Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

Analysis of Rainfall Prediction Using Parallel Hybrid Algorithm

D Karthika, K Karthikeyan*

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore, Tamil Nadu, India
E-mail: k.karthikeyan@vit.ac.in

Received: 17 May 2024; **Revised:** 19 July 2024; **Accepted:** 22 August 2024

Abstract: Precisely forecasting rainfall precipitation is an intricate and crucial challenge faced by numerous weather forecasters. In this study, we conducted an examination of different statistical models to assess their efficacy in predicting monthly rainfall precipitation. The objective of this study was to develop a combined model that could enhance the accuracy of such forecasts. To achieve this, we gathered monthly rainfall time series data spanning from January 1901 to December 2017 in Tamil Nadu, India. To enhance the accuracy of rainfall precipitation prediction, we employed a parallel hybrid strategy, combining univariate forecast models. Our proposed forecasting model was compared with other established models, including Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Holt-Winters Additive (HWA) model, Holt model, Exponential Smoothing (ETS) model, and Feed Forward Neural Network (FFNN) model. The results indicate that our proposed model outperformed the other models, demonstrating its superior forecasting capabilities. The proposed model yielded an RMSE value of 0.6403, MSE value of 0.4101, MAE value of 0.3998, NSE value of 0.5924, sMAPE value of 0.7172, and an R-value of 0.7761. A paired t-test was conducted to compare the performance metrics of the proposed model with those of the baseline models. The result shows that this model is statistically significant. Since, It p-value less than 0.05. These findings lead us to the conclusion that the proposed model is well-suited for analyzing and forecasting climatological factors and climatic extremes.

Keywords: time series analysis, SARIMA, ANN, combined forecast, rainfall prediction

MSC: 62P10

Abbreviation

ACF	Autocorrelation Function
AIC	Akaike Information Criterion
ARIMA	Autoregressive Integrated Moving Average
ETS	Exponential Smoothing
FFNN	Feed Forward Neural Network
MSE	Mean Squared Error
MLP	Multi-Layer Perceptron
PACF	Partial Autocorrelation Function
RMSE	Root Mean Square Error

DOI: https://doi.org/10.37256/cm.5320244981

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
R	Correlation Coefficient
NSE	Nash-Sutcliffe Efficiency
sMAPE	Symmetric Mean Absolute Percentage Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
HWA	Holt Winter's Additive model
TN	Tamil Nadu
HW	Holt Winters
ANN	Artificial Neural Network
OGD	Open Government Data
SD	Standard Deviation
CV	Coefficient of Variance
ADF	Augmented Dickey-Fuller Test
MA	Moving Average
SMA	Seasonal Moving Average
SAR	Seasonal Autoregressive
WN	White Noise
NNAR	Neural Network Auto Regressive
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
IoT	Internet of Things

1. Introduction

In this study, we delve into the critical realm of rainfall prediction, a pivotal area of research in environmental science and hydrology. Rainfall plays a fundamental role in shaping ecosystems, influencing agricultural productivity, and determining water resource management strategies [1]. Understanding and accurately predicting rainfall patterns are therefore essential for effective planning and decision-making in various sectors.

Our research aims to contribute to this field by exploring and evaluating robust algorithms for rainfall prediction. By harnessing advanced computational techniques and leveraging comprehensive datasets, we seek to enhance the accuracy and reliability of rainfall forecasts. This endeavor is motivated by the imperative to address the challenges posed by climate variability and change, which underscore the urgency of improving predictive capabilities in hydrological modeling [2, 3].

The impact of monsoon seasons, such as the summer monsoon affecting power generation and industrial sectors, as well as the winter monsoon influencing agricultural production, further underscores the importance of accurate rainfall predictions [4, 5]. Insufficient or excessive rainfall can significantly impact crop yields and people's daily lives.

To address this, a time-series approach is employed to statistically and graphically analyze the data, select suitable forecasting models, predict future values, and control specific methodologies [6]. Many researchers have utilized statistical techniques to address hydrological issues [7, 8], highlighting the significance of statistical modeling in testing, predicting, and decision-making based on hydrological data [9]. Univariate data, consisting of variable observations recorded at distinct time intervals, has been utilized to develop various modeling and prediction approaches for hydrological data, including rainfall precipitation [4, 5, 10].

Numerous studies have confirmed the appropriateness of statistical models for hydrologic time series modeling, particularly in areas with meteorological time series data [2, 3, 11]. When selecting forecasting models for rainfall prediction, factors such as efficiency, effort, cost, and ease of use of the model's outputs are considered [3].

The Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model, introduced by Box et al. [6] is a widely used univariate time series model for forecasting. It is known for its precision, flexibility, and ability to capture

intricate time series patterns. The SARIMA model, denoted as ARIMA (p, d, q)(P, D, Q)12, incorporates autoregressive and moving average components to identify patterns in quasi-time series [12–14]. In hydrological cycle forecasting using univariate time series data, particularly for rainfall, the SARIMA model has proven effective in various regions worldwide [3, 15]. Studies, such as Valipour [16, 17], have demonstrated the superior forecasting performance of SARIMA models compared to ARIMA models for long-term runoff forecasts.

Another model that excels in climatological prediction is the Holt-Winters (HW) model. Holt [18] modified exponentially weighted moving averages to incorporate trend and seasonal variation in the data. The Holt-Winters model is commonly employed for forecasting seasonal data, and various variants of this model, including the Holt-Winters Additive (HWA) model, have been utilized to analyze and forecast seasonal rainfall patterns [4].

Artificial Neural Networks (ANNs) are computational models modeled after the human brain, created to identify patterns and generate predictions. For rainfall prediction, ANNs can effectively model complex relationships between different meteorological variables and rainfall amounts [19]. They are especially advantageous for their capability to manage non-linear data and detect intricate patterns. Consequently, researchers have frequently employed ANN models for rainfall forecasting [20–22]. In this study, the Feed Forward Neural Network (FFNN) model is utilized for precipitation prediction.

Additionally, the exponential smoothing model, a univariate time series model, is employed to analyze the level, trend, and seasonal components of the data. The Holt model [18] is a modified version of weighted moving averages that incorporates trend and seasonal variations.

Before conducting statistical analyses, it is crucial to validate the accuracy and quality of the time series data, making any necessary corrections. This ensures robustness in subsequent modeling phases.

To further enhance forecasting accuracy and reduce vulnerability to weather changes and seasonal patterns, forecasts from multiple accurate forecasting systems can be integrated. The concept of combined forecasts, introduced by Bates and Granger [23], involves assigning weights to individual forecasting methods. Combined forecasts have been shown to reduce errors by averaging distinct forecasts [24–26], making them valuable when uncertainty surrounds the choice of the best forecasting technique. Various researchers, including Winkler and Makridakis [27], have contributed significantly to combined forecasts of univariate time series models [28, 29]. Such an approach allows for increased accuracy with minimal effort and time.

