International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I6P127
Volume 74 Issue 6, 401-420, June 2026

Original Article

Seasonal Rainfall Prediction Using SARIMAX and ARIMA Models

M. Nivedhitha[1], S. Meganathan[1*], A. Sumathi[2], R. Rajakumar[3]

[1,2,3]*Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, SASTRA University, Tamil Nadu, India.*

[*]*Corresponding Author : meganathan@src.sastra.edu*

Received: 31 December 2025 Revised: 02 April 2026 Accepted: 20 April 2026 Published: 27 June 2026

Abstract *- Analysis of rainfall and weather variability in Chennai is carried out using twenty years of historical data (2003-2023) collected from the India Meteorological Department. The study examines the rainfall response to key meteorological factors such as temperature, wind speed, visibility, and dew point related to the Northeast Monsoon during the season from October to early January. Accordingly, the data was statistically processed and divided into seasonal segments that defined the NEM season. At the same time, PRCP values were grouped into seven categories from no rain to extreme heavy rainfall to account for the dynamic changes in rainfall intensity over time. Likewise, quantile-based grouping was used for other weather variables to divide them into low, medium, and high exposure categories for easier comparison. In response to the limitations of conventional ARIMA models that neglect the seasonal effects of monsoons and the nonlinear relationships between weather variables, this paper proposes the Extreme-Sensitive Hybrid Residual Forecasting (ES-HRF) method. This approach combines SARIMAX with Fourier components for handling seasonal series and exogenous weather series for handling temporal series, and further refines residuals with nonlinear learning from Gradient Boosting and Random Forest Regression. Extreme rainfall events are defined using quantile-based thresholds, and their performance is assessed using categorical verification scores. The proposed hybrid framework shows better prediction accuracy and robust extreme event detection performance than the individual statistical and machine learning models. The results emphasize the efficacy of the proposed method in improving the reliability of rainfall prediction and facilitating disaster preparedness, flood control, and weather-smart urban planning in monsoon-dominated areas.*

Keywords *- Climate trend analysis, Northeast monsoon, Quantile classification, Rainfall pattern analysis, SARIMAX, ES-HRF, Statistical approach.*

1. Introduction

Precipitation is crucial in driving the climate on Earth, supporting stability in ecosystems, and impacting human livelihoods. Precipitation touches on agriculture, water, and energy production and is therefore a primary concern for environmental sustainability. Human-induced changes in climate have been strongly related to changes in precipitation patterns in countless places across the globe. In many urban areas, precipitation patterns have recently become less predictable regarding monsoon behaviour, which has led to issues such as flooding, drought, and impacts on agricultural and infrastructure performance [19]. The cities that experience the effects of unpredictable precipitation patterns, such as Chennai in the Indian state of Tamil Nadu, have been identified as one of the urban centres most vulnerable to variation in its monsoon and extreme weather events. The climate regime of Chennai is distinct in India because it experiences most of its annual rainfall during the Northeast Monsoon (NEM) from October to January. The Northwest is the Primary mechanism for rain for most of India, but for Chennai, it is the NEM. Chennai experiences a highly concentrated rainfall pattern, with nearly 60 per cent of its annual precipitation occurring within a limited time frame. Consequently, the Northeast Monsoon affects not only the water security of the city but also agricultural productivity, groundwater recharge, sub-surface hydrology, urban drainage, and built infrastructure. Any shifts in this climate that modify the timing, or introduce a change in precipitation, can have damaging impacts on water security, whether they be prolonged periods of dry and drought or sustained flooding that impacts the urban fabric and fabric of stability, as observed and documented in the floods that occurred in 2015 and 2023. For this reason, it is important to understand the dynamics of the monsoon in Chennai and predict it appropriately. The predictions of key climatic variables such as temperature, dew point, visibility, and wind speed. During the Northeast Monsoon (NEM) period, changes to the usually stable atmosphere that occur before amounts of rain can be identified. The values of these climatic variables are related and reflective of the system's dynamic character over the Bay of Bengal and

the Chennai coastal zone. Variation in wind speed changes moisture transport, temperature, and dew point, which determine condensation amounts, and visibility is affected by both humidity and precipitation intensity, and pollution level. All of these variables join together to create the climate profile of Chennai during the monsoon [20]. To capture such complex temporal relationships, it is necessary to use modern time-series forecasting models. Among the modern time series forecasting models, the Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX) model is of significant interest because it can model both seasonal and exogenous climatic relationships. The SARIMAX model is built upon the original ARIMA model but incorporates exogenous variables thought to relate to the target variable in a direct exogenous manner. For this reason, the SARIMAX model is an excellent candidate for predicting rainfall and climate based on these variables.

In this research, SARIMA will be used to analyse long-term climatic data and predict shifts in major atmospheric variables during the Northeast Monsoon season in Chennai. Standard precipitation forecasting methods involve relatively straightforward regression and correlation methods that cannot adequately fit nonlinear and seasonal variations. SARIMAX can independently account for both short-term dependencies and long-term seasonalities, providing a better-fitting, more interpretable method for modelling meteorological data. In particular, it can capture repeating seasonal patterns such as a consistently four-month monsoon season, whilst adapting to minor local and global climate changes on an annual basis. This period (2003 - 2023) is particularly useful when assessing climate trends in Chennai. In the last two decades, the community has encountered a litany of extreme events, including flooding in 2015 and again in 2021, as well as several years of less than average precipitation (2016 and 2018, respectively). These local microclimates are a function of urbanisation and land use, which has resulted in extremes of both rainfall and temperature. Assessing this time period is important for changing climatic characteristics of the monsoon time frame due to anthropogenic environmental impacts.

Various studies have focused on the application of statistical, machine learning, and deep learning models for climate and rainfall prediction. The hybrid WD-SARIMAX models were able to address the issues of non-stationarity and seasonality, but were limited to only single regions and further failed to capture sudden changes in climate variability [11]. Comparative studies showed that LSTM performed better than SARIMAX in modelling nonlinear and seasonal rainfall; however, neither model handled extreme events that resulted from sparse data. In addition, SARIMAX faced the problem of underfitting, while LSTM overfitted in such scenarios [12]. Another regional study was performed using ANN, RF, SVM, and LSTM, where better performance was noted with ANN, though it lacked the integration of real-time data [13]. Other approaches involve the use of GNSS-derived variables, where high detection of rainfall is noted, though it has its geographical limitations [14]. The main objective of the present study is to model and predict climatic variables associated with Chennai's Northeast Monsoon using the SARIMAX model. This will help facilitate improved understanding of the temporal dynamics associated with temperature, dew point, visibility, and wind speed, which have all been shown to influence monsoon and rainfall activity. The findings will allow for the ability to assess the strength of the atmospheric coupling of these elements, as well as the implications of coupling for the predictability of rainfall variability.

Chennai, as an urban and coastal location, complicates matters further. All weather conditions quickly change local weather parameters and subsequently increase the rate of rainfall. These patterns can be derived in a data-based manner through a SARIMA model and provide an early warning and response to the risk regime. Estimating wind speed and visibility trends specifically can aid in cyclone trajectory estimates, as well as rainfall system forecasts. This study contributes to the growing body of evidence and community of practice around data-based climate models by demonstrating that long-term meteorological data can still be used in any capacity examined through a statistical learning framework. The ability to look at multiple climate variables allows for the identification of interdependencies that are often overlooked when a single rain variable does not present itself as significant. Additionally, the use of a SARIMA model will assist in bridging the pure statistical solutions through exploring a dynamic approach, as it relates to multiple dynamics in environmental analysis. This study is to illustrate the analysis and forecasting of relevant meteorological parameters pertinent to the Northeast Monsoon in Chennai using a SARIMAX model. The assessment of the twenty years in the dataset seeks to investigate the temporal dynamics of climatic variables associated with altered rainfall in coastal urban environments. These may benefit local policymakers, meteorologists, and urban planners in aiding climate change adaptation for Chennai, which requires increased resilience to the range of impacts threatening the city from climate change. For instance, while global climate change has already influenced changes in temperature and rainfall, local studies can assess the contributions of climatic variables to urban sustainability and livability through statistical models such as SARIMAX.

Though some research works have utilized ARIMA and SARIMAX models for rainfall prediction, most of them are linear in nature and do not address the nonlinear interactions of weather and extreme rainfall variability. Recent work has started to compare hybrid and deep learning models; however, few research works have combined seasonal statistical modelling with nonlinear residual correction and extreme rainfall sensitivity. Additionally, extreme rainfall

identification is considered a separate task in most works, rather than being incorporated into the forecasting model. The current work aims to fill this research gap by introducing an Extreme-Sensitive Hybrid Residual Forecasting framework that combines seasonal patterns, exogenous weather variables, regime clustering, and stacked residual learning to improve overall forecasting performance as well as extreme rainfall detection.

2. Related Work

Kumari et al. (2025) analysed monsoon rainfall variability over Northern India from 1981 to 2020 using statistical trend analyses, rainfall intensity based on percentiles, and climate model attribution methodologies. They utilised the CMIP6 Detection and Attribution under the Climate Model Intercomparison Project Phase 6 to determine how aerosol emissions, greenhouse gas concentrations, and natural climate forcing modifications affected changes in monsoon rainfall. For estimating the impact of aerosols on monsoon dynamics, the WRF model with Aerosol Chemistry was used in the study. Their analysis revealed an inverse relationship (rainfall dipole) between northwestern India and the Indo-Gangetic plains with respect to changes in moisture transport frequency of monsoon depressions and stability in the atmosphere. Limitations include reliance on model outputs, a restricted spatial domain to only northern India, and limited consideration for urbanisation and localised land use changes that may further contribute to defining regional rainfall variability [6]. Deng et al. (2025) applied interpretable machine learning methods and the multivariate wavelet coherency method to global land monsoon regions over the period 1960-2022 to quantify which climatic teleconnections drive rainfall seasonality. They found weakening and strengthening patterns of rainfall seasonality for a number of monsoon hotspots, together with abrupt shifts in several regions related to nonlinear teleconnection behaviour. In this study, it was shown that the impact of the most important climate teleconnections on the timing and distribution of rainfall has strengthened since the 1990s. The approach is, however, limited by the reliance on large-scale teleconnection indices; the understanding of local climate drivers is only crudely accounted for, and there is considerable uncertainty in observational datasets over six decades [7].

