MAUSAM, **77**, 2 (April 2026), 485-502

MAUSAM

DOI: https://doi.org/10.54302/mausam.v77i2.6951
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No.551.553.21:551.577.13:004.8

# Analysis of summer monsoon rainfall: Sustainability of long-term modelling by machine learning methods

NAMITA GOYAL[1], APARNA N MAHAJAN[1] and K.C. TRIPATHI[2*]

[1] *Department of CSE, Maharaja Agrasen University, Baddi, Himachal Pradesh - 174103, India*

[2] *Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi – 110086, India*

(*Received 17 January 2025, Accepted 19 September 2025*)

***Corresponding author's email: kctripathi@mait.ac.in**

**सार** – भारतीय मानसून भारतीय उपमहाद्वीप में अत्यंत महत्वपूर्ण सामाजिक-आर्थिक महत्व की एक विशिष्ट मौसम संबंधी घटना है। एक सदी से अधिक समय तक फैले उपलब्ध भारतीय और क्षेत्रीय डेटासेट के विश्लेषण से वर्षा की परिवर्तनशीलता के बारे में महत्वपूर्ण जानकारी मिलती है। भारतीय ग्रीष्मकालीन मानसून वर्षा (आईएसएमआर) और केरल ग्रीष्मकालीन मानसून वर्षा (केएसएमआर) के लिए दो प्रमुख विशेषताओं - दीर्घकालिक प्रवृत्ति और परिवर्तन बिंदु - का विश्लेषण किया गया है। केरल को इसलिए चुना गया है क्योंकि यह भारत में मानसून के आरंभ का प्रतीक है। आईएसएमआर और केएसएमआर दोनों के लिए 95% विश्वास स्तर पर कोई प्रवृत्ति नहीं होने की शून्य परिकल्पना के तहत, गैर-सामान्य डेटा में एकसमान प्रवृत्तियों का पता लगाने के लिए उपयुक्त गैर-पैरामीट्रिक मान-केंडल परीक्षण का उपयोग करके प्रवृत्ति विश्लेषण किया गया। वर्षा पैटर्न में बदलाव दर्शाने वाले परिवर्तन बिंदुओं की पहचान बायेसियन परिवर्तन बिंदु पहचान विधि का उपयोग करके की गई। मौसमी वर्षा का पूर्वानुमान योजना और शासन के लिए महत्वपूर्ण है, और मशीन लर्निंग मॉडल सटीक मौसम संबंधी पूर्वानुमान के लिए सिद्ध उपकरण हैं। हालांकि, क्या ऐसे पूर्वानुमान दीर्घकालिक सांख्यिकीय विशेषताओं को बनाए रख सकते हैं, यह अभी भी अनसुलझा है। इस अध्ययन में, हमने यह आकलन करने के लिए मशीन लर्निंग मॉडल विकसित किए कि क्या पूर्वानुमान परिवर्तन बिंदु अंतरालों में देखी गई प्रवृत्तियों और मापदंडों को संरक्षित करते हैं। सर्वश्रेष्ठ प्रदर्शन करने वाले LSTM मॉडल (लैग = 12) ने 147 (प्रशिक्षण) और 167 (परीक्षण) का RMSE प्राप्त किया, जिसमें सहसंबंध $\geq$ 0.85 था। इसने 5 में से 4 प्रेक्षित ब्रेक पॉइंट्स को पुनः प्रस्तुत किया और मैन-केंडल प्रवृत्ति परिणामों से मेल खाया। निष्कर्ष दर्शाते हैं कि ML-आधारित दीर्घकालिक मौसमी मानसून मॉडलिंग बहु-दशकीय अवधियों में प्रवृत्तियों, ब्रेक पॉइंट्स और सांख्यिकीय मापदंडों को बनाए रख सकती है।

**ABSTRACT.** The Indian monsoon is a distinct meteorological phenomenon of vital socio-economic importance in the Indian subcontinent. Analysis of available Indian and regional datasets spanning over a century provides critical insights into rainfall variability. Two major characteristics - long-term trend and break points - have been analysed for Indian summer monsoon rainfall (ISMR) and Kerala summer monsoon rainfall (KSMR). Kerala is selected as it marks the onset of monsoon in India. Trend analysis was performed using the non-parametric Mann–Kendall test, suited for detecting monotonic trends in non-normal data, under the null hypothesis of no trend at the 95% confidence level for both ISMR and KSMR. Break points, indicating shifts in rainfall regimes, were identified using a Bayesian change point detection method. Prediction of seasonal rain is vital for planning and governance, and machine learning models are proven tools for accurate meteorological prediction. However, whether such predictions can sustain long-term statistical characteristics remains underexplored. In this study, we developed ML models to assess if predictions preserve the observed trends and parameters across break point intervals. The best-performing LSTM model (lag = 12) achieved RMSE of 147 (train) and 167 (test) with correlations $\geq$ 0.85, reproducing 4 of 5 observed break points and matching Mann–Kendall trend results. The findings demonstrate that ML-based long-term seasonal monsoon modelling can retain trends, break points, and statistical parameters over multi-decadal horizons.