Despite recent advancements in rainfall forecasting, current models often exhibit insufficient accuracy and reliability, particularly in regions with complex terrain and variable weather patterns. This study aims to develop a robust and scalable time-series-based rainfall prediction model. By leveraging historical weather data, geographic features, and atmospheric variables, the model seeks to enhance forecasting precision and provide timely predictions. Integrating advanced algorithms and utilizing comprehensive datasets, this research addresses existing shortcomings to improve the reliability of rainfall predictions. Ultimately, these enhancements aim to support informed decision-making and mitigate risks associated with precipitation variability.

In this study, we propose a novel combined model, employing a parallel hybrid approach that combines the SARIMA and HWA models for rainfall precipitation prediction in Tamil Nadu, India. By compensating for the shortcomings of individual models, this framework enhances the forecasting model's capabilities and optimizes the weights assigned to each model to maximize forecast accuracy.

2. Materials and methodology

2.1 *Materials*

Tamil Nadu (TN) is a state situated in the southern region of India, spanning between 8°33'0" N-12°36'0" N latitude and 73°20'0" E-80°0'0" E longitude. Figure 1 shows the study area and location of Tamil Nadu, India. Climatically, TN falls within the semi-arid to dry semi-humid zone. The state experiences two distinct monsoon seasons: the Southwest monsoon from June to September and the Northeast monsoon from October to December. The average annual rainfall in TN is approximately 945 mm, with the Northeast monsoon contributing the most precipitation. Fluctuations in rainfall

patterns significantly impact TN's water budget, making accurate rainfall prediction essential for mitigating the risks of drought and flooding.

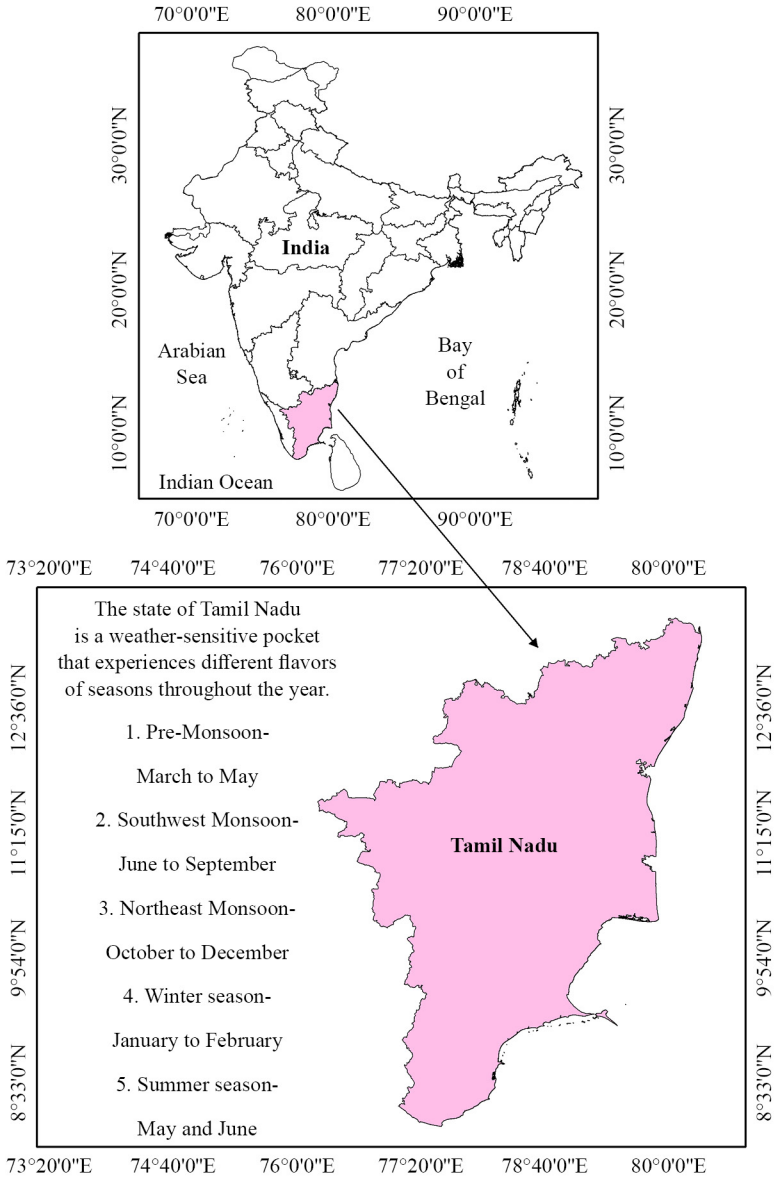

Figure 1. Study area and location of Tamil Nadu, India

Monthly rainfall precipitation data for TN, India were obtained from the Open Government Data Platform India (OGD) at https://data.gov.in/rainfall-india. The time series data spanning from January 1901 to December 2017 were utilized for forecasting purposes. R software was employed for data analysis. The rainfall time series data were divided into training and testing datasets. The training dataset consisted of 70% of the total data, encompassing the period from January 1901 to December 1982, while the testing dataset comprised the remaining 30% from January 1983 to December 2017. Table 1 provides an overview of basic statistical measures, including minimum, maximum, mean, standard deviation (SD), and coefficient of variance (CV), for the observed dataset, training dataset, and testing dataset, respectively.

Table 1. Statistical Analysis of Rainfall Data

	Minimum	Maximum	Mean	SD	CV
Rainfall data	0.1	436.1	78.451	70.074	89.321
Training data	0.1	436.1	79.344	69.96	88.172
Testing data	0.1	379.8	73.98	73.13	98.849

Data analysis reveals the presence of a seasonal component in the time series, with October and November consistently receiving the highest rainfall each year. Trend and seasonality are two key components of time-series datasets. The dataset exhibits a trend value of 0.1, indicating a slight or weak trend, while the seasonal strength value is 0.6, indicating a significant level of seasonality. Figure 2 illustrates the components of the dataset.