A comprehensive evaluation of the worldwide consequences from 2004 to 2024, where 2,091 instances of disruption caused by extreme weather, has been performed by McKinley and colleagues via qualitative content analysis, classification of each instance, as well as mapping trends, all to better understand how extreme weather conditions disrupt large group/event activities as well as what type of hazards have the most severe impact on these events. There are marked differences in the level of vulnerability, dependent on geographic regions. The highest rates of disruption are found with arts, cultural, festival-type activities, especially with the level of vulnerability among countries where English is spoken that have a mature event economy. This research links disruptions caused by extreme weather with current research regarding the attribution of the weather patterns that create these hazards; thus, there is an increase in the levels of both frequency and severity over time. However, there are limitations to be drawn from the study due to incomplete reporting from around the world, significant bias in regional reporting, as well as the inability to produce predictive modelling quantitatively [8]. Fathima et al. reconstruct long-term monsoon variability in the Andaman Sea by using paleo-proxy techniques, including planktic foraminiferal abundance analysis, magnetic susceptibility, and Red/Blue reflectance ratios from the IODP Site U1448. Their findings reveal prominent stadial-interstadial shifts in the Indian Summer and Winter Monsoon intensity driven predominantly by precession-modulated insolation and by high-latitude climatic events such as the Younger Dryas and Heinrich cycles. In addition, the outcome depicts a tight coupling between monsoon strength, productivity, CaCO content, and organic carbon level. However, these findings are limited to the regional nature of sediment core evidence, poor sampling within the Andaman Sea, and uncertainty factors associated with paleo-climate proxy interpretation [9].

Isa et al. attempted to study Isolated Breakdown Pulse (IBP) train discharges in tropical thunderstorms using high-resolution broadband electric field measurements in the range of 12.5-25 ns, and are complemented by comprehensive waveform classification techniques. Their results indicated that positive IBP trains and cloud-to-ground flash polarity and frequency were highly influenced under cool and humid conditions brought about by the monsoon. In their work, they were able to classify two different types of IBP train events, each having different pulse configurations, and came up with the conclusion that high occurrences of positive IBP trains have strong links with increased positive CG flashes. However, these findings were limited by a small sample size of 90 events, regional observation constraints, and no advanced modelling or machine learning-based prediction studies on IBP-CG interactions [10].

The authors of the article, Chacón-Maldonado et al., proposed a new technique for creating an advanced forecasting approach for estimating monsoon rainfall. The authors describe the techniques that were developed to utilize machine learning to gather those systems from a variety of datasets based on the SHAP evaluation criteria and a combination of wrapper, rank, and search algorithms, as well as an evaluation metric based on SHAP (SHapley Additive exPlanations) to generate meaningful, interpretable insight into how each feature contributes to the predictions made by a model. Their results indicate that prediction models trained on features selected using automated processes can outperform those using expert judgment. Hence, they increased the accuracy of the prediction of extreme events associated with the Western North Pacific Summer Monsoon. These findings also provide

insight into how the use of XAI can support identifying key climate indices and enhancing model transparency for decision-making. Limitations include high dependency on historical climate data, the possibility of overfitting by feature selection, and reduced generalizability when applied to different monsoon regions or various climate regimes, including both current climate change and future climate variations [15].

Prasad, Nagendra, et al- An analysis of sediment (a lacustrine core 1.2 m long) from lakes in Central India was used by Laser on HALLU (2024) to make a reconstruction of both the Late Holocene monsoonal and vegetation changes based on data obtained from pollen analysis. Reconstruction techniques were completed with various chronological dating methods and paleoclimatic interpretations. The findings of this study indicated that since approximately 2.5 ka ago, the vegetation in central India had transitioned from deciduous forests in the open to mixed, densely populated forests. The findings were also closely correlated to variations in ISMR. There is evidence, in addition to the findings, of increased human activity, particularly through the development of agriculturally based systems of the economy, that started approximately 1035 cal yr BP and has continued up to the present. However, while the findings are from a single-core data set with limited regional representation, there were no advanced modelling techniques to address the climate vs. vegetation relationship for the purpose of understanding the variance of monsoon patterns across the region [16].

Sharma et al. (2024) employed multiple linear regression, principal component analysis, and correlation analyses to study two decades of radiosonde observations with a view to investigating how a warming climate alters cloud macro-physical properties over the Indian Summer Monsoon region. Its findings indicate that an increase in high-level clouds and cloudy days, with a decrease in low-level clouds, is strongly related to global warming and ENSO-driven changes in the atmosphere. Its study identified the tropospheric expansion and upward shift of clouds as being at the root of such trends. But this analysis is solely dependent on the data from radiosondes, having limited spatial resolution without an advanced modelling technique, which could have enhanced the strength of causal attribution [17]. Subrahmanyam et al. (2023) use gridded climatological rainfall data with trend analysis, probability density distribution analysis, and diagnostics of atmospheric circulation parameters and discuss the long-term variations in ISMR from 1901 to 2022. Examples include significant temporal changes in the contributions from active and break spells, the slowdown of northward monsoon propagation speed, and changes in cloud radiative forcing and drop-size microphysics. Their findings allow for evolving monsoon dynamics driven by climatic changes that affect rainfall duration and intensity patterns. However, the results are limited by being based on inexact historical gridded datasets, the sparsity of microphysical measurements, and the absence of predictive modelling that may allow confirmation of quantitative estimates of future change [18]. A ten-day daily rainfall forecasting system was introduced by Ruhiat et al. in 2024, utilising Singular Spectrum Analysis (SSA) and Seasonal ARIMA (SARIMA) models to identify the start of the rainy season according to BMKG criteria. A total of 20 years' (2001-2020) rainfall data in the Citarum-Majalaya watershed, Indonesia, was analysed using the Thiessen polygon method. Model results were compared using MAPE, whereby SSA, with 36.8%, slightly beat the SARIMA model, with 40.0% accuracy in short-term forecasting of up to 6 months. Results revealed that the predicted start of the rainy season was correctly pinpointed by the SSA pattern to be at the end of October, obtaining the highest probability, yet having poor performances when predicted beyond 6 months by the other models due to randomness and linearity. Its drawback is being regional, dependent on historical data, and not involving other parameter variables like ENSO and IOD. Future scope would explore the utilisation of machine learning models, hybrid deep learning models, adding other variables, or other multitopic forecasting models [21].

Muzaffar et al. (2017) formulated a small-resource-demanding, IoT-enabled method for agricultural observation. This system provides sensor readings of soil moisture content, air temperature, electrical conductivity of the soil, and soil nutrients through an IoT application and integrates with an ARIMA forecasting model to provide an accurate estimate of rainfall for 7 days ahead. Using the information on what's expected for rainfall allows a farmer to determine his/her best time for irrigating crops without wasting water. Experimental calculations showed that the use of IoT and ARIMA together will increase human resource efficiency (better yields). A 15.8% MAPE means that with this combination, a farmer can increase their yield by work and time efficiency by approximately 15.8%. Furthermore, this combination may be suitable for low-resource environments due to the relative simplicity and ease of interpretation of the ARIMA model, yet it also relies on a single weather source. Also, the assumption of stationarity may not always be appropriate for predicting rainfall patterns due to changing climate conditions. The use of hybrid machine learning/deep learning models combined with multi-location weather data could potentially increase the accuracy and reliability of ARIMA rainfall forecasting for farmers [22].

Tee Huey Yin and Rosnalini Mansor (2024) offered a study with the title "Forecasting Rainfall Volume in Selangor with a Combined ARIMA Model," which was published in "Journal of Computational Innovation and Analytics." The purpose of their research work was to investigate and forecast the volume of rainfall that occurs every month in Selangor, which will help in mitigating flooding during its initial stage. The volume of rainfall from 2018 to 2022 for the Petaling and Subang rainfall stations was counted and identified using various univariate time series models such as Naïve models,

Decomposition models, Exponential Smoothing models, ARIMA models, and combined models of ARIMA. The authors of the research work developed models using combined ARIMA models based on their weighted performances. Findings from their study indicated that the models with minimal errors in forecasting rainfall during its short-term periods included ARIMA (2,0,3) and ARIMA (4,0,4).

The study showed that rainfall data and data from other variables/parameters tend to follow similar patterns that are very seasonal and also follow a linear trend. The major limitation of this research study and work was that they failed to consider other variables, such as wind speed and humidity, which influence rainfall. The short period of data also limits climate change. Future studies and research works should focus on incorporating other variables that influence rainfall or even using machine learning. The study may also help in creating a flood forecast program [23].

A hybrid framework for forecasting rainfall has been proposed by Wang et al. (2021). The proposed framework uses a combination of Wavelet Packet Decomposition (WPD) and Back Propagation Neural Network (BPNN), Generalised Method of Data Handling (GMDH), and Autoregressive Integrated Moving Average (ARIMA) to produce more accurate long-term predictions. The non-stationary time series of monthly rainfall recorded at Luoning and Zuoyu (China) stations were decomposed using WPD technology into several frequency-based components. All components were then separately forecasted using the stated individual forecasting models. Performance of the hybrid models was evaluated by RMSE, MAE, R, and NSEC values.

The results reveal that WPD-based models consistently outperform standalone models, with a WPD-BPNN forecast model giving the best results. The study also demonstrates the significant advantage of preprocessing data before applying any forecasting techniques. However, the work was limited in scope as all models (including hybrid) used only historical rainfall measurements, ignoring any influence of meteorological characteristics such as temperature and humidity. Additionally, the multi-level WPD method and combined models are computationally intensive methods to implement. Even with several different models for combined inputs, none of these models can accurately capture how rainfall varies depending on location (spatial heterogeneity). Future research could provide additional possibilities by developing greater robustness/scalability when integrating other models, using deep learning algorithms, combining multiple sources of forecast data, or employing real-time precipitation data [24].

A Seasonal ARIMA (SARIMA) forecast method for predicting future rainfall from previous years' monthly rainfall records from January 2006 through February 2016 was created by Arumugam P. & Saranya R. (2018). They used an approach to filling gaps left by missing observations in monthly data (which they did using 'Mean Substitutions'). They also looked at the presence of outliers (greater than 3 standard deviations from the mean) through examining the residual error of their fitted SARIMA models. They identified Innovative Outliers (IO) and Seasonal Additive Outliers (SAO) as a way of enhancing forecast accuracy and stabilising their models. The SARIMA (1, 1, 1)/(0, 1, 1)12 models (with a constant '12') performed best when modelling and forecasting seasonal patterns of rainfall. The researchers determined that by excluding anomalous observations and replacing missing observations with Mean Substitutes, the prediction performance of the fitted models improved significantly. However, they also stated that the SARIMA model may have problems accurately predicting severe rain speeds because it does not account for randomness. Their work only used historic monthly rainfall data and omitted other climate parameters (like temperature and relative humidity) from consideration in their models. Finally, they noted that if the missing value gap is sufficiently large, using Mean Substitutes could create bias in estimates generated from these datasets. The authors suggest that future research should explore Advanced Statistical Time-Series Modelling techniques like GARCH models and hybrid ARIMA/ML methods, and/or Deep Learning Models. In addition, opportunities to enhance forecast accuracy by incorporating Additional Weather Variables and/or Real-Time Weather Data could be examined [25].

