$

where:

$u_k(t)$ is the k-th intrinsic mode function (IMF) extracted from the original signal.

$A_k(t)$ is the instantaneous amplitude of the k-th IMF.

$\phi_k(t)$ is the instantaneous phase of the k-th IMF, where x is is the input time series variable.

The approach proposed by Dragomiretskiy and Zosso (2014) estimates the bandwidth of each mode through a variational framework. Initially, the Hilbert transform is applied to derive the analytic signal for each mode $u_k(t)$, enabling a one-sided frequency spectrum. This spectrum is shifted to the baseband using frequency adjustment based on the estimated center frequency. The bandwidth is then quantified by the squared L^2-norm of the gradient, weighted by the H^1 Gaussian smoothness of the demodulated signal. The entire process is formulated as a constrained variational optimization problem for the original time series $f(t)$.

$$\begin{cases} \min\limits_{\{u_k\},\{\omega_k\}} = \left\{ \sum_k \| \partial_t \left[\left(\delta(t) + \dfrac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega kt} \|_2^2 \right\} \\ s.t. \sum_k u_k(t) = f(t) \end{cases} \tag{3}$$

where:

$u_k(t)$ is the k-th intrinsic mode function (IMF) to be extracted from the original signal.

ω_k is the estimated center frequency of the k-th mode.

$\delta(t)$ is Dirac delta function.

$*$ is the convolution operator.

j is the imaginary unit, where $j = \sqrt{-1}$.

∂_t is partial derivative with respect to time t.

$\|.\|_2$ is the L^2 - norm, representing energy or bandwidth.

$f(t)$ is the original time – series signal to be decomposed.

Let $\{u_k(t)\} = \{u_1(t), u_2(t), ..., u_k(t)\}$ and $\{\omega_k\} = \{\omega_1, \omega_2, ..., \omega_k\}$ denote the decomposed modes and their corresponding center frequencies. Here, $*$ represents the convolution operator, t is time $j^2 = -1$ and δ is the Dirac delta function. To reformulate the constrained problem in equation (3) into an unconstrained form, a quadratic penalty term α and a Lagrange multiplier λ are introduced (Zuo et al., 2020). The augmented Lagrangian ℓ is then defined as follows:

$$\ell(\{u_k\},\{\omega_k\},\lambda) := \alpha \sum_k \left\| \partial_t \left[\left(\partial(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega k t} \right\|_2^2 + \| f(t) - \sum_k u_k(t) \|_2^2 + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle \quad (4)$$

where:

$\ell(\{u_k\},\{\omega_k\},\lambda)$ is the augmented Lagrangian function.

α is the quadratic penalty parameter controlling fidelity vs. decomposition.

$\lambda(t)$ is the Lagrange multiplier function enforcing reconstruction constraint.

$\langle .,. \rangle$ is the inner product over time domain.

$\|.\|_2^2$ is the squared L^2 -norm (energy of signal).

$f(t)$ is the original signal.

$\sum_k u_k(t)$ is the reconstructed signal from all IMFs.

In the VMD framework, the alternate direction method of multipliers (ADMM) is employed to solve the optimization problem defined in equation (4). The frequency-domain update of each mode $u_k(\omega)$ is carried out using equation (5), while the center frequency ω_k is recalculated via equation (6). The Lagrange multiplier λ is simultaneously updated by equation (7). In the time domain, each mode $u_k(t)$ is reconstructed by taking the real part of the inverse Fourier transform of $u_k(\omega)$ obtained from equation (5).

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i<k} \hat{u}_i^{n+1}(\omega) - \sum_{i>k} \hat{u}_i^n(\omega) + \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \quad (5)$$

$$\hat{\omega}_k^{n+1} = \frac{\int_0^{\infty} \omega \,|\, \hat{u}_k^{n+1}(\omega) \,|^2 \, d\omega}{\int_0^{\infty} |\, \hat{u}_k^{n+1}(\omega) \,|\, d\omega} \tag{6}$$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^{n}(\omega) + \tau\left(\hat{f}(\omega) - \sum_k \hat{u}_k^{n+1}(\omega) \right) \tag{7}$$

where:

$\hat{u}_k^{n+1}(\omega)$ is the updated frequency-domain representation of the k -th mode at iteration $n+1$.

$\hat{f}(\omega)$ is the Fourier transform of the input signal $f(t)$.

$\hat{\lambda}^{n}(\omega)$ is the Fourier-domain Lagrange multiplier at iteration n.

ω_k is the estimated center frequency of the k - th mode.

α is the quadratic penalty parameter.

ω is the frequency variable.

$\hat{\omega}_k^{n+1}$ is the updated center frequency of mode k -th mode at iteration $n+1$, computed via spectral centroid (equation 6).

τ is the parameter controlling update rate of the Lagrange multiplier.

Here, n denotes the iteration number, and τ is an iterative parameter that controls the noise tolerance in the VMD process. The terms $\hat{u}_k^{n+1}(\omega), \hat{f}(\omega)$ and $\hat{\lambda}^{n}(\omega)$ are the Fourier transforms of $\hat{u}_k^{n+1}(t)$, $f(t)$ and $\lambda^{n}(t)$, respectively. The iteration proceeds until the relative difference between two successive estimates satisfies the convergence criterion defined in equation (8), where ε represents the tolerance threshold:

$$\frac{\sum_k || \, \hat{u}_k^{n+1} - u_k^{n} \, ||^2}{|| \, \hat{u}_k^{n} \, ||_2^2} < \varepsilon \tag{8}$$

where:

$\hat{u}_k^{n+1}, \hat{u}_k^{n}$ is the frequency-domain representations of the k -th mode at iteration $n+1$ and n, respectively.

ε is the preset convergence tolerance threshold.

The effectiveness of the Variational Mode Decomposition (VMD) method depends on several key parameters, including the decomposition level (k), the quadratic penalty term (α), the noise tolerance (τ), and the convergence threshold (ε). Determining the appropriate number of intrinsic mode functions (IMFs) is not straightforward. Too few IMFs

may fail to capture essential signal features, while too many can increase computational complexity and introduce redundant components (Xu et al., 2019). The parameter α controls the bandwidth of each mode. A small α results in wider bandwidths, which may mix signal and noise, whereas a large α yields narrow bandwidths that could exclude relevant information (Xu et al., 2019). According to equations (5), (7) and (8), the Lagrange multiplier λ can hinder convergence, particularly when $\tau > 0$, leading to higher noise levels in the decomposed modes. Setting $\tau = 0$ may reduce this issue, although it might impair the accuracy of signal reconstruction. Additionally, the parameter ε influences the reconstruction error of the VMD process (Zuo et al., 2020). In this study, the VMD procedure begins with the initialization of modes, center frequencies, and the Lagrange multiplier. The components are iteratively updated using equation (5) for the mode spectrum, equation (6) for the center frequency, and equation (7) for the Lagrange multiplier, until the convergence condition in equation (8) is satisfied. The final time-domain modes are then obtained by applying the inverse Fourier transform to the frequency-domain representations (Zuo et al., 2020).

2.4 Seasonal Autoregressive Integrated Moving Average (SARIMA)

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a robust statistical tool for forecasting time series data with seasonal patterns (Sutthison, 2024; Ayiah-Mensah et al., 2025). It is widely applied in rainfall prediction. A key requirement for using SARIMA is that the time series must be stationary, exhibiting constant mean and variance. SARIMA consists of six components: Autoregressive (AR), Integration (I), Moving Average (MA), Seasonal Autoregressive (SAR), Seasonal Integration (SI), and Seasonal Moving Average (SMA), represented by the parameters p, d, q, P, D, and Q, respectively. The model is generally expressed as SARIMA (p, d, q) (P, D, Q) [m], where m denotes the seasonal period-commonly set to 12 for monthly data. The model formulation is presented in equation (9).

$$\phi_p(B)\Phi_P(B^m)\nabla^d\nabla_m^D Y_t = \theta_q(B)\Theta_Q(B^m)a_t \quad (9)$$

$\nabla^d\nabla_m^D Y_t$ represents the series adjusted for both regular and seasonal differences. B denotes the backshift operator, while a_t refers to Gaussian white noise with a mean of zero and variance σ^2. The construction process of the SARIMA model includes the following details:

2.4.1 Preparing Rainfall Time Series Data (refer to Step 2.2: Data preparation for details)

2.4.2 The Dickey-Fuller test is used to check whether time series data is stationary. If it finds non-stationarity, logarithmic transformation and differencing are used before the model is specified.