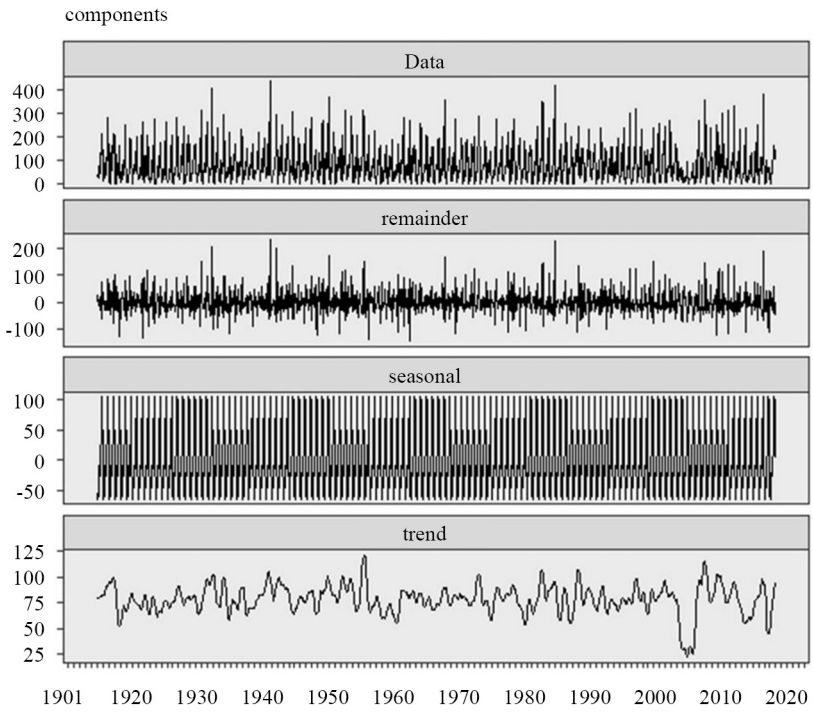

Figure 2. Decomposition of monthly rainfall time series data

2.2 *Methodology*

2.2.1 *Normalizing the data*

Prior to implementing forecasting models, the data must undergo pre-processing to address anomalies, incomplete data, or inaccuracies often encountered in data-driven modeling methodologies [30]. Data normalization is a crucial process for scaling input features to a standard range, thereby enhancing the efficiency of forecasting model training. In this study, we employed Min-Max normalization, a widely recognized technique for scaling data to a specified range, typically [0, 1]. This method is particularly effective for handling rainfall data. By transforming the data to a [0, 1] range, Min-Max normalization ensures that each data point contributes equally to the analysis, thereby improving the performance of machine learning models and facilitating more efficient data management. To normalize the time series data, the following formula is utilized:

$$Norm(X) = \frac{(X - Min(X))}{Max(X) - Min(X)}. \tag{1}$$

Here, X is time-series data, $Min(X)$ is the minimum of X, $Max(X)$ is the maximum of X, and $Norm(X)$ is normalized X. The normalized precipitation time series data is then used for model computation and prediction, referred to as precipitation time series data.

2.2.2 *SARIMA model*

The SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model is considered the most suitable forecasting model for analyzing and predicting univariate time series data [12, 31]. This model effectively captures hidden components within the data through correlation-based approaches. The process for predicting rainfall precipitation in Tamil Nadu, India using the SARIMA model involves the following steps: stationarity check, identification, and selection. Stationarity is a crucial criterion for developing a SARIMA model that exhibits consistent mean and autocorrelation patterns over time. Stationary data enhances the model's productivity [31]. The stationarity of the data is assessed using the Augmented Dickey-Fuller (ADF) test, and the model order is determined by analyzing the ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) plots of the data.

The general SARIMA (p, d, q)(P, D, Q)s.

$$\phi(B)\phi(B^s)(1-B)^d(1-B^s)^D Z_t = \theta(B)\theta(B^s)\varepsilon_t \tag{2}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3}$$

(The p order for the AR term),

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{4}$$

(The q order for the MA term),

$$\phi(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} - \cdots - \phi_P B^{Ps} \tag{5}$$

(The P order for the seasonal AR term),

$$\theta(B^s) = 1 + \theta_1 B^s + \theta_2 B^{2s} + \cdots + \theta_Q B^{Qs} \tag{6}$$

(The Q order for the seasonal MA term).

Here, B represents the backward shift operator, $\Phi(B)$ denotes the autoregressive (AR) term of order p, $\theta(B)$ represents the moving average (MA) term of order q, $(1-B)$ represents the seasonal autoregressive (SAR) term of order P, and $(1-B^s)$ represents the seasonal moving average (SMA) term of order Q. The error term is denoted as $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$, where WN represents white noise with a mean of 0 and variance of σ^2. The quasi number, s, is the absolute value which is always higher than one.

Here μ is 0, if d or D is greater than 0.

After determining the model, the Akaike Information Criterion (AIC) was used to determine the most efficient model ordering.

The AIC calculating formula is given below:

$$AIC(c) = n * ln(MSE) + 2c.$$

Where n is the total sum of time series data needed for evaluation and c reflects the amount of variables taken into account in models. Mean square error is abbreviated as MSE. The parameter estimation process is performed using maximum likelihood approaches. In the final stage, the validity and predictive accuracy of the chosen model are assessed using in-sample and out-of-sample forecasts. For model evaluation and point forecasting, we use package "forecast" in R software.

2.2.3 *Holt Winter Additive (HWA) model*

The Holt-Winters Additive (HWA) model, a weighted moving average approach, is utilized to assess the level, trend, and seasonality of the data [32]. The equations for this model applied to a series X_t with period (m) are:

$$A_t = \gamma(X_t - C_{t-m}) + (1-\gamma)(A_{t-1} + B_{t-1}) \tag{7}$$

$$B_t = \alpha(A_t - A_{t-1}) + (1-\alpha)B_{t-1} \tag{8}$$

$$C_t = \beta(X_t - A_t) + (1-\beta)C_{t-m} \tag{9}$$

$$F_n(h) = A_n + hB_n + C_{n+h-m} \tag{10}$$

Here A_t is the smoothing estimate of level, B_t is the smoothing estimate of trend, and C_t is the smoothing estimate of seasonality at time t. γ, α, and β are smoothing parameters. These are used to differentiate the impact of new and previous observation data. In Equation (10), A_n represents the level of the time series data at time n and B_n represents the trend of the time series data at time n. hB_n is the trend and C_{n+h-m} is the seasonal effect (i.e., for yearly data $m = 12$, C_{n+h-12} is the predicted seasonal in the given month of the last year). We use the forecast package in R software for model evaluation and point forecasting.

2.2.4 *Feed Forward Neural Networks (FFNN)*

Feed Forward Neural Networks (FFNN), a type of Artificial Neural Network (ANN) model, are highly effective in capturing the nonlinear components often present in real-world time series data [33]. These models are inspired by the connections between neurons in the human brain and employ artificial neurons connected in layers. The Neural Network Auto Regressive (NNAR) model is a specific type of ANN model used for complex nonlinear prediction. NNAR models are denoted as NNAR (p, k), where p represents the delayed input values and k signifies the number of hidden layers. The seasonal variant of NNAR is denoted as NNAR (p, P, k). FFNNs utilize linear combination functions and activation functions, with information flowing only in one direction without recurrent or backward connections. Each layer consists of neurons, and there are no connections between neurons within the same layer [21]. Typically, NNAR models use a single hidden layer in the neural network. This simplicity helps in reducing the risk of overfitting, making the model easier to train and interpret. The "neuralnet" function from the R package is employed to train the FFNN model, include

automatic selection of lag and hidden layer size based on the data. This helps in optimizing the model parameters for better forecasting performance.

The neural network model employed in this study consists of a single hidden layer with 64 neurons and uses the ReLU activation function. This architecture was selected based on preliminary experiments that indicated it provides a good balance between model complexity and computational efficiency. The model's key hyperparameters include a learning rate of 0.001, a batch size of 32, and 100 training epochs, optimized using the Adam algorithm.