**Key words –** Autocorrelations, Mann-Kendall test, Bayesian change point test, Break points.

## 1. Introduction

Rainfall is a crucial aspect of ecosystem. Humans, animals, birds, plants, in fact the complete biodiversity is dependent on rain for its survival. Growth of many industries like agriculture sector, fertilizers, agrochemicals, automobiles, and powerhouse etc. directly or indirectly relies on rain. Thus, even a small variation in rainfall can cause a huge impact on Indian economy. Henceforth, it is necessary to understand and foresee

monsoon trend, onset of monsoon, how long it will last, and its intensity to have beforehand planning for managing water resources, finances, power, agriculture, and daily livelihood concerns faced by the common citizens. Consequently, this subject is utmost crucial and has been the focus of the whole research community for generations. Approaches to weather forecasting (Goyal *et al*., 2023) can broadly be classified into numerical methods (Palmer *et al*., 2004; Adcroft *et al*., 2004; Mellor *et al*., 1998; Pacanowski *et al*., 1993) and statistical methods (Tripathi *et al*., 2006; Shukla *et al*., 2011; Praveen *et al*., 2020; Dash *et al*., 2019; Rajan & Desamsetti, 2021; Liyew *et al*., 2021). Numerical methods involve mathematical equation representing the complex behaviour of atmosphere and ocean. Statistical methods, on the other hand, involve the study and analysis of huge dataset using machine learning techniques to infer the future outcome. Both approaches follow different methods to foresee climatic conditions such as long-range forecast (Munot *et al*., 2007; Mitsui *et al*.,2021) which involves predicting weather and climate conditions that are several weeks to months in advance. Medium-range forecasting (Prakash *et al*., 2016) covers a period of several days to many weeks to monitor and predict the behaviour of the monsoon and short-range forecast (Kumar *et al*.,2022; Ashok *et al*.,2022) spans the time from next few hours to 2-3 days.

Many hybrid and ensemble models have been explored on basis of above-mentioned methods in various researches for forecasting weather. But still more thorough historical and paleoclimate records are required to comprehend long-term trends and inherent variability in monsoon patterns of climatic data, which will aid in the analysis and comprehension of monsoon drift. An important aspect of weather is the Indian monsoon which has attracted the attention of meteorologists across the world (Gadgil *et al*., 2003; Saha *et al*., 2017; Saha *et al*.,2021; Kumar & Singh, 2021). When investigating the Indian rainfall researchers are interested in two broad aspects- the seasonal prediction of the monsoon rain (long range forecast) and short-range forecasts. The spatial resolution ranges from All India Rainfall (AIR) to regional rainfall (north-west, north-east, central, southern peninsula) and 36 meteorological subdivisions of India (Kelkar *et al*, 2020).

The objective of this paper is to study the monsoon patterns (i.e. seasonal rainfall recorded in the months of June-July-August-September abbreviated as JJAS) of all India and all 36 subdivisions of India in last 122 years with a special focus on Kerala. The seasonal rainfall recorded in JJAS period over the All India is termed as Indian Summer Monsson Rainfall (ISMR). Hereafter the seasonal JJAS rainfall that occurs across Kerala is referred to as Kerala Summer Monsoon Rainfall (KSMR).

The study aims to find (i) any trend over time in last 122 years in ISMR and KSMR, (ii) the break points in the time series of ISMR and KSMR and (iii) if statistical models can reconstruct this deviation in trends and break points.

The statistical models developed for the above are machine learning models that employ state of the art learning algorithms for prediction and classification (Dangeti, 2017). These include Random Forest, Support Vector Machine, Recurrent Neural Network and Long Short-Term Memory. Model development to achieve accurate prediction is just a step, real objective is to ensure sustainability so that the models remain reliable and robust over extended time periods, demonstrating stable performance across multiple years and under varying climatic conditions. A seemingly good model with low RMSE may, in fact, fail to retain the quality of predictions when long term trends are considered. Such models may be “accurate” but not “sustainable”. This is explored in this study in context of ML models.