2.4.3 In step 2.4.2, stationary time series data are used to program Python commands for iteratively selecting parameters p, q, P, and Q within the range of 0 to 2. Model selection is based on minimizing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with the model yielding the lowest AIC and BIC values considered optimal.

$$AIC = 2k - 2\log(L) \tag{10}$$

$$BIC = \log(n)k - 2\log(L) \tag{11}$$

where: n is the number of observations in the data
k is the number of parameters in the model
L is the likelihood of the model

2.4.4 Following model selection in step 2.4.3, forecasting and parameter estimation are conducted. Model diagnostics are performed using Python's 'statsmodels' package for automatic adequacy assessment. Further details are provided in the Results and Discussion section.

2.4.5 The model validated in step 2.4.4 is used for forecasting and performance evaluation.

2.5 Support Vector Regression (SVR)

SVR is a machine learning technique applied in classification, regression, pattern recognition, and probability density estimation (Parviz et al., 2023; Someetheram et al., 2025). The general SVR model is defined as:

$$y = f(x) = w^T \varphi(x) + b \tag{12}$$

Here, w represents the weight vector, $\varphi(x)$ is the kernel function mapping x_i to a high-dimensional space, and b is an adjustable parameter (Someetheram et al., 2025).

$$\text{Min}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$\text{Subject to}\begin{cases} y_i - w^T\varphi(x_i) - b \le \varepsilon + \xi_i \\ y_i - w^T\varphi(x_i) - b \ge -\varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \quad i = 1,...,n \end{cases} \tag{13}$$

w denotes the weight vector, C is the penalty parameter, and ε represents the insensitive loss function. Slack variables are denoted by ξ_i and ξ_i^* with n representing the number of training samples. The penalty parameter in SVR regulates tolerance to systematic outliers. An essential step in training an SVR model is selecting the optimal kernel function. This study utilizes the Gaussian Radial Basis Function (RBF) kernel, a widely adopted method in time series used for forecasting due to its superior ability to model complex, nonlinear patterns (Parviz et al., 2023).

Monthly rainfall forecasting using an SVR model requires the transformation of the data into $d = \{x_i, y_i\}_{i=1}^{l}$ where x_i are input features and y_i are target values. The transformed data is split into a training set and a testing set (see Section 2.2 for details).

The Partial Autocorrelation Function (PACF) is used to select input variables for the SVR model (see Section 2.6 and Figure 2). The model's performance depends on selecting optimal parameters such as C , epsilon (ε), and gamma (γ), making hyperparameter tuning essential to improving accuracy. This study uses the Python package Optuna, a hyperparameter optimization library that is more efficient and faster than traditional methods like random search and grid search for finding optimal parameters. The rainfall data is scaled using MinMaxScaler to normalize values within the [0–1] range, enhancing the SVR model's learning capability. The scaled data is then restructured into time series sequences for model training and testing. Key parameters optimized using Optuna include C , randomly selected from [10^{-3}, 10^{3}]; epsilon, from [10^{-4}, 10^{3}]; and gamma, from [10^{-5}, 10^{3}]. Optuna evaluates each parameter set through an objective function using 5-fold cross-validation to ensure reliable results, with Mean Squared Error (MSE) serving as the evaluation metric. The optimization process runs for 1000 trials to comprehensively explore the parameter space. Once optimal parameters are identified, the SVR model is trained on the training dataset and applied to the testing dataset for monthly rainfall forecasting. The PACF measures the direct correlation between a time series and its lagged values while removing the influence of intermediate lags. This method helps identify significant lags for time series modeling (Xiang et al., 2018). A brief overview of PACF is presented below. The j th regression coefficient in an autoregressive model at lag k is denoted as $\phi_{k,j}$. The autoregressive model with lag k is formulated as:

$$y_t = \phi_{k1} y_{t-1} + \phi_{k2} y_{t-2} + \ldots + \phi_{kk} y_{t-k} + u_t \tag{14}$$

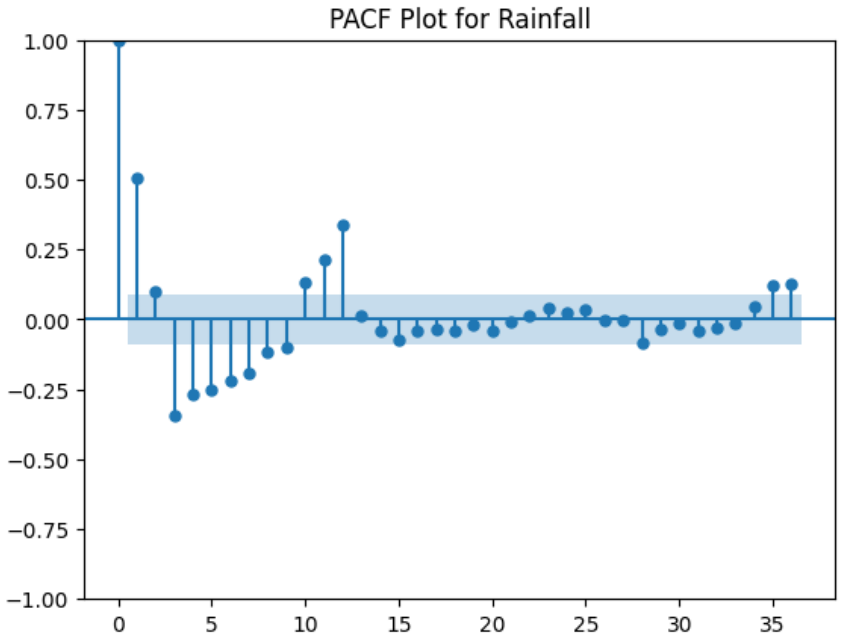


Figure 2. PACF plot for Mukdahan (MK) Rainfall Data 2.6 using Partial Autocorrelation Function (PACF)

where y_t represents the time series value at time t and $\phi_{k,k}$ is the last regression coefficient, corresponding to lag k . This coefficient, derived from PACF, measures the direct autocorrelation between y_t and y_{t-k} while excluding the influence of intermediate lags. The coefficient $\phi_{k,k}$ is calculated as follows.

$$\begin{cases} \phi_{11} = \rho_1 \\ \phi_{k+1,k+1} = (\rho_{k+1} - \sum_{j=1}^{k} \rho_{k+1-j}\phi_{kj})(1 - \sum_{j=1}^{k} \rho_j \phi_{kj})^{-1} \\ \phi_{k+1,j} = \phi_{kj} - \phi_{k+1,k+1} \bullet \phi_{k,k-j+1}, j = 1,2,...,k \end{cases} \quad (15)$$

The autocorrelation function at lag is given by:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (16)$$

where γ_k represents the covariance between y_t and y_{t-k} expressed as:

$$\gamma_k = \mathrm{cov}(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)] \quad (17)$$

This equation defines the expected value of the product of deviations from the mean, measuring the linear dependence between time series values at different lags. The partial autocorrelation function (PACF) plot can be used to identify significant input variables based on the estimated $\phi_{k,k}$ from equation (15). Input selection is determined by analyzing PACF values at different lag lengths. If the output variable is y_i and the PACF at lag k falls outside the 95% confidence interval $[-1.96/\sqrt{n}\ , 1.96/\sqrt{n}]$, then y_{i-k} is considered a relevant input variable.