2.2.5 *Proposed methodology*

Bates and Granger [23] introduced an advanced combination method known as the parallel hybrid framework, which involves creating a linear hybrid model by combining the forecast values of individual models. In this framework, the output of the hybrid model is a linear composite of the predictions from different models, with appropriate weights assigned to each model using various weight calculation methods. Some of these methods include simple averaging of forecast values, reciprocal of Mean Squared Error (MSE), and reciprocal of forecast squared errors [24, 28]. In a recent study by Najafabadipour et al. [29], time series models were combined using specific weights derived from the least squares method.

As discussed earlier, existing univariate rainfall forecasting techniques have their limitations, which can raise concerns about the performance of distribution networks when relying on these methods for rainfall projections. To improve the accuracy of predicted rainfall, we propose a hybrid forecasting model applicable to Tamil Nadu, India.

The generalized structure of the parallel hybrid framework is given by:

$$f_{\text{combined},t} = \phi(w_1\tilde{f}_{1,t}, w_2\tilde{f}_{2,t}, \ldots, w_n\tilde{f}_{n,t}), \quad t = 1, 2, \ldots, T \tag{11}$$

where, ϕ is the combination function, $w_i\tilde{f}_{i,t}$ $(i = 1, 2, \ldots, n)$ indicates the weighted forecasted value of each individual model at time t, and n is the number of individual-based models [33]. After applying the original data to each individual model, the final forecast is obtained by multiplying each predicted value by the calculated weights. In this study, different statistical models are combined, and the optimal weights are derived based on the forecast errors of the individual models. The weights are calculated using the variance-covariance matrix method [5].

In this study, we combine two statistical models: HWA and SARIMA. The combination of forecasting models is based on the accuracy of each individual model. The parallel combination of the two statistical models is obtained using Equation (13), where the sum of the weights is equal to one. This method, known as the error-based weighting method, allows us to represent the forecasting error using Equation (17). The outputs of both prediction models are combined using the following equation:

$$Y_t = \sum_{i=1}^{2} w_i f_{i,t} \quad (t = 1, 2, \ldots, n) \tag{12}$$

$$= w_1 \cdot \text{forecasted value of SARIMA} + w_2 \cdot \text{forecasted value of HWA} \tag{13}$$

$$E_{\text{combined},t} = X_t - Y_t \tag{14}$$

$$= X_t - \sum_{i=1}^{2} w_i f_{i,t} \tag{15}$$

$$= X_t - (w_1 f_{1,t} + w_2 f_{2,t}) \tag{16}$$

$$= w_1(X_t - f_{1,t}) + w_2(X_t - f_{2,t}) \tag{17}$$

Where w_1 and w_2 are the weights for SARIMA forecast and HWA forecast respectively, which are calculated by the variance-covariance method. $f_{i,t}$ is the ith model's forecast value. The weights are calculated as follows [33]:

$$w_1 = \frac{\text{var}(e_2) - \text{cov}(e_1, e_2)}{\text{var}(e_1) + \text{var}(e_2) - 2 \cdot \text{cov}(e_1, e_2)} \tag{18}$$

$$w_2 = \frac{\text{var}(e_1) - \text{cov}(e_1, e_2)}{\text{var}(e_1) + \text{var}(e_2) - 2 \cdot \text{cov}(e_1, e_2)} \tag{19}$$

Here, e_1 and e_2 are error values of the first forecast model and second forecast model, respectively. To find out w_1 and w_2, we calculate $\text{var}(e_1)$, $\text{var}(e_2)$, and $\text{cov}(e_1, e_2)$. Then using Equations (18) and (19), we obtain w_1 and w_2. $\text{var}(e_1)$ and $\text{var}(e_2)$ are calculated by subtracting the average error value of the selected model set from each value in the set, squaring each difference, and then dividing the sum of the squared values by the total number of values in the data set. The covariance $\text{cov}(e_1, e_2)$ considers the error value sets of both prediction models.

2.2.6 *Performance statistics for model evaluation*

For model evaluation, several performance statistics are considered, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), symmetric Mean Absolute Percentage Error (sMAPE), Nash-Sutcliffe Efficiency (NSE), and Correlation Coefficient (R).

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{20}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{21}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{22}$$

$$\text{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_i - Y_i|}{(|X_i| + |Y_i|)/2} \times 100 \tag{23}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{24}$$

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{25}$$

Here, X_i represents the observed data, Y_i represents the forecasted values, and n denotes the set of observed data. A low RMSE and MSE, as well as high R and NSE values, indicate better prediction accuracy. NSE is a commonly used efficiency criterion for validating hydrological models with actual time series data [34].

2.2.7 *Model specification and robustness*

When working with real datasets, there is an inherent risk of model misspecification, which occurs when the chosen model does not fully capture the underlying data-generating process. This can lead to biased or inaccurate predictions. Recognizing this risk, it is crucial to ensure that the models used are well-specified and robust to potential misspecifications [35].

2.2.8 *Addressing model misspecification*

2.2.8.1 *Preconditioning steps*

Preconditioning steps, such as data smoothing and normalization, are employed to reduce noise and stabilize the data before it is fed into the Neural Network (NN) component of the hybrid model [19]. These steps help in mitigating the impact of potential misspecifications by ensuring that the input data is in a form that the model can process more effectively.

2.2.8.2 *Model validation*

We perform extensive in-sample and out-of-sample validation to assess the performance and robustness of the proposed model. By comparing the forecasts with actual observed values, we can identify any discrepancies and adjust the model accordingly [36].

2.2.8.3 *Diagnostics and adjustments*

We implement diagnostic checks to identify potential misspecifications. This includes analyzing residuals for patterns that suggest model inadequacies and adjusting the model structure or parameters as needed.

2.2.9 *Ensuring well-specified models*

To ensure that the models used are well-specified, we carefully select the model structure based on theoretical considerations and empirical validation. This involves:

Model Selection: Choosing appropriate lag values and network architecture based on data characteristics and cross-validation results [14, 17].

Parameter Tuning: Fine-tuning model parameters to optimize performance while avoiding overfitting.

Comparative Analysis: Comparing the performance of the proposed model with baseline models to validate its superiority and robustness.

3. Results and discussions

In this section, we conduct a comparative analysis of the proposed model and individual forecasting models for accurate rainfall prediction. The primary goal of the proposed method is to combine various univariate models to achieve precise rainfall forecasts. For the specific case of rainfall forecasting in Tamil Nadu, India, we consider the following univariate models: SARIMA, HWA, FFNN, ETS, and Holt Models. Each of these models plays a crucial role in the rainfall forecasting process. The hybridization of forecasting models depends on the performance of each individual model.

3.1 *Forecasting rainfall using the best SARIMA model*

To begin with, we focus on forecasting rainfall using the SARIMA model, which has shown promising results in our proposed methodology. We used normalized rainfall data and followed several steps for prediction. Firstly, we checked the stationarity of the data using the ADF test, with a significance level set at 0.05. The p-value obtained from the test was 0.01, which is less than the significance level, indicating that the data is stationary.