Ayiah-Mensah, Bosson-Amedenu, Baah, and Addor (2025) developed a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) forecasting model to improve predictions of seasonal rainfall in the western part of Ghana. They performed a study of monthly rainfall from 2017 to 2023 using Augmented Dickey-Fuller and KPSS tests to determine whether time series data were stationary. Based on this analysis, the authors concluded that the best fit was SARIMA $(1,0,2)(2,0,0)_{12}$. Data corrections to eliminate the effects of outliers were made using the Interquartile Range (IQR) method and winsorization. Model accuracy was assessed using R-squared, Mean Absolute Percentage Error (MAPE), and Theil's U-statistic. The SARIMA model produced a high accuracy of seasonal rainfall forecast and outperformed other statistical-based methods. The limitations of this research are due to the total reliance on historical rainfall levels and the inability to incorporate any climate variable(s) such as temperature and/or moisture measure(s). Furthermore, the SARIMA model does not reflect either the long-term trends of climate change or the short-term structural changes in the data.

The research is also limited to a specific geo-location, thus limiting the external applicability of the results. Future directions for this work may use Machine Learning (ML) or hybrid models to increase the ability to predict nonlinear rainfall behaviour, incorporate climate change indicators into

the models, and expand the geographical region of the study to ensure greater reliability in forecasting results [26].

2.1. Problem Scope

The gaps that exist within the current knowledge base on rainfall pattern variability associated with the rains and the forecast of the extreme North-East Monsoon. Rainfall is still significant in accurately predicting the Regional Climate. Several earlier studies have adopted different cloud formation variability, climate modelling uncertainty, and machine learning techniques to select the feature for rainfall prediction, but do not fully accommodate prediction needs for Tamil Nadu Coastal Districts, considering the Short-Term and Location-specific nature of such rainfall prediction.

Traditional Climate models (5) do not provide adequate observational validation, while Artificial Intelligence-based climate models offer powerful techniques for developing complex, multivalued models, but, in general, have been developed without considering the seasonality of the various weather variables, dependency on lag time, or consideration of the influence brought about on one another by multi-correlated outputs Thus, research is needed in this existing gap, which would further develop a model that can generate accurate operational rainfall forecasts through identifying seasonal, trend, and multivariate relationships.

In this regard, the present research advocates the use of the Seasonal Autoregressive Integrated Moving Average with eXogenous variables framework to generate long-term and operational forecasts of extreme North-Eastern Monsoon Rains over the Coastal Tamil Nadu area. It utilises exogenously-derived weather predictors, seasonality, and enables the researcher to gain a better understanding of the possible influences affecting the rainfall prediction in that area.

2.2. Key Contribution

The proposed study comes up with a forecast framework that integrates SARIMAX in order to capture seasonal structures, lag dependencies, and other complex multivariate relationships of weather variables necessary for accurate forecasting of monsoon rainfall. The inclusion of many meteorological variables, namely temperature, dew point, visibility, wind speed, and pressure-related factors, as exogenous predictors, helped the model to identify their combined influence on extreme rainfall events over the North-East Monsoon. In addition, a broad performance comparison is carried out between the strengths and weaknesses of different techniques for monsoon prediction over Tamil Nadu. More importantly, the paper offers detailed insights into the evolving climatic pattern; trend analysis showed a rise in wind speed, increased minimum temperatures, and shifting visibility that altogether helped explain changes in the intensity of monsoons during recent decades.

3. Methods and Materials

The structure of data for this study is presented here: Table 1 consists of 14 columns, each of which represents a feature description. The dataset contains 2552 rows of daily weather records in a coastal station in Chennai on the Bay of Bengal in southern India, particularly for the North-East Monsoon. The data consists of numerical data (temperature, wind speed, and pressure) and categorical data used to analyse weather forecasting.

3.1. About the Dataset

Historical weather data for Chennai (12.994414,80.180517) from the year 2003 to 2023 have been procured for this study from the NOAA GSOD Version 7 dataset archived by the National Climatic Data Centre (NCDC), Federal Climate Complex, USA. The World Meteorological Organisation-endorsed dataset contains daily meteorological observations for Chennai, a major coastal station along the Bay of Bengal. Variables include temperature, dew point, sea-level pressure, visibility, wind speed, maximum gust, precipitation, and weather indicators such as fog, rain, snow, hail, thunder, and tornado flags.

The data originally reported in standard units (°F for temperature, knots for wind speed, and inches for precipitation) were converted to their respective values in Celsius, km/h, and millimetres. Only the months of the Northeast Monsoon-October, November, December, and January were extracted to create a refined dataset with 2552 daily observations and approximately 14 meteorological variables for further analysis. In that, refer to some of the references based on IMD [1-5].

3.2. Dataset Preprocessing

For this research, historical weather data for Chennai (12.994414, 80.180517) from 2003 to 2023 were obtained from the National Oceanic and Atmospheric Administration (NOAA) Global Surface Summary of Day (GSOD) Version 7 held by the National Climatic Data Centre (NCDC) at the Federal Climate Complex, which is located in the United States of America. This data set is endorsed by the World Meteorological Organisation (WMO) and consists of daily records for Chennai, which is a major coastal station located on the southeastern shore of India along the Bay of Bengal. The dataset has the following weather variables: temperature, dew point, sea-level pressure, visibility, wind speed, maximum wind gust, precipitation, fog, rain, snow, hail, thunder, and tornado indicators.

The data were immediately entered in standard units of reporting temperature in Fahrenheit, wind speed in knots, and precipitation in inches. These units were converted to Celsius, km/h, and mm, etc., accordingly for consistency. Principal Component Analysis (PCA) was performed to select significant feature variables, as well as manage missing and invalid data

(i.e., 9999.9 or 99.99). To consider periods focused on the Northeast (NE) Monsoon, only the months October, November, December, and January were selected. The data were later divided into 21 blocks, leading to 2552 rows and approximately 14 columns, each column representing a specific meteorological variable or derived feature. This formatted and cleaned set of data was effective in providing a framework for statistical time series forecasting with SARIMAX to investigate patterns of precipitation, wind speed, and visibility during the NE Monsoon.

Table 1. Dataset features description

S.no	Feature Name	Description	Unit	Data Type
1	TEMP_C	Mean Temperature	°C	Float
2	DEWP_C	Mean Dew Point	°C	Float
3	SLP	Mean Sea Level pressure	mb	Float
4	STP	Mean Station Pressure	mb	Float
5	VISIB	Mean Visibility	Km	Float
6	WDSP_kmph	Mean Wind Speed	Km/h	Float
7	MXSPD_kmph	Maximum Sustained Wind Speed	Km/h	Float
8	TEMP_C_CAT	Temperature Category	-	Object
9	DEWP_C_CAT	Dewpoint Category	-	Object
10	WDSP_KMPH_CAT	Windspeed Category	-	Object
11	VISIB_CAT	Visibility Category	-	Object
12	PRCP_mm	Precipitation Amount	mm	Float
13	THUNDER	Thunder Occurrences	Binary (0 / 1)	Integer
14	RainCategory	Categories of rainfall	-	Object

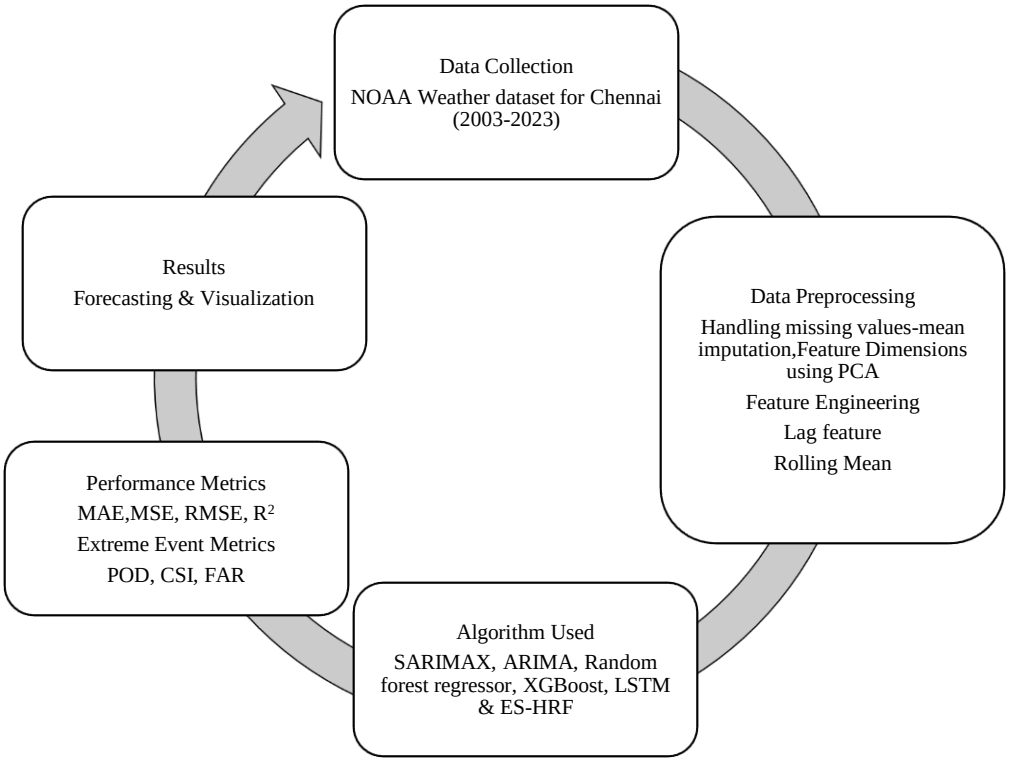

Fig. 1 Methodological framework for NE monsoon

Figure 1 illustrates the overall methodological workflow adopted in this study. It shows the sequence of steps starting from data collection and preprocessing, followed by model development, and evaluation.