2.6 Hybrid proposed model

The SARIMA-EVMD-SVR model was developed to improve the accuracy of forecasting nonlinear data. SARIMA is utilized to capture trends and seasonal patterns, while VMD-SVR is employed to decompose and predict the nonlinear components that traditional linear models cannot effectively handle (see Figure 3). The model development process consists of the following key steps:

2.6.1 Applying a Linear Model (SARIMA) to Forecast Monthly Rainfall Trends and Seasonality ($\hat{L}_t$) (see details in 2.4).

2.6.2 Calculating Residual Subseries (e_t): The residual subseries (e_t) is calculated using the actual rainfall values (y_t) and the forecasted values from SARIMA ($\hat{L}_t$) based on the following equation :

$$e_t = y_t - \hat{L}_t \quad (18)$$

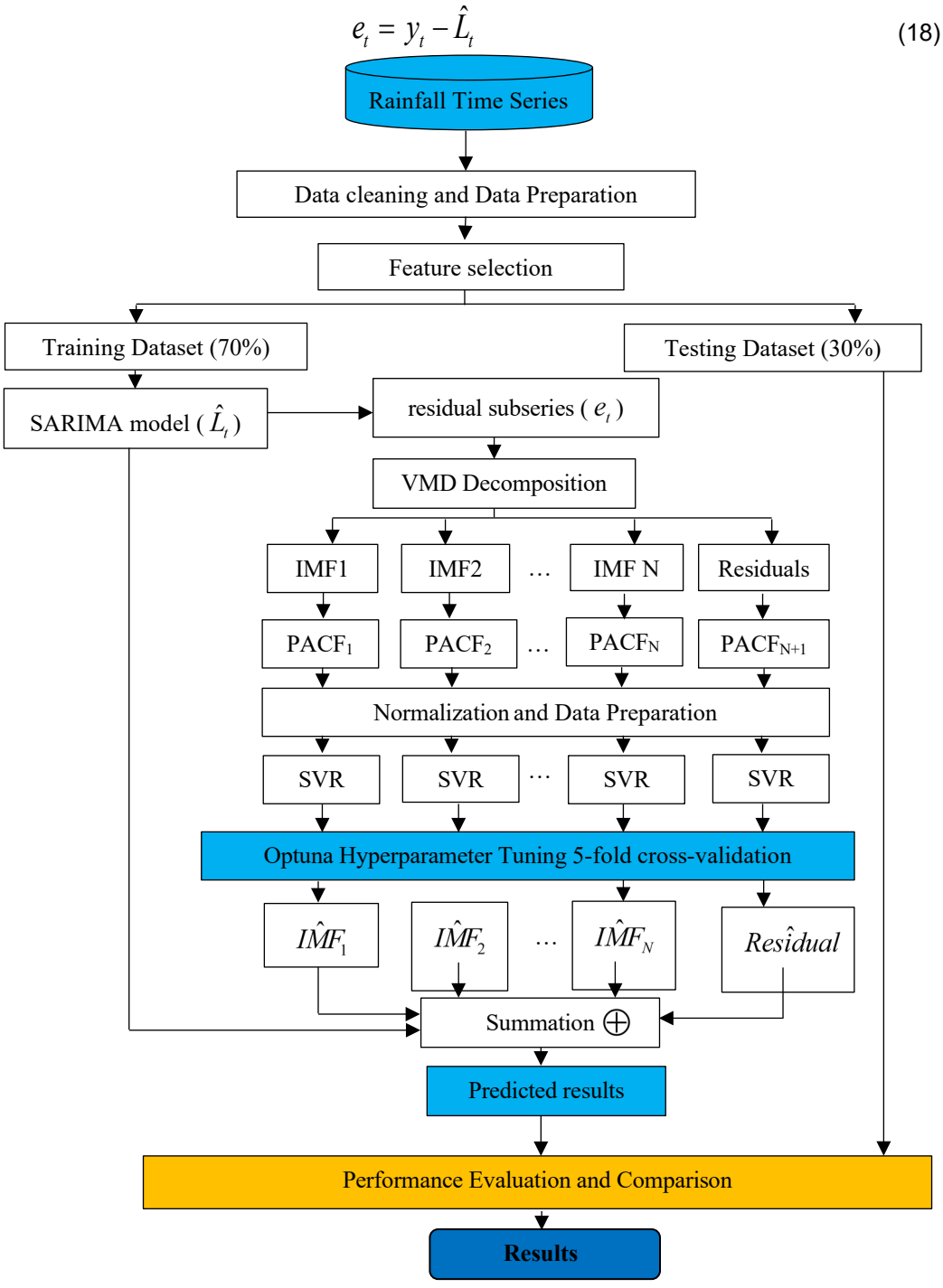


Figure 3. Framework of the SARIMA-ESVMD-SVR Hybrid Model for monthly rainfall forecasting

2.6.3 Decomposing Nonlinear Components using Variational Mode Decomposition (VMD): From step 2.7.2 the residual subseries (e_t) is decomposed into K Intrinsic Mode. Functions (IMFs) using VMD, and the remaining residual component ($Residual_t$) is computed as follows:

$$Residual_t = e_t - \sum_{i=1}^{K} IMF_t \tag{19}$$

2.6.4 Forecasting Nonlinear Components Using Support Vector Regression (SVR): SVR is used to forecast the nonlinear components, including IMFs and the remaining residual ($Residual_t$) obtained from the VMD process. The model parameters are optimized using Optuna Hyperparameter Tuning, and its performance is evaluated through 5-fold cross-validation (see details in 2.5).

2.6.5 Combining Forecasts from Linear and Nonlinear Models: The final forecast is obtained by combining the predicted values from SARIMA ($\hat{L}_t$) and the nonlinear components forecasted using SVR. The combined forecast is given by the following equation:

$$\hat{y}_t = \hat{L}_t + \sum_{i=1}^{K} \hat{IMF}_{i,t} + \hat{Residual}_t \tag{20}$$

2.6.6 Model Accuracy Evaluation: The forecasted values are compared with the actual data and assessed based on the criteria specified in 2.8 to evaluate the model's accuracy and reliability in rainfall forecasting.

2.7 Evaluation of the model performance

The performance of the proposed SARIMA-EVMD-SVR hybrid model is evaluated using multiple statistical metrics, including root mean square error (RMSE), relative root mean square error (RRMSE), residual predictive deviation (RPD), mean absolute error (MAE), accuracy improvement (AI) and coefficient of determination (r^2). These metrics collectively assess the model's forecasting accuracy, reliability, and comparative performance against conventional models (Parviz et al., 2023).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{21}$$

$$RRMSE = \frac{RMSE}{\bar{y}} \times 100 \tag{22}$$

$$RPD = \frac{SD}{RMSE} \tag{23}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\, y_t - \hat{y}_t \,| \tag{24}$$

$$AI = \frac{S - S_h}{S} \times 100 \tag{25}$$

$$r^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y}_t)^2} \tag{26}$$

where y_t represents the observed value at time t, $\hat{y}_t$ is the predicted value, and $\overline{y}_t$ is the mean of observed values. S denotes the MAE of the baseline model, while S_h is the MAE of the improved hybrid model. SD refers to the standard deviation of observed values. The minimum values of $RMSE$, $RRMSE$, and MAE indicate a higher similarity between forecasted and observed values. Accuracy Improvement (AI) is used to assess the effectiveness of the hybrid model, where $AI > 0$ suggests that the hybrid model outperforms the baseline model, whereas $AI \leq 0$ implies no significant improvement (Parviz et al., 2023). RPD measures the predictive reliability of the model, particularly for datasets with varying levels of variability. RPD values greater than 2.0 indicate strong predictive performance, while values between 1.5 and 2.0 suggest moderate accuracy. RPD below 1.5 represents poor model performance. The coefficient of determination (r^2) quantifies how well the model explains the variance in observed values. An r^2 value close to 1.0 indicates a highly reliable forecast.