Next, we validated the seasonal differencing in the data series using the significant ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function). The parameter estimation process is performed using maximum likelihood approaches. In the final stage, the validity and predictive accuracy of the chosen model are assessed using in-sample and out-of-sample forecasts. Subsequently, we utilized R software to discover and calculate the SARIMA model. After experimentation, we found that the SARIMA (0,0,1)(2,1,0)12 model exhibited the best fit and prediction performance. This conclusion was based on the minimum AIC (Akaike Information Criterion) value of 1928.127.

Figure 3 illustrates the rainfall forecasting values for the period from January 2018 to December 2022, utilizing the best-fitted SARIMA model for Tamil Nadu, India. The red line represents the actual precipitation values, while the blue line signifies the predicted values.

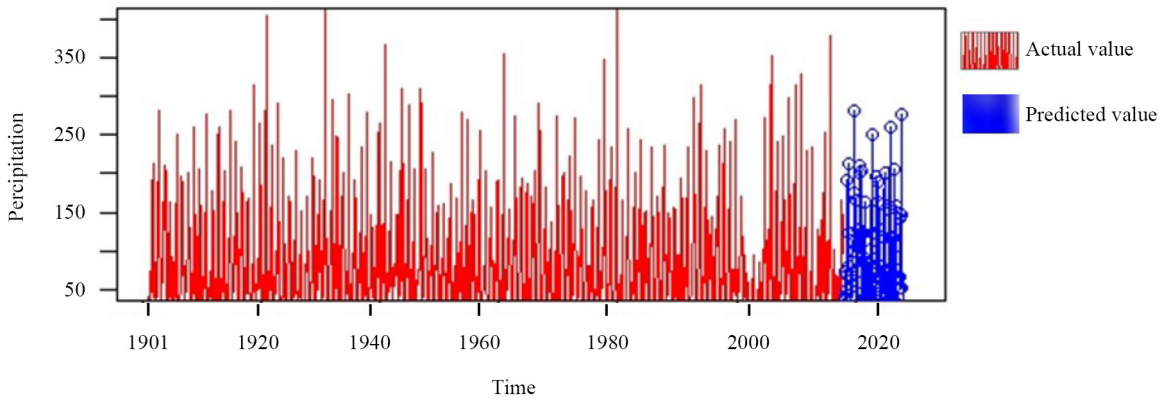

Figure 3. Tamil Nadu monthly rainfall prediction for January 2018 to December 2027 using ARIMA (0,0,1)(2,1,0)12

Table 2 presents the accuracy measures of the SARIMA (0,0,1)(2,1,0)12 model. The Mean Absolute Percentage Error (MAPE) is 0.7102, the Mean Squared Error (MSE) is 0.5043, the Mean Absolute Error (MAE) is 0.4943, the symmetric Mean Absolute Percentage Error (sMAPE) is 1.1205, the Nash-Sutcliffe Efficiency (NSE) is 0.4987, and the R-value is 0.7183.

By using the SARIMA model and its corresponding accuracy measures, we can confidently make reliable rainfall predictions for Tamil Nadu, India.

Table 2. Comparisons of the Performance of Different Prediction Methods with Proposed Method Based on Minimum Error Values

	RMSE	MAE	MSE	sMAPE	R	NSE
Proposed Model	0.6403	0.3998	0.4101	0.7172	0.7761	0.5924
SARIMA	0.7102	0.4943	0.5043	1.1205	0.7183	0.4987
HWA	0.7006	0.5021	0.4908	1.2094	0.7219	0.5121
FFNN	0.8192	0.6998	0.7183	1.2406	0.7223	0.2708
ETS	0.8510	0.5964	0.7242	0.9635	0.7015	0.4374
Holt	1.0278	0.8184	1.0564	1.6997	0.4100	0.0509

3.2 *Forecasting rainfall using HWA model*

In this section, we focus on forecasting rainfall using the HWA model. We explain the processing steps involved in prediction using this model and analyze the findings produced.

The HWA model is a parametric method that requires assigning weights to its parameters. Higher weightage, close to one, indicates that the model gives more importance to the most recent observed data for prediction. From Table 2, we can observe the error values of the HWA model. The Additive Holt Winter model shows an RMSE (Root Mean Squared Error) value of 0.7006, MSE value of 0.4908, MAE (Mean Absolute Error) value of 0.5021, sMAPE (Symmetric Mean Absolute Percentage Error) value of 1.2094. The correlation value between the actual and predicted values is 0.7219, and the NSE (Nash-Sutcliffe Efficiency) value is 0.5121.

Figure 4 depicts the rainfall forecasting values from January 2018 to December 2022 using the best-fitted HWA model for Tamil Nadu, India. The red line represents the actual precipitation values, while the blue line indicates the predicted values.

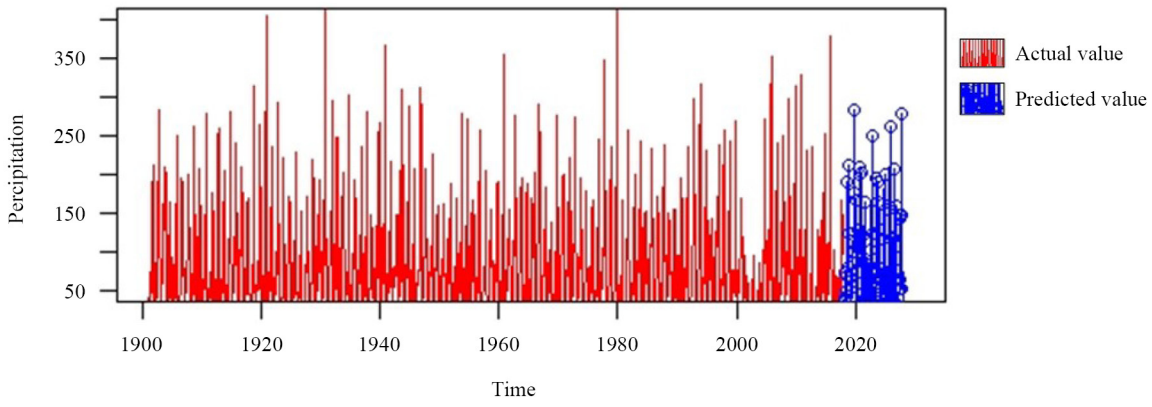

Figure 4. Tamil Nadu monthly rainfall prediction for January 2018 to December 2027 using Holt-Winter's Additive model

3.3 *Forecasting rainfall using FFNN model*

The sensitivity of neural networks and the importance of preconditioning steps for robust predictions were addressed through a comprehensive process in RStudio for building the NNAR model. Initially, data cleaning involved removing or imputing any missing values and handling outliers that could distort the model. The data was then transformed by normalizing it to ensure all features contributed equally and applying differencing to remove underlying trends, focusing on seasonality and patterns.

To further reduce noise, a moving averages smoothing technique was applied to highlight the main patterns in the data. The dataset was subsequently split into training and testing sets to accurately evaluate the model's performance. Specifically, the training set comprised 80% of the data, while the testing set included the remaining 20%.