The framework summarises how raw climate data is progressively transformed into a structured forecasting output, highlighting each major stage in the NE monsoon prediction process.

3.3. Data Engineering Workflow

The data engineering pipeline starts with the collection of twenty years (2003-2023) of daily meteorological data for the Chennai region from the India Meteorological Department.

The data was cleaned systematically by addressing missing values, inconsistent data, and by representing all meteorological variables as numerical variables. Seasonal division was done to separate the Northeast Monsoon season (October to January) for in-depth analysis. Feature engineering methods were used, which included lag features, rolling mean statistics, Fourier-based seasonal features, and quantile-based classification of rainfall and meteorological variables. An 80-20 train-test split was used to maintain temporal integrity.

3.4. Conventional Analysis

In the conventional analysis, both statistical and machine learning algorithms were employed to evaluate baseline rainfall prediction performance.

3.4.1. ARIMA Model

AutoRegressive Integrated Moving Average (ARIMA) is a statistical model used for time series forecasting, which combines three components: Autoregression (AR), differencing (I) to achieve stationarity, and Moving Average (MA). It predicts future values of a time series based on its own past values, past errors, and the number of differences applied.

The general ARIMA model is written as ARIMA (p, d, q), where: p = order of autoregression (number of lag observations included), d = degree of differencing (number of times the data is differenced to achieve stationarity), q = order of moving average (size of the moving average window). The Mathematical steps are as follows:

AR(p) Component

$$X_t = \phi_1 X_{\{t-1\}} + \phi_2 X_{\{t-2\}} + \cdots + \phi_p X_{\{t-p\}} + \varepsilon_t \quad (1)$$

MA(q) Component

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{\{t-1\}} + \theta_2 \varepsilon_{\{t-2\}} + \cdots + \theta_q \varepsilon_{\{t-q\}} \quad (2)$$

ARIMA (p, d, q) Model. After differencing d times to make the series stationary ($y_t = \nabla^d X_t$):

$$\nabla^d X_t = \phi_1 \nabla^d X_{\{t-1\}} + \cdots + \phi_p \nabla^d X_{\{t-p\}} + \varepsilon_t + \theta_1 \varepsilon_{\{t-1\}} + \cdots + \theta_q \varepsilon_{\{t-q\}} \quad (3)$$

Where:

$\nabla^d X_t = 1 - B^d X_t$ (difference operator, B is the backward shift operator)

ϕ_i = AR coefficients
θ_i = MA coefficients
ε_t = White noise (error term)
X_t = Observed value of the time series at time t
∇^d = Series differenced d times
$\varepsilon_{\{t-1\}}$ = Lagged error items
s = Seasonal period (e.g., $s = 4$ for monthly data)

3.4.2. SARIMAX Model

The Seasonal ARIMA with Exogenous variables (SARIMAX) extends ARIMA by incorporating seasonal terms and external predictors:

$$\Phi_P(L^s)\phi_p(L)(1-L)^d(1-L^s)^D\, y_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t + \beta X_t \quad (4)$$

Where:

y_t Is the observed value at time t,

ε_t is white noise,

t = Time index,

L is the lag operator,

P = Order of Seasonal AutoRegressive (SAR) component,
Q = Order of Seasonal Moving Average (SMA) component,

$\phi_p(L)$ and $\theta_q(L)$ Represent the non-seasonal AR and MA polynomials,

$(1-L)^d$ = Non-seasonal differencing term
$(1-L)^D$ = Seasonal differencing term

$\Phi_P(L^s)$ and $\Theta_Q(L^s)$ represent the seasonal AR and MA polynomials,

d and D are the orders of non-seasonal and seasonal differencing, respectively.

βX_t Denotes the influence of the exogenous regressors at time t.

SARIMAX improves seasonal representation but still assumes mostly linear relationships.

3.4.3. Random Forest Regressor

Random Forest Regressor prediction is the average of multiple decision trees:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i\ (x) \quad (5)$$

Where:

$T_i(x)$ = Prediction from the i-th tree
N = Number of trees

It captures nonlinear relationships but does not inherently model temporal structure.

3.4.4. XGBoost

XGBoost is a gradient boosting framework that minimizes a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k \quad (6)$$

Where:

l = Loss function
Ω = Regularization term
f_k = Decision trees

3.4.5. LSTM

The Long Short-Term Memory network models temporal sequences using gating mechanisms:

Forget Gate:

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \quad (7)$$

Input Gate:

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \quad (8)$$

Output Gate:

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \quad (9)$$

Where:

x_t is the input at time t
h_{t-1} is the previous hidden state
W and U are weight matrices
σ is the sigmoid activation function

3.5. Proposed Work

The Extreme-Sensitive Hybrid Residual Forecasting (ES-HRF) model combines seasonal statistical forecasting with nonlinear residual correction and extreme event sensitivity.

Unlike traditional models, the new model combines linear temporal patterns, weather factors, and nonlinear learning in a single framework.

The initial rainfall forecasting is done using a SARIMAX model with Fourier seasonal terms added to account for periodic monsoon patterns. The seasonal terms are expressed as:

$$\sin\left(\frac{2\pi t}{S}\right), \cos\left(\frac{2\pi t}{S}\right) \quad (10)$$

where S denotes the seasonal cycle. The base prediction is expressed as:

$$\hat{y}_t^{base}$$

To enhance sensitivity toward high-intensity rainfall, extreme thresholds are defined using quantile-based criteria:

$$SPIKE = 1 \text{ if } y_t \geq Q_{90}$$

$$EXTREME = 1 \text{ if } y_t \geq Q_{97} \quad (11)$$

Next, residual learning is applied to capture nonlinear deviations not explained by the statistical model. The residual is computed as:

$$R_t = y_t - \hat{y}_t^{base} \quad (12)$$

Gradient Boosting and Random Forest Regressors are used to model these residual nonlinearities. The final hybrid prediction is obtained as:

$$\hat{y}_t^{final} = \hat{y}_t^{base} + \hat{R}_t \quad (13)$$

This multi-stage hybrid architecture enables the model to retain seasonal interpretability while improving nonlinear and extreme rainfall representation. Overall, the proposed ES-HRF framework enhances predictive stability, improves extreme event detection capability, and provides a robust forecasting solution suitable for monsoon-dominated regions.

4. Results and Discussion

The climate parameters of Temperature (TEMP), Dew Point (DEWP), Visibility (VISIB), Wind Speed (WDSP), and Rain (PRCP) were characterised by a Low, Medium, and High categorisation based on the quantile ranges defined for these parameters in the provided table from the dataset. This categorisation enabled the examination of how Low, Medium, or High levels of each climate parameter were associated with Rainfall intensity during the Northeast monsoon months. The analysis showed that high temperature months predominantly corresponded to lower or medium intensity Rainfall. In contrast, low-visibility months were generally associated with heavily rainy months, supporting the assertion that heavy rain reduced visibility significantly. Furthermore, high dew point months related to medium to high intensity Rainfall corresponded to the naturally higher moisture content of the

atmosphere, and wind speed had moderate variation in predicting months with rainfall. The dimension of maintaining the categorisation of each of the climate parameters in this way allows for comparisons of the rainfall seasonality of climate parameters, weather pattern trends of climate, and parameters with rainfall, as well as possibly predicting severe weather events. Figure 2 shows a clear upward trend in temperature across the blocks. As temperatures rise, warm air can hold more moisture, increasing atmospheric humidity. Higher moisture content enhances the likelihood of cloud formation and rainfall. Thus, the rising temperature trend may indicate increased chances of rain in later blocks. Figure 3 shows a steady rise in dew point across the blocks. A higher dew point indicates higher atmospheric moisture. This increased moisture supports cloud formation and precipitation. Hence, the upward trend in dewpoint suggests an increased probability of rainfall in the later blocks. Table 7 provides a summary of long-spell rainfall events derived from the dataset that are increasing rapidly. It shows the number of consecutive days with reported rainfall monitored throughout the course of the monsoon season. As shown in Figure 4, wind speed shows a gradual increase. Stronger winds promote vertical lifting of moist air, leading to cloud formation and potential rainfall. Therefore, the increasing wind speed trend suggests higher rainfall chances in later blocks. Figure 5 shows a consistent decline in visibility from Block 1 to Block 21. Reduced visibility often indicates increased moisture, fog, or rainfall. This suggests that later blocks may experience wetter and more humid atmospheric conditions. As Figure 6 shows, although light rainfall remains relatively stable, heavy and very heavy rainfall events exhibit increasing peaks across blocks. These highlight the rising variability and intensity of rainfall during the NE monsoon.