3. Results and Discussion

3.1 Performance of SARIMA: Baseline model analysis

This section evaluates SARIMA as a baseline model for rainfall forecasting, analyzing stationarity, parameter tuning, prediction accuracy, and residuals. The findings serve as a reference for hybrid modeling.

3.1.1 Stationarity analysis, model selection, and parameter optimization

To evaluate the suitability of SARIMA modeling, the stationarity of the 12 datasets was first assessed using the Augmented Dickey-Fuller (ADF) test. All datasets showed strong evidence of stationarity, as indicated by ADF statistics ranging from -4.498 to -6.439 and p-values less than 0.001. Consequently, all datasets were confirmed to be stationary without the need for any data transformation (e.g., logarithmic), allowing for direct modeling. Subsequently, SARIMA models were constructed for each dataset. Model parameters were optimized based on the minimum values of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The best-performing models varied across datasets. For instance, the MK dataset was best modeled using SARIMA $(2,0,0)(0,1,2)_{12}$ with AIC = –525.489 and BIC = –506.419, while the UBA dataset fitted best with SARIMA

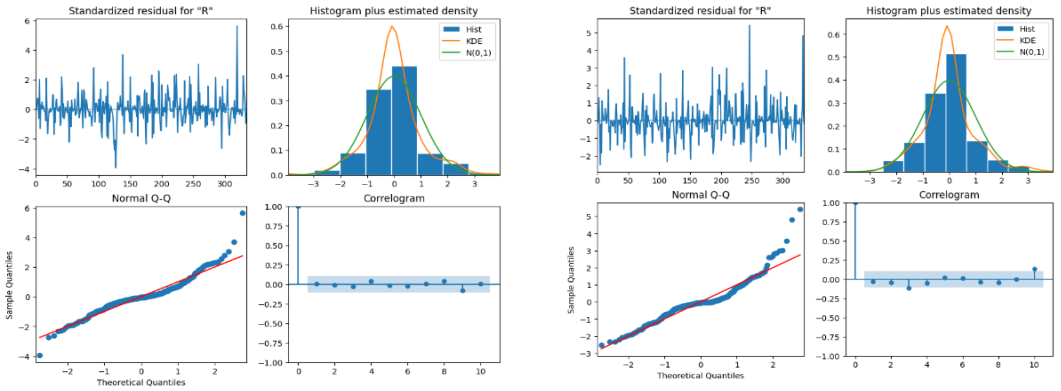
$(0,0,2)(1,0,1)_{12}$, yielding the lowest AIC = –573.515 and BIC = –554.268. All selected models incorporated seasonal components with a period of 12, indicating yearly seasonality. The results shown in Table 2 emphasize how well combining ADF-based stationarity testing and information criteria for model selection works to create strong SARIMA models suited for specific datasets.

3.1.2 Residual analysis and diagnostics for SARIMA

Following the stationarity analysis, model selection, and parameter optimization described in Section 3.1.1, a residual analysis was conducted to validate the SARIMA model using Python's statsmodels package. The diagnostics examined randomness, normality, and independence of residuals to ensure reliable forecasting. Two representative datasets, shown in Figure 4 illustrate standardized residuals for randomness, a Q-Q plot for normality, a histogram with KDE for density estimation, and a correlogram for autocorrelation. Since residual patterns across all 12 datasets exhibited similar characteristics, these examples provide a representative evaluation of SARIMA's overall performance. The results indicate that while SARIMA effectively captures linear trends and seasonality, it still exhibits nonlinear dependencies that may require hybrid modeling to enhance forecasting performance, which will be further explored in Section 3.3.

Table 2. ADF test results, data stationarity assessment, and SARIMA model selection with parameter optimization.

Data Set	ADF Statistics	p-Values	SARIMA Model
MK	-4.498	<0.001	SARIMA $(2,0,0)(0,1,2)_{12}$ AIC = -525.489, BIC = -506.419
RE	-5.155	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -392.244, BIC = -384.616
SK	-4.909	<0.001	SARIMA $(0,0,1)(0,1,1)_{12}$ AIC = -367.594, BIC = -356.151
UBC	-6.117	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -345.282, BIC = -337.654
UBA	-5.546	<0.001	SARIMA $(0,0,2)(1,0,1)_{12}$ AIC = -573.515, BIC = -554.268
SURT	-4.792	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -402.631, BIC = -395.003
SUR	-6.111	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -421.501, BIC = -413.873
SURA	-6.439	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -435.327, BIC = -427.699
NKR	-6.104	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -530.881, BIC = -523.253
NKRP	-5.465	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -356.237, BIC = -348.609
NKRC	-6.090	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -394.901, BIC = -387.273
BURN	-4.784	<0.001	SARIMA $(0,0,0)(0,1,1)_{12}$ AIC = -322.001, BIC = -314.373



1. MK data set 2. RE data set

Figure 4. Residual diagnostics for SARIMA Model based on MK and RE datasets

3.2 Performance evaluation of SVR model

Based on the parameter tuning results of the Support Vector Regression (SVR) model, lag values for input variables were identified using Partial Autocorrelation Function (PACF) analysis, while the core hyperparameters were optimized via Optuna. The results revealed significant variations in the optimal parameters across datasets, reflecting SVR's capability to adapt to the structural characteristics of different time series in each region (as detailed in Table 3). For the parameter C, which governs the trade-off between model complexity and prediction error, the UBA dataset exhibited the highest value (C = 428.344409), indicating that a high degree of flexibility was required for effective fitting. In contrast, the SK dataset yielded the lowest C value (13.625763), suggesting a relatively simple underlying data structure that could be effectively modeled under stricter constraints. Regarding ε, which defines the margin of tolerance where no penalty is given for errors, only minor variation was observed across datasets, with values ranging approximately from 0.0001 to 0.0011. This narrow margin indicates that the SVR models were generally calibrated to achieve high precision in fitting. The γ parameter, which determines the curvature of the kernel function, was consistently small across all datasets (less than 0.001), implying that the data exhibited only mild nonlinearity or that the models employed relatively smooth kernel functions to avoid overfitting. In terms of lag selection from PACF, most datasets involved a wide range of lag inputs, particularly NKRC and MK, both of which selected more than ten lags. This suggests more complex temporal dependencies in these series, aligning with the residual analysis results that indicated potential long-term autocorrelation pattern.

These findings are consistent with those of Xiang et al. (2018), who demonstrated that SVR is highly effective for modeling short-term components with high variability. The authors emphasized that SVR is sensitive to parameters such as C, ε, and γ, and that careful tuning can substantially improve forecasting accuracy. When combined with signal decomposition techniques like Ensemble Empirical Mode Decomposition (EEMD) and lag selection through PACF, SVR demonstrated superior performance over standalone ANN models—particularly for short-period components (e.g., IMF1), which are often volatile and difficult to model using linear or single-network approaches.