The remaining structure of the paper is organised as follows: section 2, summarizes the related work in monsoon prediction studies. Section 3 explains the dataset used for this work. Section 4, presents the methodology of the proposed work. Section 5 examines the results and finally paper Section 6 concludes the paper.

### 1.1 *Literature Review*

Extensive research has focused on developing models to simulate and predict the complex behaviour of the Indian summer monsoon under varying climatic influences Conventional approaches such as regression analysis and empirical orthogonal functions have laid the groundwork for understanding monsoon behaviour, yet they often struggle to effectively represent the system’s non-linear characteristics and rare extreme events. In recent developments, machine learning and deep learning techniques have shown considerable promise in enhancing prediction accuracy by utilizing vast datasets and uncovering intricate patterns among climatic variables. A comprehensive review of recent literature, summarized in Table S1 (supplementary), outlines the methodologies, datasets used, performance indicators, and key findings across various studies. Despite these advancements, concerns persist regarding the reliability and sustainability of these predictive models-particularly in the context of evolving climate conditions. Ensuring long-term effectiveness requires models that are not only accurate but also stable, adaptable, and transparent, to support dependable decision-making in diverse and dynamic scenarios.

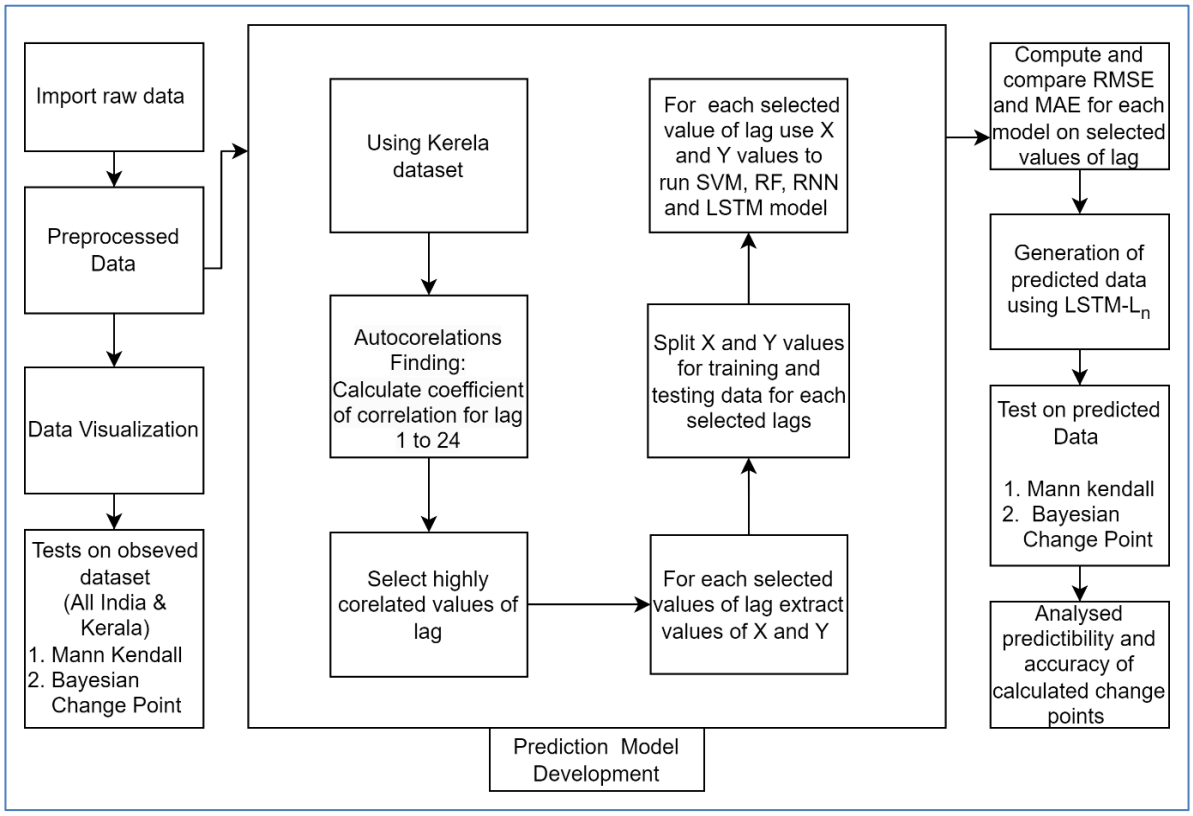


**Fig.1** Proposed Methodology Framework

## 2. Data and methodology

### 2.1 *Data*

The rainfall time series is taken from India Meteorology Department (IMD) (https://imdpune.gov.in). The website contains comprehensive data sets at various spatial and temporal resolutions. For the present study we have used the 122 years (1901-2022) monthly record of subdivisional rainfalls. The subdivisions have been defined in section 1. There are multiple observatories (rain gauge stations) that give daily, monthly, or seasonal rainfall measurements. The following methodology of calculating rainfall from rain stations is used in preparing the data (https://www.imdpune.gov.in, https://data.gov.in):

Monthly rainfall: Sum of daily rainfalls observed at a station for all the days of the month.