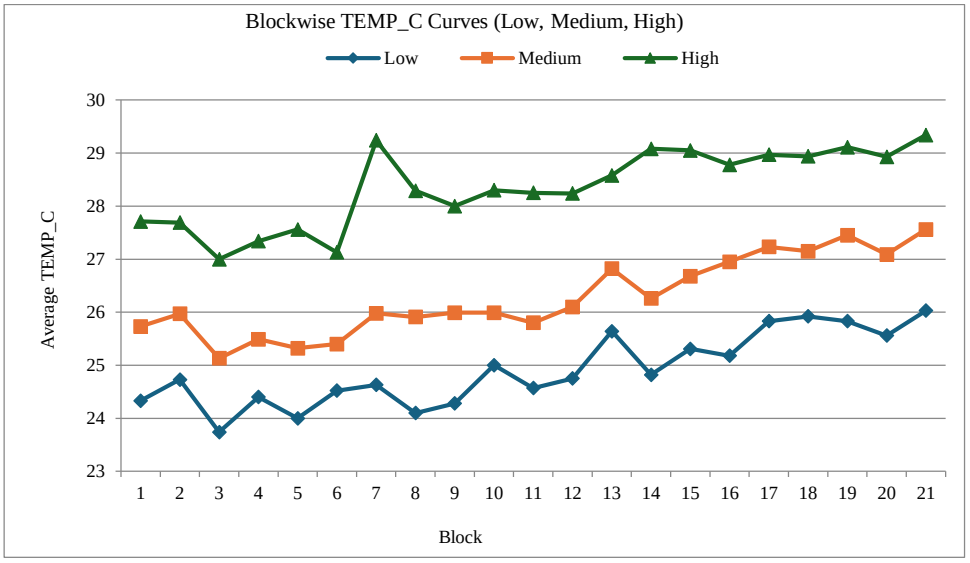

Fig. 2 Visualization of temperature classification

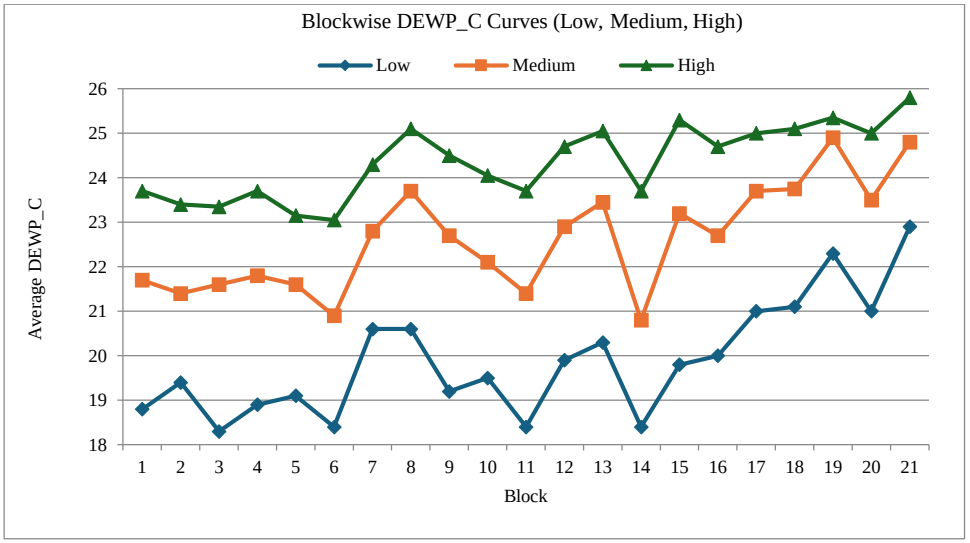

Fig. 3 Visualization of dew point classification

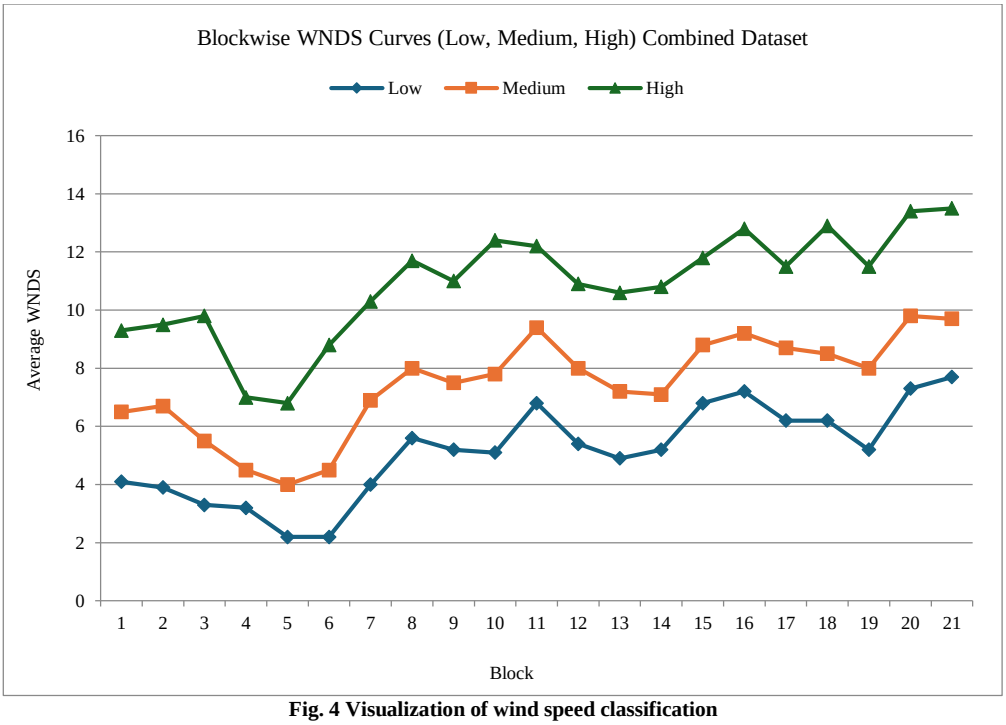

Fig. 4 Visualization of wind speed classification

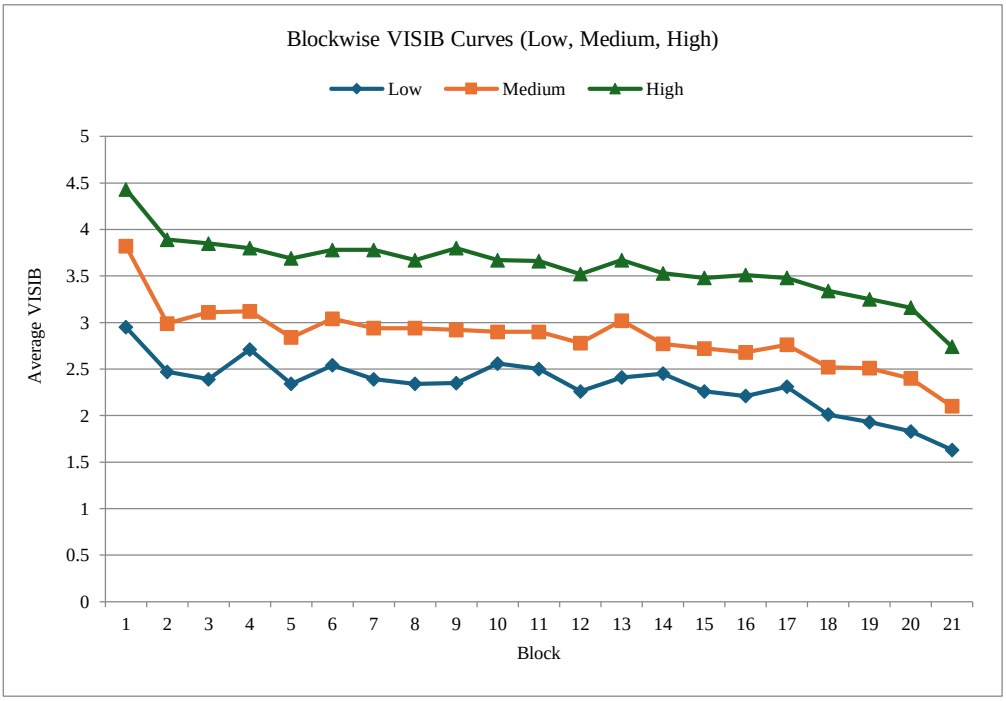

Fig. 5 Visualization of visibility during NE

Figure 7 shows a consistent increase in dewpoint, with forecasts indicating rising humidity. While this suggests moist air, it may not necessarily translate into increased rainfall.

Figure 8 indicates that high rainfall events show significant inter-year variability, with peaks in 2006, 2008, and 2016. Medium and low rainfall remain more stable.

This suggests an increasing frequency of extreme rainfall events. Figure 9 shows a gradual rise in temperature over the years, with forecasts indicating further warming in the next five years.

This warming may increase evaporation and reduce soil moisture.