Table 3. PACF results and optimized SVR parameters (C, ε, and γ) for each dataset

Data Set	Input Variable	C	ε	γ
MK	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 28, 35, 36]	288.032723	0.000904	0.000154
RE	[0, 1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 24, 36]	93.072743	0.000125	0.000093
SK	[0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 23, 24, 35, 36]	13.625763	0.00054	0.000874
UBC	[0, 1, 3, 4, 5, 6, 7, 8, 11, 12, 17, 18, 19, 21, 23, 24, 29, 35, 36]	59.215929	0.000302	0.000997
SURT	[0, 1, 3, 4, 5, 6, 7, 8, 11, 12, 13, 24, 25, 36]	39.06143	0.000456	0.000155
SUR	[0, 1, 3, 4, 5, 6, 7, 8, 11, 12, 13, 17, 18, 19, 24, 29, 36]	742.49453	0.000119	0.000024
SURA	[0, 1, 3, 4, 5, 6, 7, 8, 11, 12, 13, 17, 18, 19, 24, 25, 35, 36]	14.635374	0.000195	0.000926
NKR	[0, 1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 18, 19, 21, 24, 26, 28, 36]	323.076783	0.000727	0.000388
NKRP	[0, 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 18, 20, 21, 22, 24, 25, 26, 27, 30, 33, 35, 36]	236.096679	0.00017	0.000111
NKRC	[0, 1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 22, 23, 24, 26, 30, 36]	326.194488	0.000896	0.000065
BURN	[0, 1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 17, 23, 24, 27, 33, 36]	49.945764	0.000996	0.000995

In summary, the SVR model demonstrated an ability to tailor its parameters to the unique characteristics of each dataset and proved more capable of capturing nonlinear data structures than SARIMA, which is generally limited in modeling complex temporal and nonlinear patterns. However, a comprehensive comparative performance evaluation of SARIMA, SVR, and hybrid models is presented in the following sections.

3.3 Performance evaluation of the hybrid model

The SARIMA–EVMD–SVR hybrid model was designed to integrate the strengths of both linear and nonlinear models, aiming to enhance the accuracy and reliability of monthly rainfall forecasting across 12 monitoring stations in northeastern Thailand. This hybrid framework, as detailed in Section 2.7, begins with the application of SARIMA to capture linear trend and seasonality and to extract residuals that contain nonlinear patterns. These residuals are then processed using Ensemble Variational Mode Decomposition (EVMD), a robust signal decomposition technique that isolates nonlinear features into a set of Intrinsic Mode Functions (IMFs) and a final residual component (Chen et al., 2021). Subsequently, each IMF and the final residual are forecasted using Support Vector Regression (SVR), with lag values determined by the Partial Autocorrelation Function (PACF). To ensure optimal performance, SVR parameters are fine-tuned using Optuna, a metaheuristic optimization framework. The final forecast is obtained by aggregating the SARIMA prediction with the SVR-predicted components, thereby reconstructing the signal with contributions from both linear and nonlinear dynamics. Figures 5 to 7 present a representative case using the Mukdahan (MK) dataset to illustrate the decomposition, forecasting, and reconstruction stages of the proposed hybrid model.

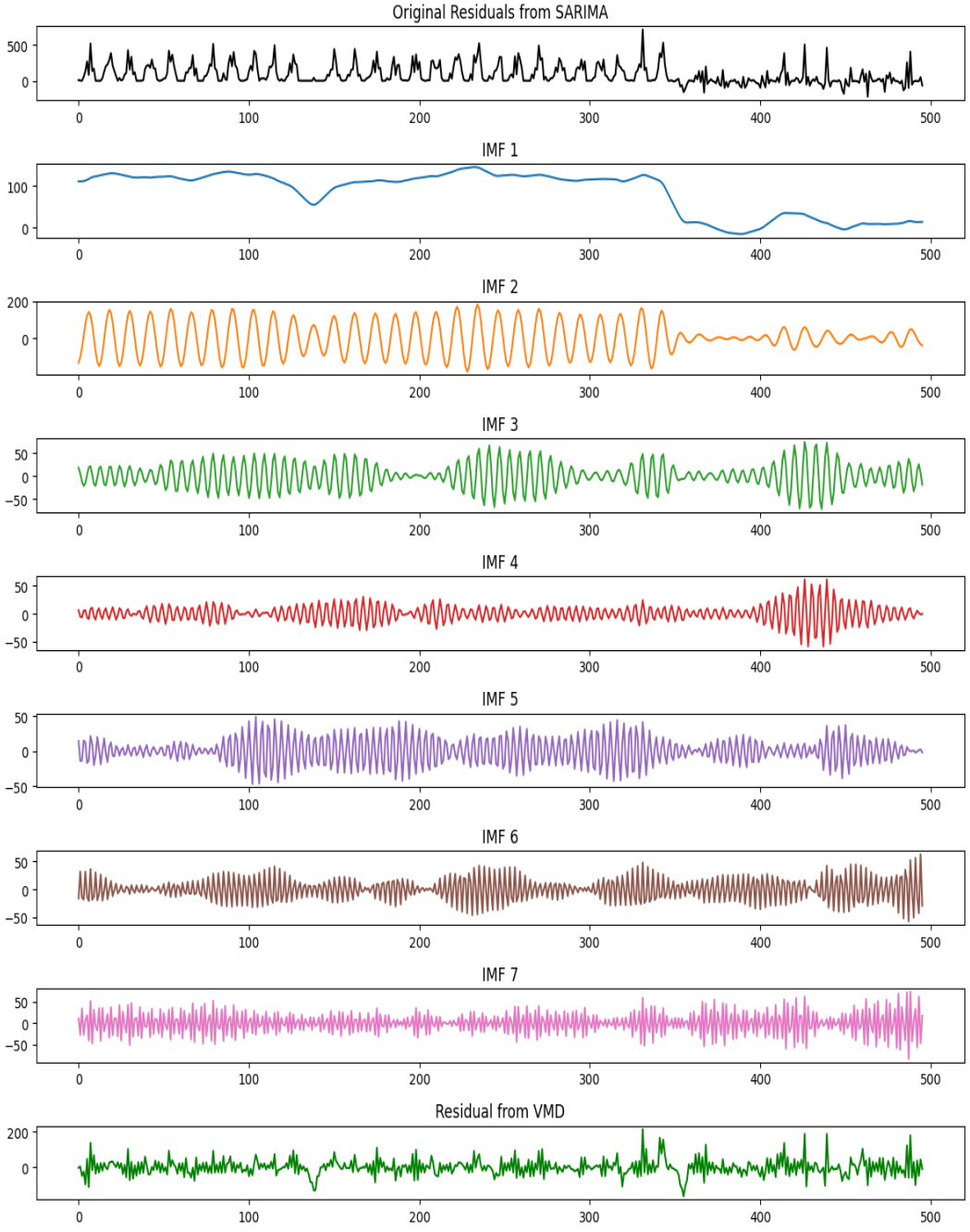


Figure 5. Residual decomposition of the SARIMA model for the Mukdahan (MK) dataset using VMD