District series: Simple arithmetic mean of the monthly rainfall data of all the available stations in the district for a particular month.

Subdivision and state series: Calculated by district area weighted average method.

All India series (AIR): Calculated by subdivision area weighted average method.

The weights in subdivisional and AIR are proportional to the geographical area.

$$R_{All\ India} = \sum_{s=1}^{36} \beta_s R_s \quad (1)$$

Where $R_{All\ India}$: All India rainfall, $\beta_s$: weight assigned to subdivision 's', $R_s$: Rainfall of subdivision 's'

This research work presents a holistic framework to identify drift in Indian monsoon using rainfall data of all 36 subdivisions of India. Monsoon patterns of the subdivisions have been investigated for significant trend, shifts or notable changes. An attempt has been made to construct machine learning models for long-range seasonal forecasting of Kerala subdivisions. The framework is presented in fig. 1.

In the sub sections below, we have discussed the methodology in the light of the framework depicted above.

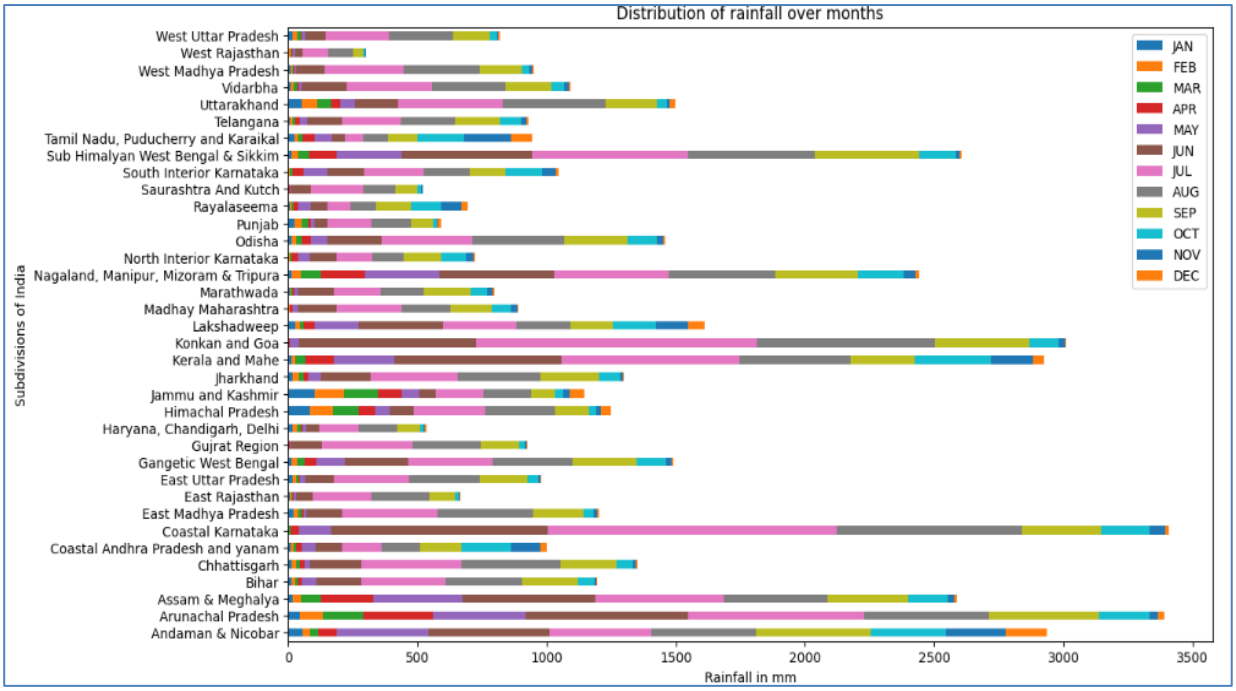


**Fig. 2.** Average monthly rainfall of 36 subdivisions of India

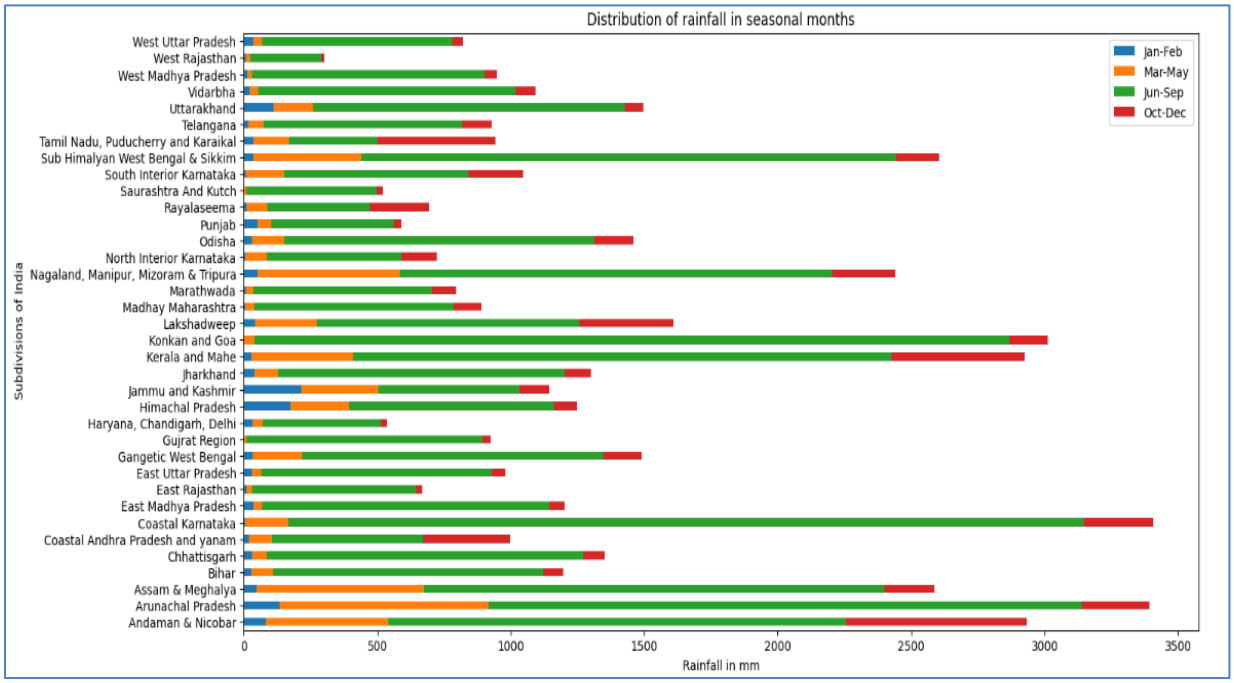


**Fig. 3.** Average seasonal rainfall of 36 subdivisions of India

### 2.1.1 *Raw data*

As listed in section 3, the raw data is taken from the IMD website (https://imdpune.gov.in) which is freely available.

### 2.1.2 *Data Preprocessing*

With daily, monthly, and annual rainfall levels recorded, the data is rich in diversity and detail. Kerala and Mahe's data have been extracted from the dataset of 36 subdivisions.