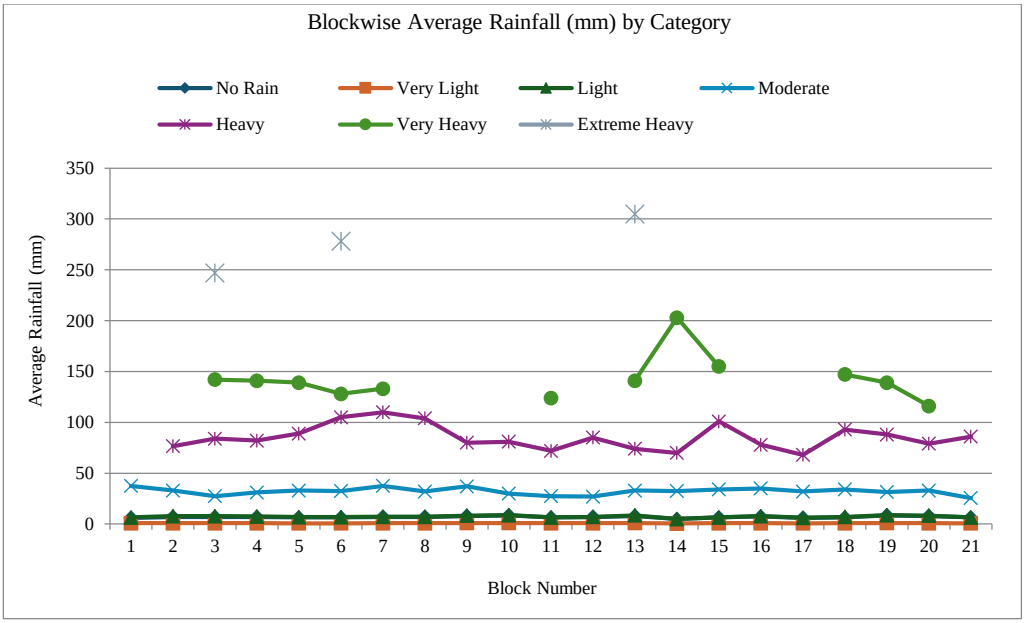

Fig. 6 Visualization of rainfall intensity for NE monsoon

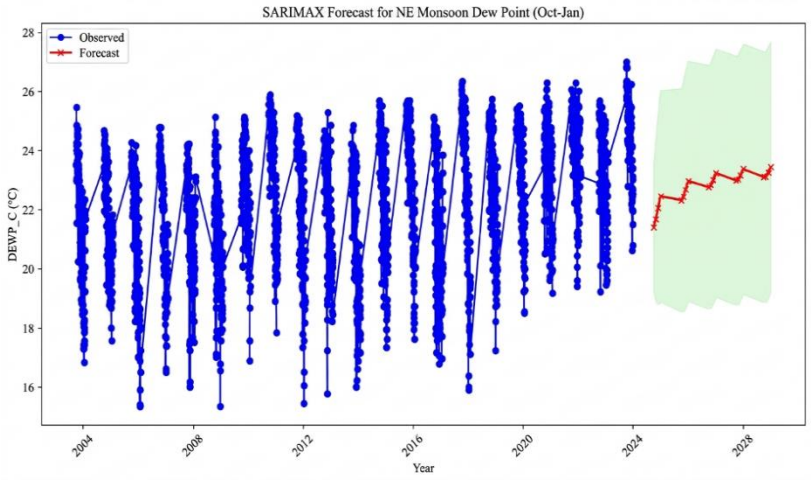

Fig. 7 SARIMAX forecast for NE monsoon dew point

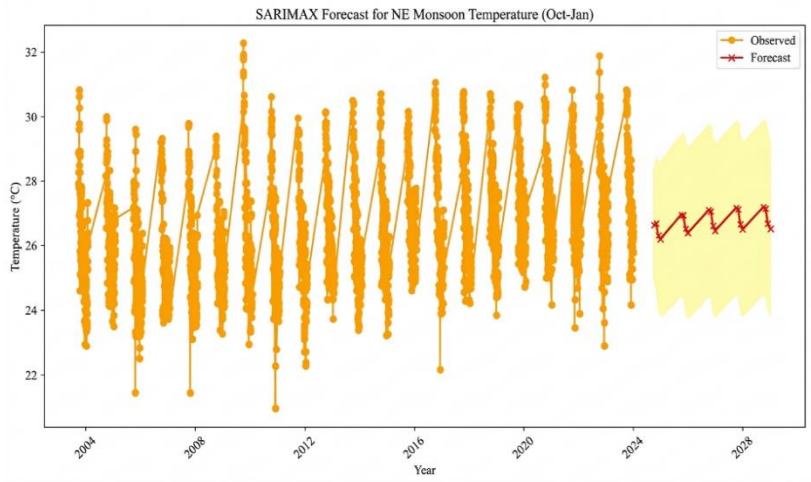

Fig. 8 Visualization of yearly rainfall range for NE monsoon

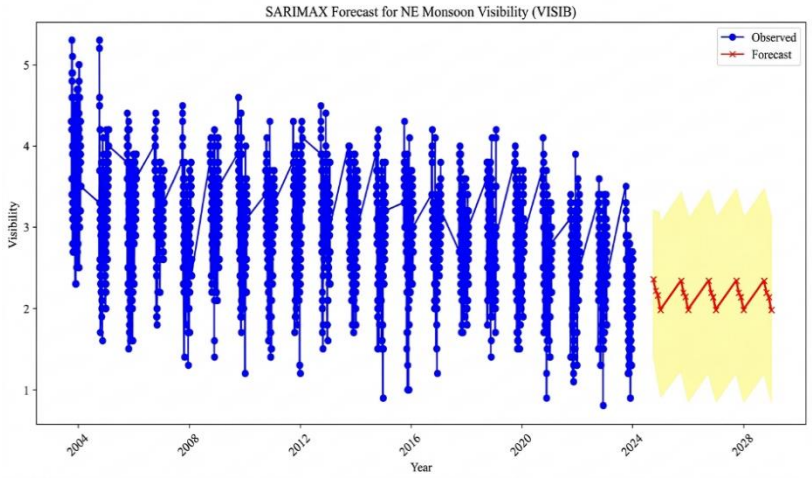

Fig. 9 SARIMAX forecast for NE monsoon temperature

Table 2. Temperature range distribution during NE monsoon

Year	Low	Medium	High
2003	22.89 - 25.11	25.11 - 26.28	26.28 - 30.83
2004	23.50 - 25.40	25.40 - 26.64	26.64 - 30.00
2005	21.44 - 24.44	24.44 - 25.72	25.72 - 29.61
2006	23.61 - 25.06	25.06 - 25.83	25.83 - 29.33
2007	21.44 - 24.61	24.61 - 26.11	26.11 - 29.78
2008	23.28 - 25.06	25.06 - 25.81	25.81 - 29.39
2009	22.94 - 25.28	25.28 - 26.95	26.95 - 32.28
2010	20.94 - 24.83	24.83 - 26.75	26.75 - 30.61
2011	22.28 - 25.40	25.40 - 26.56	26.56 - 29.94
2012	23.72 - 25.56	25.56 - 26.81	26.81 - 30.17
2013	23.39 - 25.24	25.24 - 26.56	26.56 - 30.50
2014	23.22 - 25.44	25.44 - 26.83	26.83 - 30.72
2015	24.61 - 26.28	26.28 - 27.33	27.33 - 30.17
2016	22.17 - 25.68	25.68 - 27.11	27.11 - 31.06
2017	24.22 - 26.00	26.00 - 27.39	27.39 - 30.78
2018	23.83 - 26.18	26.18 - 27.67	27.67 - 30.72
2019	24.72 - 26.61	26.61 - 27.75	27.75 - 30.39
2020	24.17 - 26.50	26.50 - 27.72	27.72 - 31.22
2021	23.44 - 26.78	26.78 - 27.97	27.97 - 30.83
2022	22.89 - 26.35	26.35 - 27.86	27.86 - 31.89
2023	24.17 - 26.89	26.89 - 28.23	28.23 - 30.83

Table 3. Dew point range distribution during NE monsoon

Year	Low	Medium	High
2003	16.83 -20.39	16.83 - 20.39	16.83 - 20.39
2004	17.56 -20.29	20.29 - 22.53	22.53 - 24.67
2005	15.33 -20.01	20.01 - 22.56	22.56 - 24.28
2006	16.5 - 20.11	20.11 - 23.06	23.06 - 24.78
2007	16.0 - 20.46	20.46 - 22.2	22.2 - 24.22
2008	15.33 -19.85	19.85 - 21.92	21.92 - 25.11
2009	16.89 -21.67	21.67 - 23.64	23.64 - 25.11
2010	17.83 -22.28	22.28 - 24.61	24.61 - 25.89
2011	15.44 -21.08	21.08 - 23.78	23.78 - 25.17
2012	15.78 - 20.7	20.7 - 23.22	23.22 - 25.28
2013	16.0 - 19.89	19.89 - 22.89	22.89 - 24.83

2014	17.33 -21.25	21.25 - 23.94	23.94 - 25.67
2015	17.61 -22.18	22.18 - 24.47	24.47 - 25.67
2016	16.78 -19.83	19.83 - 22.06	22.06 - 25.11
2017	15.89 -21.93	21.93 - 24.34	24.34 - 26.33
2018	17.22 -21.64	21.64 - 23.59	23.59 - 25.72
2019	18.5 - 22.24	22.24 - 24.56	24.56 - 25.5
2020	19.17 -22.61	22.61 - 24.39	24.39 - 26.28
2021	19.39 -23.79	23.79 - 24.94	24.94 - 26.28
2022	19.22 -22.57	22.57 - 24.47	24.47 - 25.67
2023	20.61 -24.17	24.17 - 25.17	25.17 - 27.0

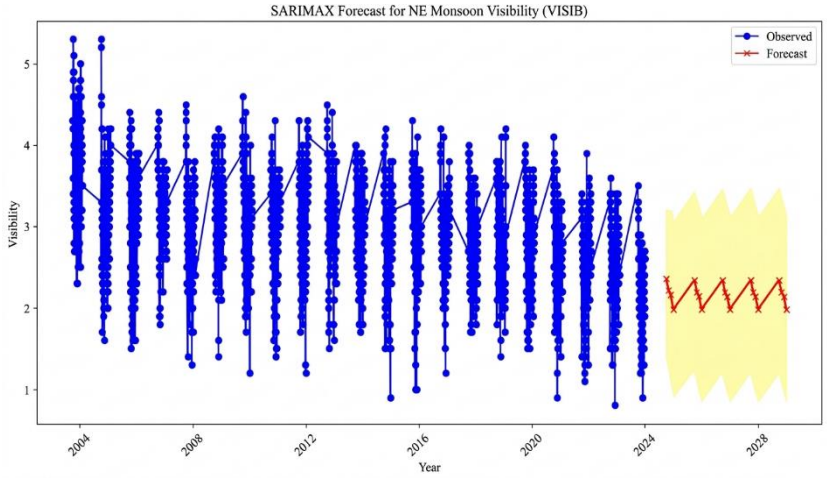

Fig. 10 SARIMAX forecast for NE monsoon visibility

Figure 10 shows a declining visibility trend, continuing into the forecast period. This suggests hazier or more humid conditions in future NE monsoons. As shown in the figure, the wind speed is fluctuating but gradually decreasing, with forecasts suggesting further weakening. Reduced wind may limit moisture movement and influence rainfall.

Table 4. Wind speed range distribution during the NE monsoon

Year	Low	Medium	High
2003	0.93 - 5.56	5.56 - 7.32	7.32 - 14.82
2004	0.0 - 5.74	5.74 - 7.87	7.87 - 13.89
2005	0.37 - 4.49	4.49 - 6.58	6.58 - 23.71
2006	0.93 - 3.7	3.7 - 5.0	5.0 - 10.93
2007	0.37 - 3.15	3.15 - 4.73	4.73 - 16.48
2008	0.0 - 3.33	3.33 - 5.19	5.19 - 26.67
2009	0.37 - 5.56	5.56 - 8.15	8.15 - 14.08
2010	2.59 - 6.85	6.85 - 9.26	9.26 - 18.71
2011	2.41 - 6.34	6.34 - 8.52	8.52 - 25.37
2012	0.93 - 6.34	6.34 - 9.26	9.26 - 31.3
2013	3.7 - 8.33	8.33 - 10.37	10.37 - 17.78
2014	0.37 - 6.9	6.9 - 8.7	8.7 - 16.11
2015	2.78 - 5.93	5.93 - 8.33	8.33 - 24.45
2016	2.78 - 5.93	5.93 - 7.96	7.96 - 40.74
2017	3.33 - 7.96	7.96 - 9.45	9.45 - 16.3
2018	5.0 - 8.33	8.33 - 10.0	10.0 - 19.63
2019	3.7 - 7.41	7.41 - 9.63	9.63 - 14.45
2020	3.15 - 7.22	7.22 - 9.45	9.45 - 40.37
2021	2.04 - 6.53	6.53 - 8.89	8.89 - 21.11
2022	4.82 - 8.7	8.7 - 10.37	10.37 - 28.34
2023	5.93 - 8.7	8.7 - 10.75	10.75 - 31.11

Table 5. Rainfall intensity classification for NE monsoon

Sno	Year	No Rain = 0	Very Light Rain 0.1<2.4	Light Rain 2.5 -15.5	Moderate Rain 15.6-64.4	Heavy Rain 64.5-115.5	Very Heavy Rain 115.6-204.4	Extreme Heavy Rain >=204.5
1	2003	86	18	12	7	0	0	0
2	2004	81	15	12	14	1	0	0
3	2005	71	8	15	21	4	2	2
4	2006	77	15	14	13	2	2	0
5	2007	87	12	15	7	1	1	0
6	2008	85	7	14	12	2	2	1
7	2009	80	13	14	13	2	1	0
8	2010	72	15	22	13	1	0	0
9	2011	84	7	12	16	4	0	0
10	2012	89	9	11	12	2	0	0
11	2013	83	14	17	7	1	1	0
12	2014	85	13	13	7	5	0	0
13	2015	70	11	17	19	1	3	2
14	2016	100	13	3	5	1	1	0
15	2017	81	14	12	13	1	2	0
16	2018	89	10	12	10	2	0	0
17	2019	80	9	16	16	2	0	0
18	2020	80	9	15	16	2	1	0
19	2021	65	14	15	22	4	3	0
20	2022	73	16	16	14	3	1	0
21	2023	50	16	15	8	3	0	0