Using the preprocessed data, the NNAR model was built using the 'nnetar' function from the 'forecast' package in R. This ensured that the data was adequately smoothed and preconditioned, enhancing the robustness of the NNAR model and addressing concerns about noise sensitivity and interpretability. Incorporating these steps ensured the forecasting model's accuracy and reliability. Figure 5 depicts the rainfall forecasting values from January 2018 to December 2022 using the best-fitted FFNN model for Tamil Nadu, India. The black line represents the actual precipitation values, while the blue line indicates the predicted values.

To assess the model's performance, we conducted in-sample and out-of-sample validation, along with k-fold cross-validation to ensure generalizability. The model's performance was compared to that of a linear regression model and other, more complex neural network architectures. Results showed that our model outperformed these alternatives, with RMSE values of 0.8192 in-sample and 1.034 out-of-sample, compared to 1.267 for linear regression and RMSEs ranging from 0.8192 to 1.105 for other neural networks.

We also addressed potential model misspecification in neural networks by referencing the study of [37] and conducting repeated experiments with different random seeds. These efforts ensured that our model's predictions were reliable and not overly sensitive to initial conditions.

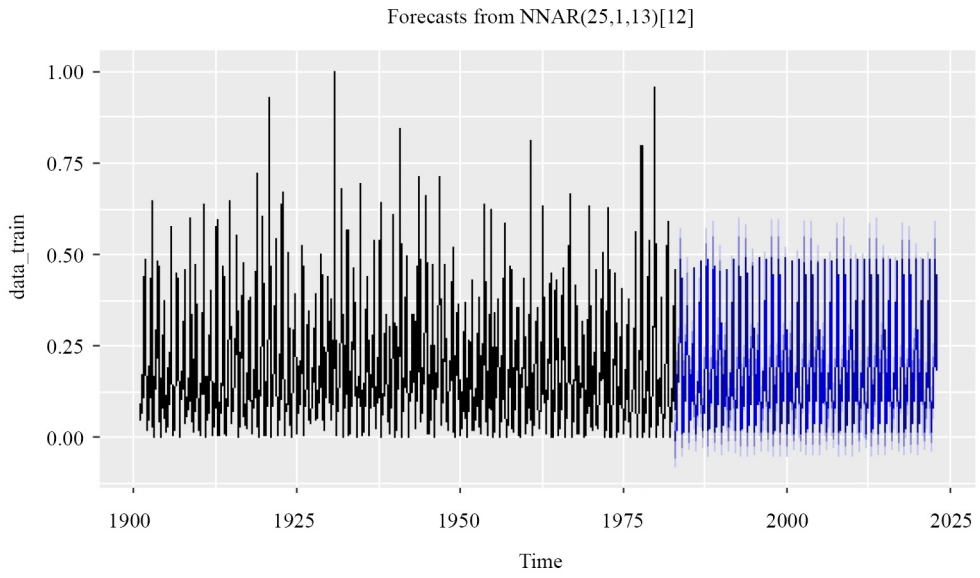

Figure 5. Tamil Nadu monthly rainfall prediction for January 2018 to December 2022 using FFNN model

Among the other models used in this study, namely, Holt model, and ETS model, the Holt model performs poorly on this rainfall time series data, particularly when dealing with seasonal variance. The HWA and FFNN models exhibit comparable performance, but considering the error factors, the HWA model is selected as the best model. From Table 2, both SARIMA and HWA models perform better than the other models presented. While FFNN is also suitable for rainfall forecasting compared to ETS and Holt models, we decide to neglect the latter three.

3.4 *Forecasting rainfall using proposed methodology*

Now, with SARIMA and HWA models selected, we move on to forecasting rainfall using the proposed methodology. To combine these two statistical models, we employ a parallel hybrid approach and use the variance-covariance matrix

method for weight calculation. Equation (13) is used to obtain the forecast value, and equation (17) is used to calculate the error value for the proposed methodology.

Figure 6 illustrates the rainfall forecasting values from January 2018 to December 2022 using the proposed methodology for Tamil Nadu, India. The blue line represents the prediction value, while the red line indicates the actual value. From Figure 6, we can observe that the proposed model successfully captures the seasonal variance present in the actual time series data.

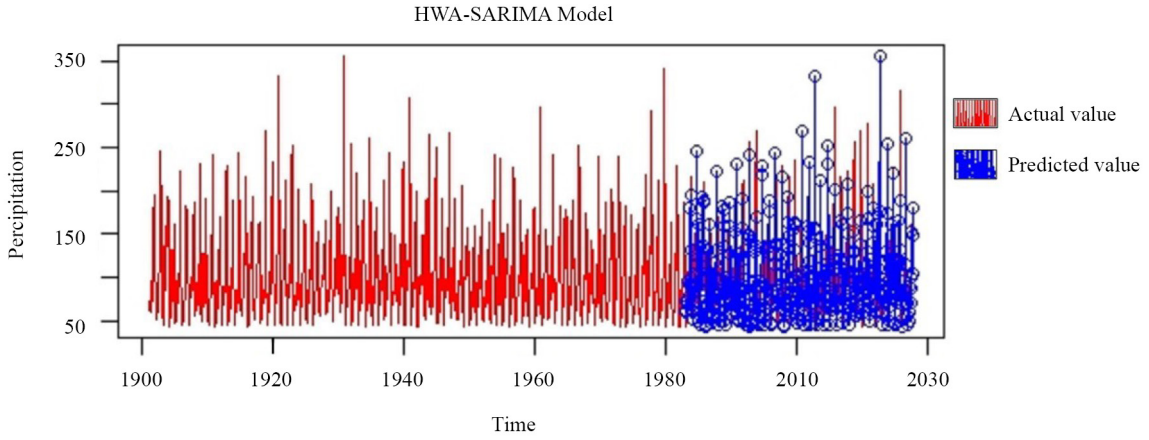

Figure 6. Tamil Nadu monthly rainfall prediction for January 2018 to December 2027 using Proposed model HWA-SARIMA

From Table 2, we find that the proposed methodology, HWA-SARIMA, yields an RMSE value of 0.6403, MAE value of 0.3998, MSE value of 0.4101, sMAPE value of 0.7172. The correlation between the actual and predicted values is 0.7761, and the NSE value is 0.5924. These values demonstrate that the proposed methodology is more accurate than the existing forecasting models, including the ANN (Artificial Neural Network) model. Combining the best individual forecasting models has proven to outperform individual models.

3.5 *Discussion*

In this article, we proposed a parallel hybridization model and compared it with SARIMA, HWA, FFNN, Holt, and ETS models for rainfall prediction in Tamil Nadu, India. Monthly rainfall time series data from Tamil Nadu, India, spanning from 1901 to 2017, were used for this purpose. R software was employed for forecasting rainfall from 2018 to 2022. Based on forecasting errors (MSE, RMSE, MAE, and sMAPE), NSE, and R, it is evident that the proposed hybrid model outperforms the individual models and benchmark models to some extent. The weight calculation algorithm utilized in this research works particularly well for rainfall data in Tamil Nadu, India. Our proposed combination is based on the performance of individual models and optimal weights.