#### 2.1.3 *Data Visualization*

Data has been visualized from different perspectives to understand seasonality and trend. Fig. 2 depicts average monthly rainfall distribution of 36 subdivisions of India. It is evident that the observations are dominated by precipitation in June, July, August, and September. Seasonal rainfall is calculated by simply adding up the rainfall in each of the months that contribute to the season at the specific geographical area. Fig. 3 represents average seasonal rainfall of the subdivisions. The monsoon season (JJAS) depicts highest records.

#### 2.1.4 *Tests on observed dataset (identifying trends and change points)*

##### 2.1.4.1 *Trend in the series*

To quantify the trend in the time series, Mann Kendal (MK) (Sudarsan & Lasitha, 2023) test was performed. This test is preferred in the current scenario over the regression line trend for few reasons- (i) MK test is non-parametric that does not assume any specific distribution or form, (ii) it uses rank correlation to assess the direction and significance, (iii) it is better suited for non-linear data (iv) outliers are better treated, (v) sensitive to order rather than magnitude and (vi) trend direction and strength are better quantified. At a later stage when we build model of prediction, the tests and prediction results are analysed. The following assumptions are used:

(i) Observations are independent, meaning thereby that observation recorded at any point in time has no bearing on the observation recorded at subsequent times. It does not mean the time series is independent.