Table 6. Visibility range distribution during NE monsoon

Year	Low	Medium	High
2003	2.3 - 3.5	3.5 - 4.0	4.0 - 5.3
2004	1.6 - 2.9	2.9 - 3.4	3.4 - 5.3
2005	1.5 - 3.0	3.0 - 3.5	3.5 - 4.4
2006	1.8 - 3.1	3.1 - 3.5	3.5 - 4.4
2007	1.3 - 2.7	2.7 - 3.2	3.2 - 4.5
2008	1.4 - 2.9	2.9 - 3.4	3.4 - 4.2
2009	1.2 - 2.8	2.8 - 3.4	3.4 - 4.6
2010	1.4 - 2.8	2.8 - 3.3	3.3 - 4.3
2011	1.2 - 2.7	2.7 - 3.4	3.4 - 4.3
2012	1.5 - 2.9	2.9 - 3.25	3.25 - 4.5
2013	1.7 - 2.9	2.9 - 3.2	3.2 - 4.0
2014	0.9 - 2.7	2.7 - 3.15	3.15 - 4.2
2015	1.0 - 2.9	2.9 - 3.4	3.4 - 4.3
2016	1.2 - 2.7	2.7 - 3.1	3.1 - 4.2
2017	1.7 - 2.6	2.6 - 3.1	3.1 - 4.0
2018	1.4 - 2.5	2.5 - 3.0	3.0 - 4.2
2019	1.5 - 2.7	2.7 - 3.0	3.0 - 4.0
2020	0.9 - 2.4	2.4 - 2.8	2.8 - 4.1
2021	1.1 - 2.4	2.4 - 2.8	2.8 - 3.9
2022	0.8 - 2.23	2.23 - 2.8	2.8 - 3.6
2023	0.9 - 2.0	2.0 - 2.4	2.4 - 3.5

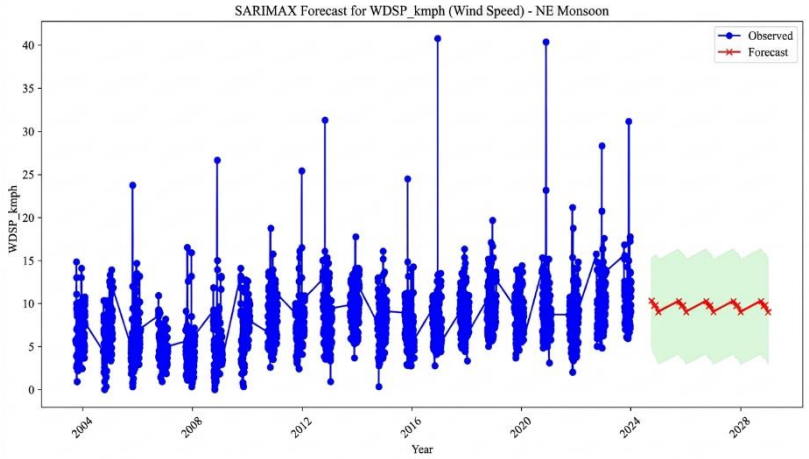

Fig. 11 SARIMAX forecast for NE monsoon wind speed

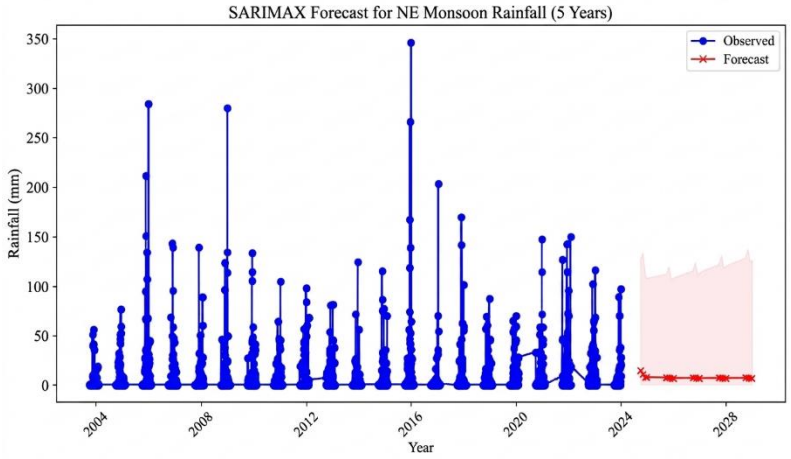

Fig. 12 SARIMAX forecast for NE monsoon PRCP

Table 7. Long spill rainfall for NE monsoon

Year	Count
2003	4
2004	1
2005	6
2006	3
2007	6
2008	3
2009	4
2010	5
2011	3
2012	2
2013	3
2014	5
2016	2
2017	3
2018	4
2019	5
2020	5
2021	5
2022	4
2023	5

Figure 12 indicates varying rainfall patterns across years, with significant peaks in 2015-2016. Forecasts suggest a reduction in rainfall in the coming years, implying a potential weakening of the NE monsoon. Rising temperatures and reduced wind speeds may contribute to this trend.

4.1. Performance Metrics Used in this Study of NE Monsoon

Mean Absolute Error (MAE): MAE measures the average magnitude of prediction errors without considering their direction.

$$MAE \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (14)$$

Where:

y_i = actual observed value
$\hat{y}_i$ = predicted value
n = number of observations

Mean Squared Error (MSE): MSE computes the average squared difference between actual and predicted values, giving larger errors more weight.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (15)$$

Where:
y_i = actual value
$\hat{y}_i$ = predicted value
n = number of observations

Root Mean Squared Error (RMSE): RMSE is the square root of MSE and expresses the error in the same units as the original data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (16)$$

Where:
y_i = actual observed value
$\hat{y}_i$ = predicted value
n = number of observations

Coefficient of Determination (R^2): R^2 measures the proportion of variance in the dependent variable explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (17)$$

Where:
y_i = actual observed value
$\hat{y}_i$ = predicted value
n = number of observations
$\bar{y}$ = mean of actual values

Figure 13 illustrates the comparison between observed and predicted daily rainfall using the proposed ES-HRF model. The predicted values closely follow the actual rainfall pattern, particularly capturing major extreme peaks, indicating strong agreement and effective extreme-event detection.

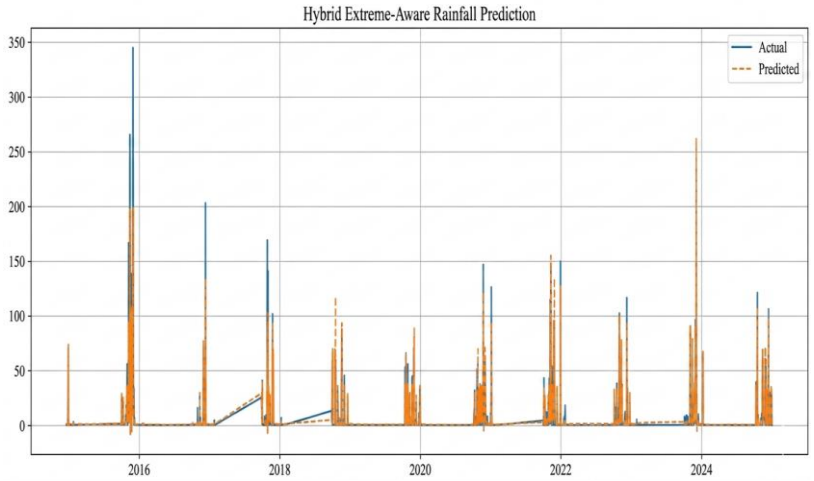

Fig. 13 Hybrid extreme- aware rainfall prediction

4.2. Extreme Event Detection Performance Analysis

Probability of Detection (POD): the fraction of observed extreme events that the model correctly predicted. A value close to 1 indicates high sensitivity.

$$POD = {}^{H}/_{H+M} \quad (18)$$

Where:
H = number of correctly predicted extreme events (Hits)
M = Number of missed extreme events (Misses)

False Alarm Ratio (FAR): FAR measures the proportion of predicted extreme events that did not actually occur. A value close to 0 indicates fewer false warnings.

$$FAR = {}^{F}/_{H+F} \quad (19)$$

Where:

F = Number of falsely predicted extreme events (False Alarms)

H = number of correctly predicted extreme events

Critical Success Index (CSI): CSI evaluates the overall skill of extreme event prediction by jointly considering hits, misses, and false alarms. Higher values indicate better prediction performance.

$$CSI = {}^{H}/_{H+M+F} \quad (20)$$

Where:

H = Number of correctly predicted extreme events (Hits)

M = Number of missed extreme events (Misses)

F = Number of falsely predicted extreme events (False Alarms)

The categorised ranges for the temperature, dew point, wind speed, and visibility during the North-East Monsoon Season from 2003 to 2023. Each parameter was classified into three categories with respect to year as low, medium, and high, for the purpose of examining trends and anomalies associated with seasonality and intensity.Table 2, 3, 4, and 6 present the quantile-based classification of temperature, dew point, wind speed, and visibility. Table 5 presents the observations of Precipitation (PRCP), which were binned into specific categories of no rain, very light rain, light rain, moderate rain, heavy rain, very heavy rain, and extreme heavy rain, for the likes of clarity and great detail as to how rainfall is distributed. After applying a quantile analysis to separate the parameters into low, medium, and high categories, it was possible to better interpret the variability the climatic parameters exhibited through an identified relationship attributable to consistent climatic patterns. For example, as the rainfall became heavier, the visibility remained lower, and the wind speed moderated, and during months of drier precipitation, higher temperatures ensued.

The SARIMAX models were evaluated based on diagnostic tests: Ljung-Box Q statistics for autocorrelation; Jarque-Bera tests for the normality of residuals; and tests for residuals, heteroskedasticity, skewness, and kurtosis. And also examined the model performance using MAE, MSE, R^2, and RMSE to support the reliability of forecast evaluations, but R^2 is low.