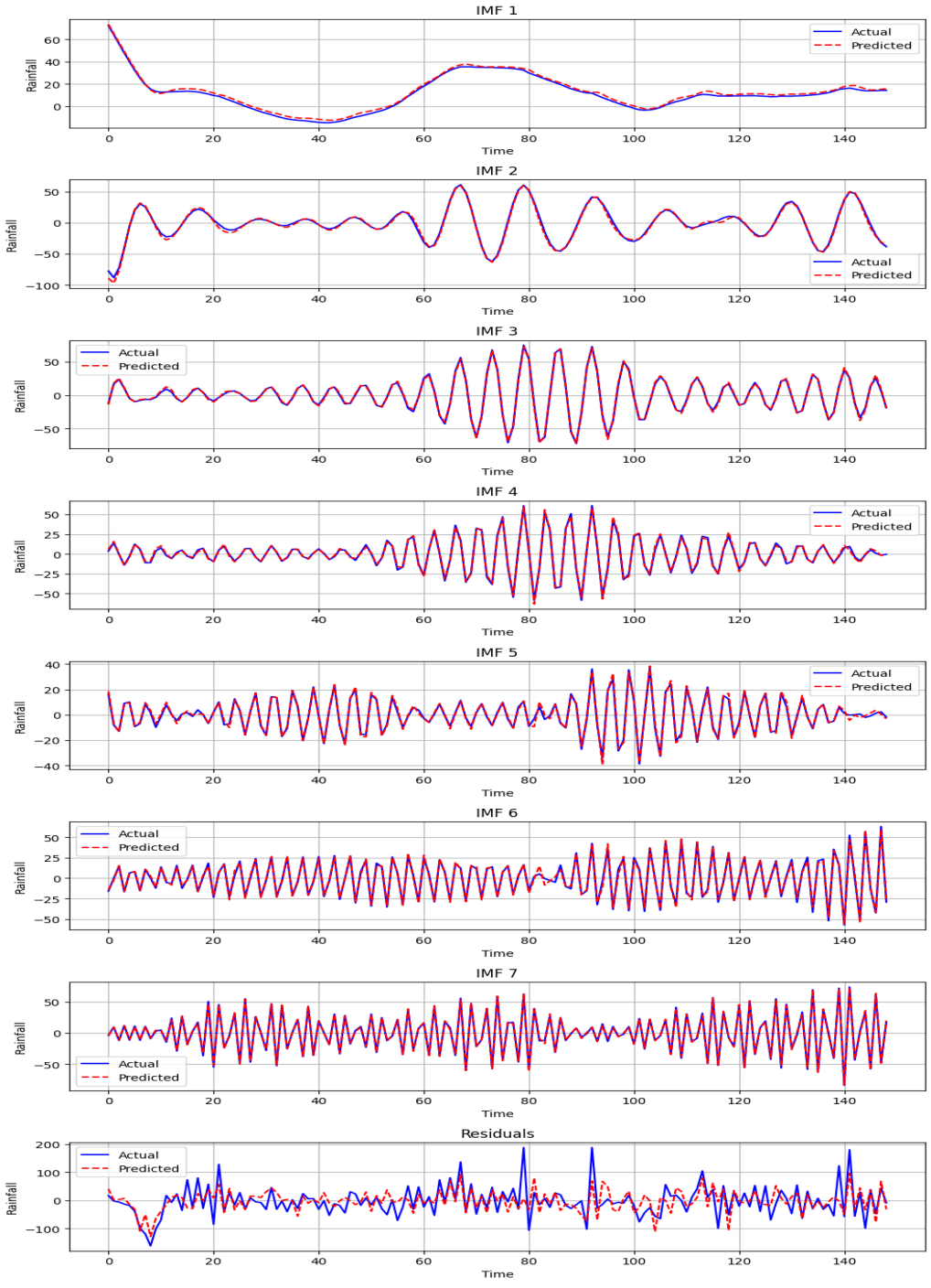


Figure 6. Forecasting results of IMFs and residuals using SVR for the Mukdahan (MK) dataset

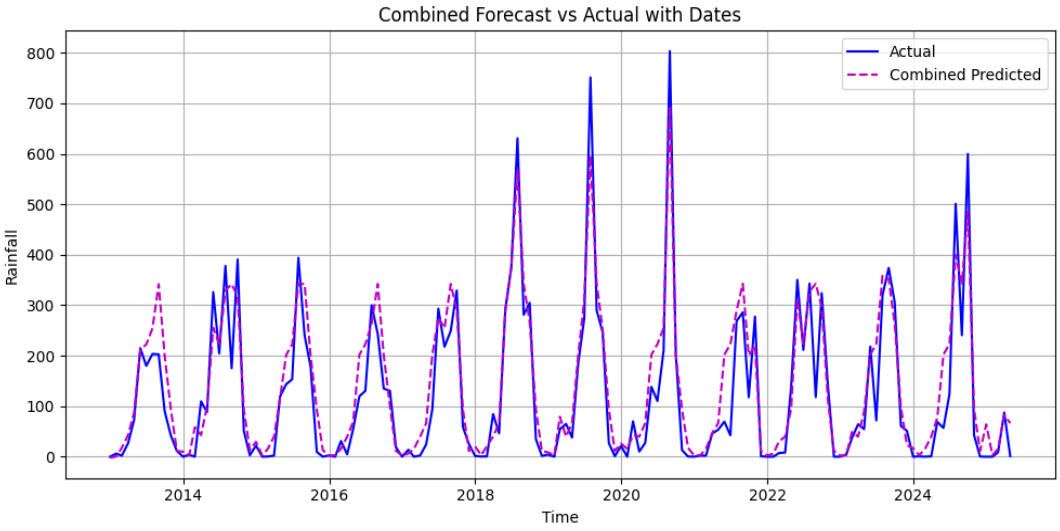


Figure 7. Final combined forecast (SARIMA, IMFs, and Residuals) vs. actual rainfall for the Mukdahan (MK) dataset

The same modeling procedure was applied to the remaining 11 datasets to ensure consistency and robustness. This multi-layered process aligns with recent advancements in hybrid modeling, which combine signal decomposition, machine learning, and parameter tuning techniques to improve predictive performance (Mehr et al., 2024; Parsaie et al., 2024). Model performance was assessed using multiple accuracy metrics, including RMSE, PRMSE, MAE, RPD, and R^2, along with the Accuracy Improvement (AI) percentage. As presented in Table 4, the hybrid model consistently outperformed the baseline SARIMA and SVR models across all stations. For instance, at Mukdahan station (MK), RMSE was significantly reduced from 101.754 (SARIMA) and 82.610 (SVR) to 59.341 with the hybrid model. Similarly, MAE dropped to 41.376, reflecting a consistent error reduction pattern across all stations.

These findings align with previous studies, such as that by Wang et al. (2024b) who proposed the SABO–VMD–SVR model and found that combining VMD-based IMFs with a hybrid model significantly improved forecasting accuracy over SARIMA. Similarly, Parviz and Ghorbanpour (2024) demonstrated that integrating SVR and ARIMA using PSO in their IARIMA–C–PSO model reduced RMSE by up to 35% and achieved RPD values greater than 2. Dotse et al. (2024) also confirmed that hybrid models incorporating signal decomposition and parameter tuning significantly enhanced prediction accuracy in complex datasets. In terms of reliability, the hybrid model achieved higher RPD and R^2 values than the base models. At Sisaket station (SK), the hybrid model yielded RPD = 3.215 and R^2 = 0.903, outperforming SARIMA and SVR with statistically significant margins. This result is in line with Mehr et al. (2024) who reported NSE = 0.98 for their VMD–SANN model enhanced with a stabilizer component for ANN. Moreover, the hybrid model demonstrated substantial accuracy improvement compared to the baselines, with AI gains ranging from 12.83% to 36.11% over SARIMA and 18.04% to 39.26% over SVR. Stations such as SK, UBC, and MK showed the highest improvements, exceeding 30%, echoing the findings of Chen et al. (2021), who used VMD to decompose error series before combining with ARIMA and AI models for forecasting under high uncertainty.