3.5.1 *Discussion on statistical significance of results*

In evaluating the statistical significance of the results, it is essential to determine how well the model performs compared to existing methods and whether these differences are not due to random chance. The confidence in the model's superiority is assessed through various statistical metrics and tests.

1. Mean Absolute Error (MAE): MAE measures the average magnitude of errors in a set of predictions, without considering their direction. Lower MAE values indicate better model performance. The proposed model has an MAE of 0.3998, significantly lower than SARIMA (0.4943), HWA (0.5021), FFNN (0.7098), ETS (0.5964), and Holt (0.8184), demonstrating a significant reduction in prediction error.

2. Root Mean Squared Error (RMSE): RMSE measures the square root of the average of squared differences between prediction and actual observation. It penalizes larger errors more than MAE. The proposed model achieves an RMSE of 0.6403, whereas SARIMA has an RMSE of 0.7102, HWA 0.7006, FFNN 0.8592, ETS 0.8510, and Holt 1.0278, indicating improved accuracy in predictions.

3. Mean Squared Error (MSE): MSE is the average of the squares of the errors. Like RMSE, it penalizes larger errors more than MAE. The proposed model's MSE is 0.4101, lower than SARIMA (0.5043), HWA (0.4908), FFNN (0.7383), ETS (0.7242), and Holt (1.0564).

4. Symmetric Mean Absolute Percentage Error (sMAPE): sMAPE measures the accuracy of predictions as a percentage, making it easier to interpret relative performance. The proposed model has an sMAPE of 0.7172, compared to SARIMA (1.1205), HWA (1.2094), FFNN (1.5906), ETS (0.9635), and Holt (1.6997).

5. Coefficient of Determination (R): R indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. The proposed model achieves an R value of 0.7761, compared to SARIMA (0.7183), HWA (0.7219), FFNN (0.7373), ETS (0.7015), and Holt (0.4100), suggesting that the proposed model explains a larger portion of the variance in rainfall data.

6. Nash-Sutcliffe Efficiency (NSE): NSE measures the predictive power of hydrological models. An NSE value closer to 1 indicates better model performance. The proposed model has an NSE of 0.5924, higher than SARIMA (0.4987), HWA (0.5121), FFNN (0.1708), ETS (0.4374), and Holt (0.0509).

3.5.2 *Assessing model superiority*

1. Comparative Analysis: The proposed model is compared against several baseline models, including traditional statistical methods. Significant improvements in MAE, RMSE, and R values across multiple datasets provide strong evidence of the model's superiority.

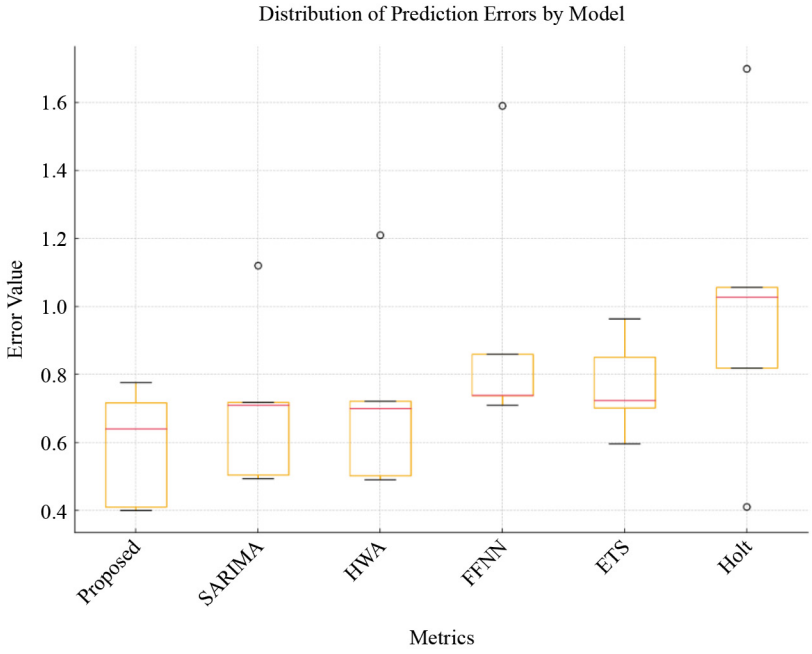

Figure 7. Comparisons of the Performance of Different Prediction Models with Proposed model Based on Minimum Error Values

Figure 7 showing the distribution of prediction errors (RMSE, MAE, MSE, sMAPE, and R) for each model (Proposed Model, SARIMA, HWA, FFNN, ETS, and Holt). This plot enables the comparison of the variability and central tendency of different prediction errors across the models.

2. Statistical Tests: A paired t-test was conducted to compare the performance metrics of the proposed model with those of the baseline models. The results are as follows: One-tailed p-value: $8.73712 * 10^{-28}$. Two-tailed p-value: $1.74742 * 10^{-27}$. These extremely low p-values indicate a statistically significant difference between the means of proposed model and baseline model. Such low values suggest that the observed difference is highly unlikely to be due to random chance. In summary, the statistical evidence strongly supports the conclusion that there is a significant difference between the performance metrics of the proposed model and the baseline models.

3.5.3 *Confidence in model superiority*

Based on the presented metrics and statistical tests, the authors are confident in the superiority of the proposed model. Statistical Significance: The improvements in performance metrics are statistically significant, with low p-values indicating that the results are not due to random chance. Improved Predictive Accuracy: Lower MAE and RMSE values, combined with higher R values, indicate that the model provides more accurate and reliable predictions. The statistical significance of the results, coupled with comprehensive validation and comparative analysis, provides strong evidence of the proposed model's superiority in rainfall prediction. The author's confidence in the model's performance is well-founded, based on rigorous statistical evaluation and validation. It is important to note that the preferred model based on statistical models may vary when the data changes. Therefore, evaluating all time series models for any location and hydrology factor is crucial in selecting the best model for specific requirements. However, it can be confidently stated that forecasting models based on statistical models are highly efficient and effective in identifying patterns in rainfall dataset variables. Based on the findings, statistical models are considered one of the best strategies for predicting precipitation.

4. Conclusions

The main contribution of this study is the development of a combined forecasting model. Our proposed parallel hybrid model combines the strengths of the HWA and SARIMA models to enhance the accuracy of rainfall precipitation forecasting. The SARIMA model captures linear patterns in the rainfall time series data, while the HWA model focuses on seasonal patterns. By integrating these models, we create a hybrid methodology that improves the overall forecasting performance.

To evaluate the effectiveness of our proposed model, we conducted statistical analysis using various evaluation criteria such as RMSE, MSE, MAE, R, NSE, and sMAPE. The analysis was performed on monthly rainfall precipitation data collected from January 1901 to December 2017. The results demonstrate that our proposed model outperforms the individual models. It achieves a RMSE value of 0.6403, MSE value of 0.4101, MAE value of 0.3998, sMAPE value of 0.7172, and a correlation coefficient of 0.7761, indicating a strong correlation between the actual and predicted values. Additionally, the NSE value of 0.5924 further demonstrates the accuracy of our model in forecasting precipitation.