(ii) The nature of trend is not assumed.

Further, we have used the JJAS seasonal rainfall and not the entire time series. The reason for this is there is a large variance among the rainfalls of individual months. The JJAS accounts for about 70% of total annual precipitation. Hence the data is largely biased towards JJAS. Keeping only the seasonal data maintains uniformity in the time series.

The argument in the MK test in the present case is all about existence of a positive trend, negative trend, or no trend. A brief description of is presented below:

$$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} sgn(x_j - x_i) \tag{2}$$

where the inner function is defined as:

$$sgn(x_j - x_i) = \begin{cases} 1, & x_j > x_i \\ -1, & x_j < x_i \\ 0, & x_j = x_i \end{cases} \tag{3}$$

The MK test has been applied on the Indian summer monsoon rainfall (ISMR) data as well as the Kerala summer monsoon rainfall (KSMR). The reason for choosing Kerala is that the summer monsoon strikes the Indian coast from Kerala and any major deviation in onset of KSMR can lead to significant deviations in other sub-divisions.

##### 2.1.4.2 *Break points in the series*

Change points (Gallagher *et al*., 2013), also called break points or structural breaks, reflect the points in time where in vital statistics of the time series, such as the mean, variance, or trends, undergo significant changes. In the context of Indian monsoon change points may indicate a shift in climate pattern (like monsoon onset or retreat). Detection of change points is crucial for understanding shifts in monsoon drift and rain patterns. These points may reflect on the future behaviour of the parameter. Sudden changes in the future behaviour during the prediction process may render the prediction inaccurate, if there are change points in the near future. Knowledge of the future behaviour of the system based on change points coupled with the estimation of the future points by statistical or numerical models may give accurate picture of the future behaviour.

Consider a time series $X = \{X_t : 1 \leq t \leq T\}$. A change point is a temporal point '$\tau$' such that some statistical property of $X$ change after $\tau$ i.e.

$$X_t = \begin{cases} f_1(\boldsymbol{\theta_1}), t = 1,2, \dots, \tau \\ f_2(\boldsymbol{\theta_2}), t = \tau + 1, \tau + 2, \dots T \end{cases} \tag{4}$$

where $f_1$ and $f_2$ are the governing distributions before and after the change point $\tau$. $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are the sets of parameters of the distributions. Usually, $f_1$ and $f_2$ are same and only the parameters may change. Further, for practical purposes we may assume $f_1$ and $f_2$ to be normal distributions and $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ to be mean and standard deviations i.e. $\boldsymbol{\theta_1} = (\mu_1, \sigma_1)$ and $\boldsymbol{\theta_2} = (\mu_2, \sigma_2)$ **where** $\mu_1$and $\mu_2$ are means and $\sigma_1$ and $\sigma_2$ are standard deviations. We have used Bayesian method to detect break points in the ISMR and KSMR time series.

#### 2.1.5 *Autocorrelation and lag correlation analysis*

Autocorrelation (Tabari *et al*., 2011) measures the

**Table 1**

**Key hyperparameters used in the training of the ML models used**

| Model | Hyperparameter | Value | Function |
|---|---|---|---|
| **SVM** | Kernel | 'rbf' | Captures non -linear patterns in data |
| **Random Forest** | n_estimators | 100 | Specifies number of decision trees in the ensemble. |
| | random_state | 42 | Ensures reproducibility of results |
| **RNN** | units (SimpleRNN) | 64 | Number of units in the SimpleRNN layer to capture temporal dependencies |
| | input_shape | (1, 1) | for one time step and one feature |
| | optimizer | 'adam' | for adaptive learning |
| | loss | 'mean_squared_error' | for regression |
| | epochs | 1000 | Number of training iterations |
| | batch_size | 1 | Training with one sample at a time. |
| **LSTM** | units (LSTM) | 64 | Number of units in the Simple RNN layer to capture temporal dependencies |
| | input_shape | (1, 1) | for one time step and one feature |
| | optimizer | 'adam' | for adaptive learning |
| | loss | 'mean_squared_error' | for regression |
| | epochs | 1000 | Number of training iterations |
| | batch_size | 1 | Training with one sample at a time. |

strength of association between $\{X_i: 1 \leq i \leq n\}$ and $\{X_{i+l}: 1 \leq l \leq \tau;\ 1 \leq i \leq n+l\}$, the length of the sequence being $L = n + \tau$. Utilizing autocorrelation can help predictive model to learn more about temporal dependencies required to decide predictors when modelling a time series. Autocorrelation is a widely accepted and interpretable method that allows us to detect repeated patterns or persistence over time, which is particularly relevant for data like rainfall that often exhibit seasonal or cyclical trends. Other methods over autocorrelation could have also be chosen but fig. 8 in section 5 shows significant correlation values like 0.78, 0.77, 0.88, 0.77, 0.76 and 0.87 at lag 1, 11, 12, 13, 23 and 24 respectively which makes it sufficient to determine lag predictors.