Table 4. Comparison of forecasting accuracy and AI improvement with SARIMA and SVR as baselines

Data Set	Model	RMSE	RR MSE	RPD	MAE	R^2	AI SARIMA (%)	AI SVR (%)
MK	SARIMA	101.754	83.201	1.505	60.389	0.558		
	SVR	82.610	64.282	1.693	50.486	0.651	31.48	18.04
	Hybrid	**59.341**	**48.521**	**2.580**	**41.376**	**0.849**		
RE	SARIMA	76.413	66.555	1.650	49.576	0.633		
	SVR	87.874	72.832	1.451	54.953	0.525	31.98	38.63
	Hybrid	**49.982**	**43.534**	**2.523**	**33.722**	**0.843**		
SK	SARIMA	95.317	65.083	1.753	58.749	0.675		
	SVR	82.827	59.510	1.747	56.510	0.672	36.11	33.58
	Hybrid	**51.961**	**35.479**	**3.215**	**37.536**	**0.903**		
UBC	SARIMA	97.396	71.275	1.664	58.247	0.639		
	SVR	80.349	57.333	1.821	53.051	0.698	32.90	26.33
	Hybrid	**57.705**	**42.229**	**2.808**	**39.082**	**0.873**		
UBA	SARIMA	76.267	65.345	1.756	50.558	0.675		
	SVR	102.925	78.537	1.447	58.354	0.523	28.11	37.71
	Hybrid	**51.158**	**43.831**	**2.616**	**36.348**	**0.854**		
SURT	SARIMA	58.209	61.294	1.731	41.486	0.666		
	SVR	74.358	61.582	1.576	50.149	0.597	12.83	27.89
	Hybrid	**52.049**	**54.806**	**1.936**	**36.164**	**0.733**		
SUR	SARIMA	68.314	59.441	1.777	43.054	0.683		
	SVR	84.083	63.511	1.533	51.341	0.574	24.76	36.90
	Hybrid	**45.641**	**39.713**	**2.659**	**32.394**	**0.859**		
SURA	SARIMA	63.527	57.423	1.790	41.727	0.688		
	SVR	86.192	66.188	1.533	52.260	0.574	23.92	39.26
	Hybrid	**47.519**	**42.952**	**2.393**	**31.744**	**0.825**		
NKR	SARIMA	65.136	67.273	1.562	45.731	0.590		
	SVR	68.057	69.938	1.400	42.624	0.490	34.25	29.45
	Hybrid	**39.482**	**40.778**	**2.577**	**30.070**	**0.849**		
NKRP	SARIMA	59.538	65.474	1.401	45.375	0.491		
	SVR	64.352	61.769	1.438	44.882	0.517	26.14	25.33
	Hybrid	**46.842**	**51.512**	**1.780**	**33.514**	**0.685**		
NKRC	SARIMA	53.814	62.708	1.577	38.194	0.598		
	SVR	58.314	61.869	1.515	39.618	0.564	29.72	32.24
	Hybrid	**38.505**	**44.869**	**2.204**	**26.844**	**0.794**		
BURN	SARIMA	65.989	62.419	1.604	45.387	0.611		
	SVR	63.355	58.427	1.558	42.047	0.588	33.78	28.52
	Hybrid	**41.275**	**39.042**	**2.565**	**30.055**	**0.848**		

4. Conclusions

This study proposed a hybrid forecasting model integrating SARIMA, Ensemble Variational Mode Decomposition (EVMD), and Support Vector Regression (SVR) to enhance the accuracy of monthly rainfall prediction in northeastern Thailand. The evaluation across 12 meteorological stations confirmed that the hybrid model consistently outperformed baseline models in terms of forecasting accuracy and reliability. By combining linear and nonlinear modeling approaches with signal decomposition, the SARIMA–EVMD–SVR model effectively reduced forecasting errors and demonstrated high predictive capability for complex rainfall time series. These findings highlight the model's potential applicability to other climate-related variables or regions with similar characteristics. Future research may explore adapting this framework to real-time forecasting systems or extending its use to support water resource planning and climate adaptation strategies.

5. Acknowledgements

The authors would like to express their sincere gratitude to the Faculty of Science, Ubon Ratchathani Rajabhat University, for providing financial support for this research. Sincere appreciation is also extended to the Northeastern Regional Office of the Thai Meteorological Department (TMD) for kindly providing the monthly rainfall data collected from 12 meteorological stations, which were used in this study. The authors also wish to thank the Mathematics Program, Faculty of Science, Ubon Ratchathani Rajabhat University, for the continued opportunity and academic support throughout the research process.

6. Authors' Contributions

Thanakon Sutthison: designed research, data curation, performed research, analyzed data, wrote the paper.; Somporn Thepchim: conceptualization, data curation; Yaovaruk Thongphum: coordinated research.

7. Conflicts of Interest

The authors declare no conflict of interest.

8. AI Declaration

The preparation of this manuscript by ChatGPT (OpenAI) and QuillBot in order to enhance conceptual clarity, verify formulas, support initial drafting, and improve grammar and check language errors has been reviewed and edited as needed by Thanakon Sutthison, Somporn Thepchim, and Yaovaruk Thongphum. The authors take full responsibility for the content and accuracy of the publication.

References

Abebe, W. T., & Endalie, D. (2023). Artificial intelligence models for prediction of monthly rainfall without climatic data for meteorological stations in Ethiopia. *Journal of Big Data*, *10*(1), Article 2. https://doi.org/10.1186/s40537-022-00683-3

Akhtar, M., Shatat, A. S. A., Ahamad, S. A. H., Dilshad, S., & Samdani, F. (2023). Optimized cascaded CNN for intelligent rainfall prediction model: A research towards statistic-based machine learning. *Theoretical Issues in Ergonomics Science*, *24*(5), 564-592. https://doi.org/10.1080/1463922X.2022.2135786

Ali, M., Prasad, R., Xiang, Y., & Yaseen, Z. M. (2020). Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of Hydrology*, *584*, Article 124647. https://doi.org/10.1016/j.jhydrol.2020.124647

Alqahtani, F., Abotaleb, M., Subhi, A. A., El-Kenawy, E.-S. M., Abdelhamid, A. A., Alakkari, K., Badr, A., Al-Mahdawi, H. K., Ibrahim, A., & Kadi, A. (2023). A hybrid deep learning model for rainfall in the wetlands of southern Iraq. *Modeling Earth Systems and Environment*, *9*(4), 4295-4312. https://doi.org/10.1007/s40808-023-01754-x

Ayiah-Mensah, F., Bosson-Amedenu, S., Baah, E. M., & Addor, J. A. (2025). Advancements in seasonal rainfall forecasting: A seasonal auto-regressive integrated moving average model with outlier adjustments for Ghana's Western Region. *Scientific African*, *28*, Article e02632. https://doi.org/10.1016/j.sciaf.2025.e02632

Chen, W., Xu, H., Chen, Z., & Jiang, M. (2021). A novel method for time series prediction based on error decomposition and nonlinear combination of forecasters. *Neurocomputing*, *426*, 85-103. https://doi.org/10.1016/j.neucom.2020.10.048

Dotse, S.-Q., Larbi, I., Limantol, A. M., & De Silva, L. C. (2024). A review of the application of hybrid machine learning models to improve rainfall prediction. *Modeling Earth Systems and Environment*, *10*(1), 19-44. https://doi.org/10.1007/s40808-023-01835-x

Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, *62*(3), 531-544. https://doi.org/10.1109/TSP.2013.2288675

Guo, S., Sun, S., Zhang, X., Chen, H., & Li, H. (2023). Monthly precipitation prediction based on the EMD–VMD–LSTM coupled model. *Water Supply*, *23*(11), 4742-4758. https://doi.org/10.2166/ws.2023.275

He, R., Zhang, L., & Chew, A. W. Z. (2022). Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning. *Knowledge-Based Systems*, *251*, Article 109125. https://doi.org/10.1016/j.knosys.2022.109125

He, R., Zhang, L., & Chew, A. W. Z. (2024). Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning. *Expert Systems with Applications*, *235*, Article 121160. https://doi.org/10.1016/j.eswa.2023.121160

He, Z., & Huang, J. (2023). A novel non-ferrous metal price hybrid forecasting model based on data preprocessing and error correction. *Resources Policy*, *86*, Article 104189. https://doi.org/10.1016/j.resourpol.2023.104189

Hou, S., Geng, Q., Huang, Y., & Bian, Z. (2024). Rainfall prediction model based on CEEMDAN-VMD-BiLSTM network. *Water, Air, and Soil Pollution*, *235*(8), Article 482. https://doi.org/10.1007/s11270-024-07299-8