Table 8. Comparison of various algorithms

Algorithms	MAE	MSE	RMSE	R^2
ARIMA	11.840	401.330	20.033	0.0005
XGBOOST	6.4235	264.450	16.2619	0.27291
LSTM	9.1471	279.062	16.705	0.230
RANDOM FOREST REGRESSOR	11.8622	525.965	22.933	0.1917
SARIMAX	4.4006	200.970	14.176	0.5637
Proposed ES-HRF	3.542	97.080	9.853	0.8508

Table 9. Extreme rainfall forecast skill metrics

Extreme Event Metrics	Values
POD	0.994
FAR	0.019
CSI	0.975

Table 8 shows the comparative results of ARIMA, SARIMAX, XGBoost, LSTM, Random Forest Regressor, and the proposed ES-HRF model on rainfall forecasting. The result shows that ARIMA has a very poor forecasting ability (R^2 = 0.0005), whereas SARIMAX performs better (R^2 = 0.642) because of the inclusion of seasonal and exogenous features.

The machine learning models have a moderate level of accuracy but higher error values. On the other hand, the proposed ES-HRF model performs best, with RMSE = 9.853 and R^2 = 0.8508, indicating a substantial improvement in forecasting accuracy. Moreover, the performance of the proposed ES-HRF model for extreme rainfall event detection is shown in Table 9, with excellent detection capability (POD = 0.994, FAR = 0.019, and CSI = 0.975). The comparative study clearly shows that the traditional statistical model ARIMA is unsuited for modelling the nonlinear and seasonal variations for this monsoon rainfall, as evident from its poor performance in R^2. Although the SARIMAX model performs relatively better by taking into account the seasonal and exogenous factors of climate, its performance is still not satisfactory in dealing with the nonlinear and extreme rainfall patterns. The performance of individual machine learning models, XGBoost, LSTM, and Random Forest Regressor, is moderate in rainfall prediction, but they are not able to capture the temporal patterns and extreme rainfall events.

In contrast, the performance of the proposed Extreme-Sensitive Hybrid Residual Forecasting (ES-HRF) model is significantly better in rainfall prediction in terms of R^2 value and RMSE. Additionally, the extreme event detection analysis shows that the model performs exceptionally well in terms of POD, FAR, and CSI values. This clearly shows that not only

does the proposed hybrid model perform better in rainfall prediction, but it also performs well in extreme rainfall event detection, which is very important for flood preparedness and climate risk management.

4.3. Model Performance Justification

The reasons for the better performance of the proposed framework can be attributed to its hybrid model. The SARIMAX model captures the structured seasonality and exogenous influences of meteorological variables, while the Fourier components improve the periodic modelling. The extreme event detection based on quantiles improves the sensitivity of the model to extreme rainfall events. The residual nonlinearities that are not modeled using statistical models are efficiently modeled using ensemble models, thus improving the prediction and reducing the errors.

4.4. Practical Implications

The proposed framework for forecasting has numerous applications in regions where the monsoon is a dominant factor. It is possible to use the proposed framework to predict the intensity of rainfall and extreme events, which can be used for flood warning, drainage systems, and reservoir management. The extreme sensitivity of the proposed framework is also useful in regions such as Chennai, where the monsoon rains cause extreme disruptions to infrastructure and socio-economic activities due to the high intensity of rainfall.

5. Conclusion and Future Scope

The ES-HRF model has been introduced in this paper to improve the accuracy of monsoon rainfall forecasting. The proposed ES-HRF model combines seasonal patterns, weather regime clustering, spike-sensitive detection, and stacked residual learning to overcome the drawbacks of traditional rainfall forecasting methods. The comparison analysis shows that traditional statistical models, such as ARIMA and even more advanced models such as SARIMAX, are not capable of handling the nonlinear rainfall variability and extreme values. Conversely, machine learning models such as ARIMA-ML and SARIMAX-ML are also ineffective in showing consistent performance in extreme value detection. The proposed hybrid model integrates linear seasonal modelling and nonlinear residual correction and performs better than the existing models in terms of R^2 and error metrics.

Moreover, the extreme value analysis also confirms the high sensitivity and accuracy of the proposed model, which performs outstandingly well in extremely high-intensity rainfall event detection with a low false alarm rate. Future research can be done to further improve the framework by incorporating large-scale climate patterns such as ENSO, IOD, and Madden-Julian Oscillation indices. The addition of probabilistic forecasting techniques would also be beneficial for early warning systems. More advanced deep learning models, such as attention-based recurrent networks and transformers, can be used to further improve the learning of long-term temporal dependencies. The addition of spatial extension of the model to multi-station or gridded datasets would enable regional-scale rainfall forecasting.

The addition of real-time data assimilation and model updating techniques would also be beneficial for disaster management agencies.

Conflict of Interest

The authors disclose no conflicts of interest and agree to publish the research under academic ethics.

References

[1] Climate Data, Data and Graphs for Weather and Climate in Chennai, 2021. [Online]. Available: https://en.climate-data.org/asia/india/tamil-nadu/chennai-1003222/

[2] India Meteorological Department Ministry of Earth Sciences Government of India, Rainfall Information Meteorological Sub-divisions, 2026. [Online]. Available: https://mausam.imd.gov.in/responsive/rainfallinformation_msd.php

[3] Government of India Earth System Science Organisation Ministry of Earth Sciences India Meteorological Department, Report on Northeast Monsoon, 2024. [Online]. Available: https://mausam.imd.gov.in/chennai/mcdata/ne_monsoon_2024.pdf

[4] Government of India Earth System Science Organisation Ministry of Earth Sciences India Meteorological Department, Report on Northeast Monsoon, 2023. [Online]. Available: https://mausam.imd.gov.in/chennai/mcdata/ne_monsoon_2023.pdf

[5] SANDRP, South Asia Network on Dams, Rivers and People, SANDRP, 2025. [Online]. Available: https://sandrp.in/2025/10/01/sw-monsoon-2025-district-wise-rainfall-in-india/

[6] Amita Kumari et al., "Recent Decades have Witnessed a Strong East-West Gradient of Monsoon Precipitation Changes Over Northern India," *Atmospheric Research*, vol. 318, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Shulin Deng et al., "Rainfall Seasonality Changes and Underlying Climatic Causes in Global Land Monsoon Regions," *Atmospheric Research*, vol. 326, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shawna McKinley et al., "Mapping the Impact of Extreme Weather on Global Events and Mass Gatherings: Trends and Adaptive Strategies," *International Journal of Disaster Risk Reduction*, vol. 127, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rinu Fathima et al., "A Strong Influence of the Precession and Northern High Latitude Climate on the Monsoon Seasonality and Productivity in the Andaman Sea," *Global and Planetary Change*, vol. 254, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Nur Asyiqin Isa et al., "Impact of Ground-Level Cool Conditions During Monsoon Seasons: Occurrences of Isolated Breakdown Pulse Trains Discharges and their Relationship with Ground Flashes Events in Tropical Thunderstorms," *Electric Power Systems Research*, vol. 251, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ahmed M. Elshewey et al., "A Novel WD-SARIMAX Model for Temperature Forecasting using Daily Delhi Climate Dataset," *Sustainability*, vol. 15, no. 1, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Rudi Setyo Prihatin, and Eri Zuliarso, "Rainfall Prediction using the SARIMAX and LSTM Methods in Semarang City," *Jurnal Inovtek Polbeng - Informatics Series*, vol. 10, no. 3, pp. 1391-1401, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Md. Mehedi Hassan et al., "Machine Learning-based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness," *IEEE Access*, vol. 11, pp. 132196-132222, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Zhuoya Liu et al., "A Novel Linear Rainfall Forecast Model based on GNSS Observations and CAPE," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] A.M. Chacón-Maldonado et al., "Improving Monsoon Forecasting based on Feature Selection and Explainable Artificial Intelligence," *Applied Soft Computing*, vol. 185, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nagendra Prasad et al., "Late Holocene Vegetation History and Monsoonal Climate Change from the Core Monsoon Zone of India," *CATENA*, vol. 246, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Saloni Sharma et al., "Observational Evidence of Changing Cloud Macro-Physical Properties Under Warming Climate Over the Indian Summer Monsoon Region," *Science of the Total Environment*, vol. 947, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Kandula V. Subrahmanyam, M.V. Ramana, and Prakash Chauhan, "Long-Term Changes in Rainfall Epochs and Intensity Patterns of Indian Summer Monsoon in Changing Climate," *Atmospheric Research*, vol. 295, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] P. Umamaheswari, and V. Ramaswamy, "An Integrated Framework for Rainfall Prediction and Analysis using a Stacked Heterogeneous Ensemble Model (SHEM)," *Expert Systems with Applications*, vol. 256, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Umamaheswari P, and V. Ramaswamy, "Optimized Preprocessing using time Variant Particle Swarm Optimization (TVPSO) and Deep Learning on Rainfall Data," *Journal of Scientific and Industrial Research*, vol. 81, pp. 1317-1325, 2022. [CrossRef] [Google Scholar]

[21] D. Ruhiat et al., "Ten Daily Rainfall Forecasting using SSA Algorithms and Seasonal Arima Model to Determine the Beginning of the Rainy Season," *IOP Conference Series: Earth and Environmental Science*, vol. 1314, no. 1, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Amierul Aiman Hafidz Bin Muzaffar, R Kanesaraj Ramasamy, and Venushini Rajendran, "A Lightweight IoT-Enabled Agricultural Monitoring System with Arima-based Rainfall Forecasting for Precision Irrigation," *SSRN*, pp. 1-42, 2025. [Google Scholar] [Publisher Link]

[23] Huey Yin Tee, and Rosnalini Mansor, "Forecasting Rainfall Volume in Selangor with a Combined ARIMA Model," *Journal of Computational Innovation and Analytics (JCIA)*, vol. 3, no. 1, pp. 83-103, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Wenchuan Wang et al., "A Comparison of BPNN, GMDH, and ARIMA for Monthly Rainfall Forecasting based on Wavelet Packet Decomposition," *Water*, vol. 13, no. 20, pp. 1-24, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Paramasivan Arumugam, and R. Saranya, "Outlier Detection and Missing Value in Seasonal ARIMA Model using Rainfall Data," *Materials Today: Proceedings*, vol. 5, no. 1, pp. 1791-1799, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Francis Ayiah-Mensah et al., "Advancements in Seasonal Rainfall Forecasting: A Seasonal Auto-Regressive Integrated Moving Average Model with Outlier Adjustments for Ghana's Western Region," *Scientific African*, vol. 28, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]