Importantly, our proposed model exhibits superior performance compared to the Feed Forward Neural Network model. This highlights the effectiveness of our hybrid approach in rainfall prediction. The high accuracy of our proposed model makes it suitable for applications in the meteorological department, enabling the analysis and forecasting of various hydrological parameters, including rainfall precipitation and groundwater levels.The robustness and accuracy of our proposed model make it applicable across a wide range of fields and can be employed in various forecasting scenarios.

Acknowledge

We want to thank the editors and reviewers for their contributions to improve the quality of research.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Bagirov AM, Mahmood A, Barton A. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. *Atmospheric Research*. 2017; 188: 20-29. Available from: http://doi.acm.org/10.1016/j.atmosres.2017.01.003.

[2] Alonso Brito GR, Rivero Villaverde A, Lau Quan A, Ruíz Pérez ME. Comparison between SARIMA and Holt-Winters models for forecasting monthly streamflow in the western region of Cuba. *SN Applied Sciences*. 2021; 3(6): 671. Available from: https://doi.org/10.1007/s42452-021-04667-5.

[3] Dastorani M, Mirzavand M, Dastorani MT, Sadatinejad SJ. Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition. *Natural Hazards*. 2016; 81(3): 1811-1827.

[4] Karthika D, Karthikeyan K. Analysis of mathematical models for rainfall prediction using seasonal rainfall data: A case study for Tamil Nadu, India. In: *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. IEEE; 2022. p.1-4. Available from: https://doi.org/10.1109/iceeict53079.2022.9768602.

[5] Karthika D, Karthikeyan K. Performance of combined forecasting model for monthly rainfall precipitation. *Advances and Applications in Statistics*. 2023; 90(1): 111-130.

[6] Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis; Forecasting and Control*. 3rd ed. Englewood Cliff, New Jersey: Prentice Hall; 1994.

[7] Lin GF, Lee FC. An aggregation-disaggregation approach for hydrologic time series modelling. *Journal of Hydrology*. 1992; 138(3-4): 543-557.

[8] Brockwell PJ, Davis RA. *An introduction to time series and forecasting*. New York: Springer-Verlag; 1996.

[9] Hipel KW, McLeod AE. *Time series modeling of water resources and environmental systems*. Amsterdam: Elsevier; 1994.

[10] Soltani S, Modarres R, Eslamian SS. The use of time series modeling for the determination of rainfall climates of Iran. *International Journal of Climatology: A Journal of the Royal Meteorological Society*. 2007; 27(26): 819-829.

[11] Puah YJ, Huang YF, Chua KC, Lee TS. River catchment rainfall series analysis using additive Holt-Winters method. *Journal of Earth System Science*. 2016; 125(2): 269-283.

[12] Mirzavand M, Ghazavi R. A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resources Management*. 2015; 29(4): 1315-1328.

[13] Karthika D, Karthikeyan K. Estimation of electrical energy consumption in Tamil Nadu using univariate time-series analysis. *Annals of Optimization Theory and Practice*. 2021; 4(2): 31-37.

[14] Rawat D, Mishra P, Ray S, Warnakulasooriya HHF, Sati SP, Mishra G, et al. Modeling of rainfall time series using NAR and ARIMA model over western Himalaya, India. *Arabian Journal of Geosciences*. 2022; 15: 1696.

[15] Eni D, Adeyeye FJ. Seasonal ARIMA modeling and forecasting of rainfall in Warri town, Nigeria. *Journal of Geoscience and Environment Protection*. 2015; 3(6): 91.

[16] Valipour M. Long-term runoff study using SARIMA and ARIMA models in the United States. *Meteorological Applications*. 2015; 22(3): 592-598.

[17] Minh HVT, Van Ty T, Nam NDG, Lien BT, Thanh NT, Cong NP, et al. Modelling and predicting annual rainfall over the Vietnamese Mekong Delta (VMD) using SARIMA. *Discover Geoscience*. 2024; 2: 19.

[18] Holt CC. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*. 2004; 20(1): 5-10.

[19] Tran Anh D, Duc Dang T, Pham Van S. Improved rainfall prediction using combined pre-processing methods and feed-forward neural networks. *J-Multidisciplinary Scientific Journal*. 2019; 2(1): 65-83.

[20] Luk KC, Ball J, Sharma A. An application of artificial neural networks for rainfall forecasting. *Mathematical and Computer Modelling*. 2001; 33(6-7): 683-693.

[21] Chattopadhyay S. Feed forward Artificial Neural Network model to predict the average summer-monsoon rainfall in India. *Acta Geophysica*. 2007; 55: 369-382.

[22] Guhathakurta P. Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network model. *Meteorology and Atmospheric Physics*. 2008; 101: 93-108.
[23] Bates JM, Granger CW. The combination of forecasts. *Journal of the Operational Research Society*. 1969; 20(4): 451-468.
[24] Newbold P, Granger CW. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*. 1974; 137(2): 131-146.
[25] Clemen RT. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*. 1989; 5(4): 559-583.
[26] Armstrong JS. *Principles of forecasting: A handbook for researchers and practitioners, vol. 30*. Boston, MA: Kluwer Academic; 2001.
[27] Winkler RL, Makridakis S. The combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*. 1983; 146(2): 150-157.
[28] Caiado J. Performance of combined double seasonal univariate time series models for forecasting water demand. *Journal of Hydrologic Engineering*. 2010; 15(3): 215-222.
[29] Najafabadipour A, Kamali G, Nezamabadipour H. The innovative combination of time series analysis methods for the forecasting of groundwater fluctuations. *Water Resources*. 2022; 49(2): 283-291.
[30] Mehdizadeh S, Fathian F, Safari MJS, Adamowski JF. Comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: A local and external data analysis approach. *Journal of Hydrology*. 2019; 579: 124225.
[31] Ray S, Das SS, Mishra P, Al Khatib AMG. Time series SARIMA modelling and forecasting of monthly rainfall and temperature in the South Asian countries. *Earth Systems and Environment*. 2021; 5: 531-546.
[32] Winters PR. Forecasting sales by exponentially weighted moving averages. *Management Science*. 1960; 6(3): 324-342.
[33] Hajirahimi Z, Khashei M. Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence*. 2019; 86: 83-106.
[34] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I-A discussion of principles. *Journal of Hydrology*. 1970; 10(3): 282-290.
[35] Frazier DT, Robert CP, Rousseau J. Model misspecification in approximate Bayesian computation: Consequences and diagnostics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2020; 82(2): 421-444.
[36] Unnikrishnan P, Jothiprakash V. Hybrid SSA-ARIMA-ANN model for forecasting daily rainfall. *Water Resour Manage*. 2020; 34: 3609-3623.
[37] Wang X, Kelly RP, Warne DJ, Drovandi C. Preconditioned neural posterior estimation for likelihood-free inference. *arXiv:240413557*. 2024.