Jamei, M., Ali, M., Malik, A., Karbasi, M., Rai, P., & Yaseen, Z. M. (2023). Development of a TVF-EMD-based multi-decomposition technique integrated with Encoder-Decoder-Bidirectional-LSTM for monthly rainfall forecasting. *Journal of Hydrology*, *617*, Article 129105. https://doi.org/10.1016/j.jhydrol.2023.129105

Jiang, X. (2023). A combined monthly precipitation prediction method based on CEEMD and improved LSTM. *PLOS ONE*, *18*(7), Article e0288211. https://doi.org/10.1371/journal.pone.0288211

Johny, K., Pai, M. L., & S., A. (2022). A multivariate EMD-LSTM model aided with Time dependent intrinsic cross-correlation for monthly rainfall prediction. *Applied Soft Computing*, *123*, Article 108941. https://doi.org/10.1016/j.asoc.2022.108941

Mehr, A. D., Shadkani, S., Abualigah, L., Safari, M. J. S., & Migdady, H. (2024). A novel stabilized artificial neural network model enhanced by variational mode decomposing. *Heliyon*, *10*(13), Article e34142. https://doi.org/10.1016/j.heliyon.2024.e34142

Parsaie, A., Ghasemlounia, R., Gharehbaghi, A., Haghiabi, A., Chadee, A. A., & Nou, M. R. G. (2024). Novel hybrid intelligence predictive model based on successive variational mode decomposition algorithm for monthly runoff series. *Journal of Hydrology*, *634*, Article 131041. https://doi.org/10.1016/j.jhydrol.2024.131041

Parviz, L., & Ghorbanpour, M. (2024). A hybrid EMD and MODWT models for monthly precipitation forecasting using an innovative error decomposition method. *Stochastic Environmental Research and Risk Assessment*, *38*(10), 4107-4130. https://doi.org/10.1007/s00477-024-02797-x

Parviz, L., Rasouli, K., & Torabi Haghighi, A. (2023). Improving hybrid models for precipitation forecasting by combining nonlinear machine learning methods. *Water Resources Management*, *37*(10), 3833-3855. https://doi.org/10.1007/s11269-023-03528-7

Pinheiro, E., & Ouarda, T. B. M. J. (2023). Short-lead seasonal precipitation forecast in northeastern Brazil using an ensemble of artificial neural networks. *Scientific Reports*, *13*(1), Article 20429. https://doi.org/10.1038/s41598-023-47841-y

Pirone, D., Cimorelli, L., Del Giudice, G., & Pianese, D. (2023). Short-term rainfall forecasting using cumulative precipitation fields from station data: A probabilistic machine learning approach. *Journal of Hydrology*, *617*, Article 128949. https://doi.org/10.1016/j.jhydrol.2022.128949

Rezaiy, R., & Shabri, A. (2024). Improving drought prediction accuracy: A hybrid EEMD and support vector machine approach with standardized precipitation index. *Water Resources Management*, *38*(13), 5255-5277. https://doi.org/10.1007/s11269-024-03912-x

Shao, P., Feng, J., Zhang, P., & Lu, J. (2024). Interpretable spatial-temporal attention convolutional network for rainfall forecasting. *Computers and Geosciences*, *185*, Article 105535. https://doi.org/10.1016/j.cageo.2024.105535

Skarlatos, K., Bekri, E. S., Georgakellos, D., Economou, P., & Bersimis, S. (2023). Projecting Annual Rainfall Timeseries Using Machine Learning Techniques. *Energies*, *16*(3), Article 1459. https://doi.org/10.3390/en16031459

Someetheram, V., Marsani, M. F., Kasihmuddin, M. S. M., Jamaludin, S. Z. M., Mansor, Mohd. A., & Zamri, N. E. (2025). Hybrid double ensemble empirical mode decomposition and K-Nearest Neighbors model with improved particle swarm optimization for water level forecasting. *Alexandria Engineering Journal*, *115*, 423-433. https://doi.org/10.1016/j.aej.2024.12.035

Sutthison, T. (2024). Forecasting Thai durian exports using a hybrid time series SARIMA-SVR approach. *Suranaree Journal of Science and Technology*, *31*(5), 030229(1-13).

Wang, H., Chen, S., & Zhai, W. (2024a). Variational generalized nonlinear mode decomposition: Algorithm and applications. *Mechanical Systems and Signal Processing*, *206*, Article 110913. https://doi.org/10.1016/j.ymssp.2023.110913

Wang, S., Cao, B., Bai, R., & Liu, G. (2024b). Optimization of waterproofing and drainage measures for open-pit mines based on seasonal rainfall time series prediction. *Environmental Modelling and Software*, *173*, Article 105957. https://doi.org/10.1016/j.envsoft.2024.105957

Wang, Y., Yuan, Z., Liu, H., Xing, Z., Ji, Y., Li, H., Fu, Q., & Mo, C. (2022). A new scheme for probabilistic forecasting with an ensemble model based on CEEMDAN and AM-MCMC and its application in precipitation forecasting. *Expert Systems with Applications*, *187*, Article 115872. https://doi.org/10.1016/j.eswa.2021.115872

Waqas, M., Humphries, U. W., Hlaing, P. T., Wangwongchai, A., & Dechpichai, P. (2024). Advancements in daily precipitation forecasting: A deep dive into daily precipitation

forecasting hybrid methods in the tropical climate of Thailand. *MethodsX*, *12*, Article 102757. https://doi.org/10.1016/j.mex.2024.102757

Wu, H., Du, P., & Heng, J. (2024a). Gated convolution with attention mechanism under variational mode decomposition for daily rainfall forecasting. *Measurement, 237*, Article 115222. https://doi.org/10.1016/j.measurement.2024.115222

Wu, S., Dong, Z., Guzmán, S. M., Conde, G., Wang, W., Zhu, S., Shao, Y., & Meng, J. (2024b). Two-step hybrid model for monthly runoff prediction utilizing integrated machine learning algorithms and dual signal decompositions. *Ecological Informatics*, *84*, Article 102914. https://doi.org/10.1016/j.ecoinf.2024.102914

Xiang, Y., Gou, L., He, L., Xia, S., & Wang, W. (2018). A SVR–ANN combined model based on ensemble EMD for rainfall prediction. *Applied Soft Computing*, *73*, 874-883. https://doi.org/10.1016/j.asoc.2018.09.018

Xu, B., Zhou, F., Li, H., Yan, B., & Liu, Y. (2019). Early fault feature extraction of bearings based on Teager energy operator and optimal VMD. *ISA Transactions*, *86*, 249-265. https://doi.org/10.1016/j.isatra.2018.11.010

Yin, Y., He, J., Guo, J., Song, W., Zheng, H., & Dan, J. (2024). Enhancing precipitation estimation accuracy: An evaluation of traditional and machine learning approaches in rainfall predictions. *Journal of Atmospheric and Solar-Terrestrial Physics*, *255*, Article 106175. https://doi.org/10.1016/j.jastp.2024.106175

Zhang, X., & Wu, X. (2023). Combined forecasting model of precipitation based on the CEEMD-ELM-FFOA coupling model. *Water*, *15*(8), Article 1485. https://doi.org/10.3390/w15081485

Zheng, Z., Zhang, X., Yin, Q., Liu, F., Ren, H., & Zhao, R. (2024). A novel optimization rainfall coupling model based on stepwise decomposition technique. *Scientific Reports*, *14*(1), Article 15617. https://doi.org/10.1038/s41598-024-66663-0

Zhou, Z., Ren, J., He, X., & Liu, S. (2021). A comparative study of extensive machine learning models for predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors. *Hydrological Processes*, *35*(11), Article e14424. https://doi.org/10.1002/hyp.14424

Zuo, G., Luo, J., Wang, N., Lian, Y., & He, X. (2020). Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. *Journal of Hydrology*, *585*, Article 124776. https://doi.org/10.1016/j.jhydrol.2020.124776