As pointed out earlier, considering the importance of Kerala and Mahe in the Indian monsoon, the model will be built using the Kerala and Mahe dataset. Therefore, the autocorrelations have been computed exclusively for the Kerala and Mahe subdivision.

#### 2.1.6 *Prediction models*

We have developed four machine learning models for long term modelling of the rainfall- Random Forest (RF) (Uddin *et al*., 2022), Support Vector Machine (SVM) (Hayaty *et al*., 2023), Recurrent Neural Network (RNN) (Kang *et al*., 2020) and Long Short-Term Memory (LSTM) (Poornima & Pushpalatha, 2019).

The key hyperparameters used in the training of above-mentioned models has been briefed in table 1. The selected models reflect both conventional machine learning (RF, SVM) and deep learning (RNN, LSTM) methodologies. This makes sure that methods with various strengths like interpretability (RF), flexibility (SVM), and the capacity to describe sequential dependencies (RNN, LSTM) are thoroughly evaluated.

Previous studies pertaining to time-series forecasting and rainfall prediction have made substantial use of these models. They are deemed suitable for this domain because of their well-documented performance in the literature. As discussed in section 4.5 autocorrelation analysis has been done to select the predictors.

The data set is prepared as $D = \{(x_t, x_{t+l}) | l = 1,11,12,13,23,24\}$ where the first element is input and the second target. After training, models have been compared with respect to two performance metrics Root Mean Square Error and Mean Absolute Error.

## 3. Results and discussion

### 3.1 *MK test applied on Indian summer monsoon rainfall (ISMR)*

Z-score is calculated by standard statistical methods (Meyer, 1965). The null hypothesis $H_o$ is taken to mean "the time series doesn't display any trend" which may be construed to mean "the data is randomly ordered".

The critical Z-values are approximately 1.96 for 95% confidence and 2.58 for 99% confidence.

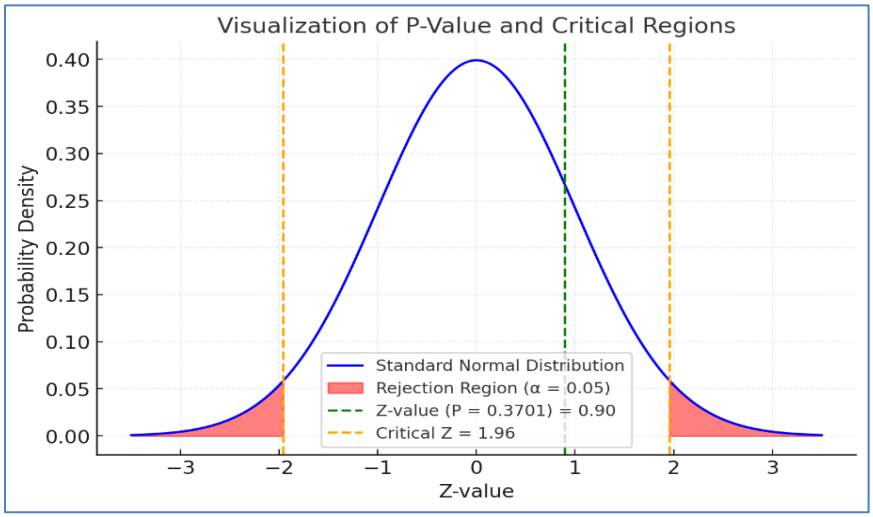


**Fig. 4**. Two tailed test result on ISMR data set. The p-value (0.3701) lies within the acceptance region

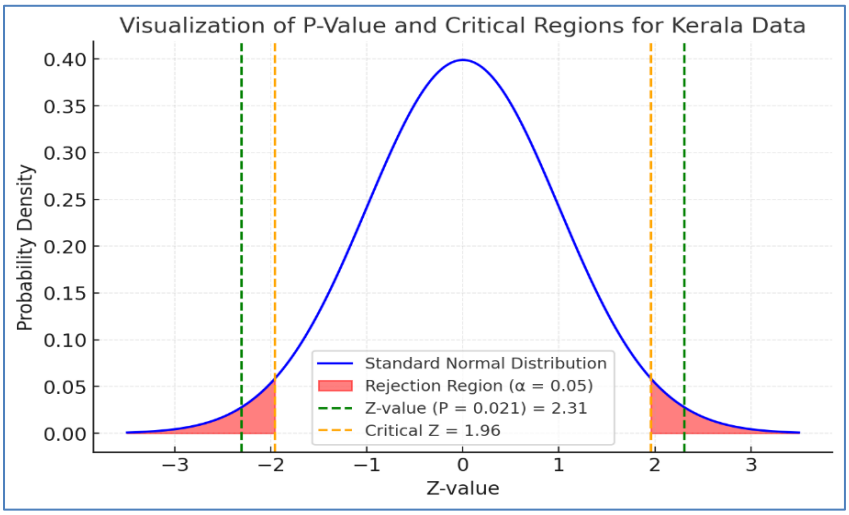


**Fig. 5**. Two tailed test result on KSMR data set. The p-value (0.021) lies within the rejection region

In the present case we have worked with probability values (or p-values). Since the null hypothesis is about the existence of trend and not about the magnitude, we performed two-tail test. A p-value of 0.3701 is obtained. Interpretation of this p-value along with standard regions for 95% confidence (or significance $\propto = 0.05$) is shown in fig. 4. The critical z-value at 95% confidence is 1.96 as shown in the fig. 4. The z-value corresponding to the obtained p-value (0.3701) is 0.90. As the p-value is clearly inside the acceptable range, with z-value significantly less than the critical value, the hypothesis is accepted, and we deduce that the data does not show any trends. The value of Kendall's $\tau$ is -0.0549 indicating a very weak negative trend. Since the p-value test is in the acceptance region this slight negative trend may be due to random and local variations rather than system dynamics. These two observations lead to the conclusion that null hypothesis cannot be rejected or that there is no significant trend. This is a vital conclusion as the calculations were performed on 122 years long seasonal time series. Similar analysis on the KSMR gives the p-value 0.021 and $\tau = -0.14$. The result is presented in fig. 5.

As seen from fig. 5, the hypothesis of "no trend" must be rejected. The value $\tau = -0.14$ indicates weak negative trend. This negative trend, though weak, is statistically significant. The reason for this trend cannot be statistically ascribed to local processes or random events. The dynamics of the system must be investigated for

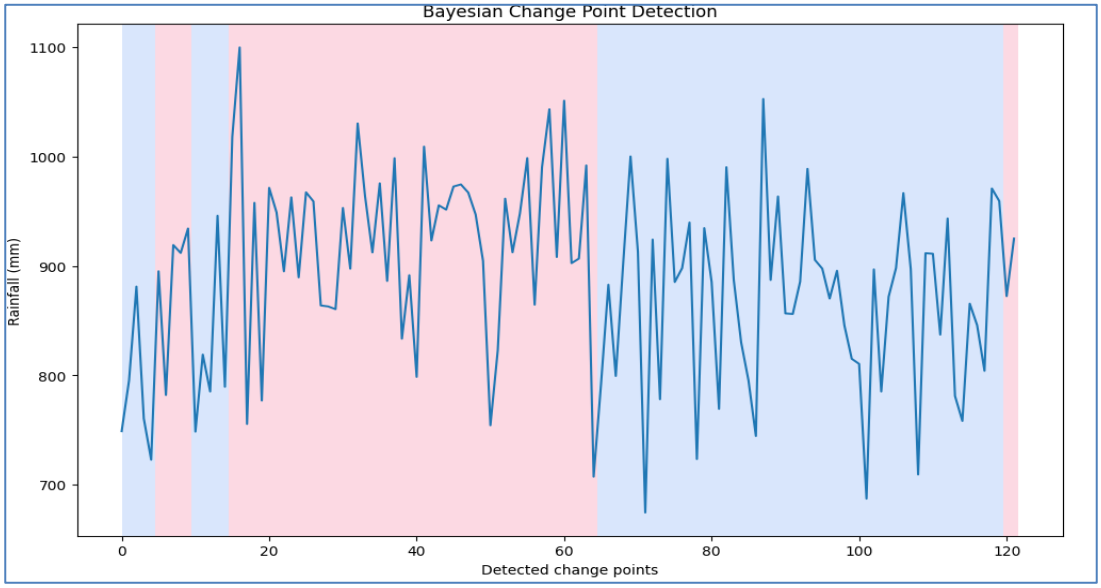


**Fig. 6**: Detection of Multiple change points in ISMR Bayesian change point analysis

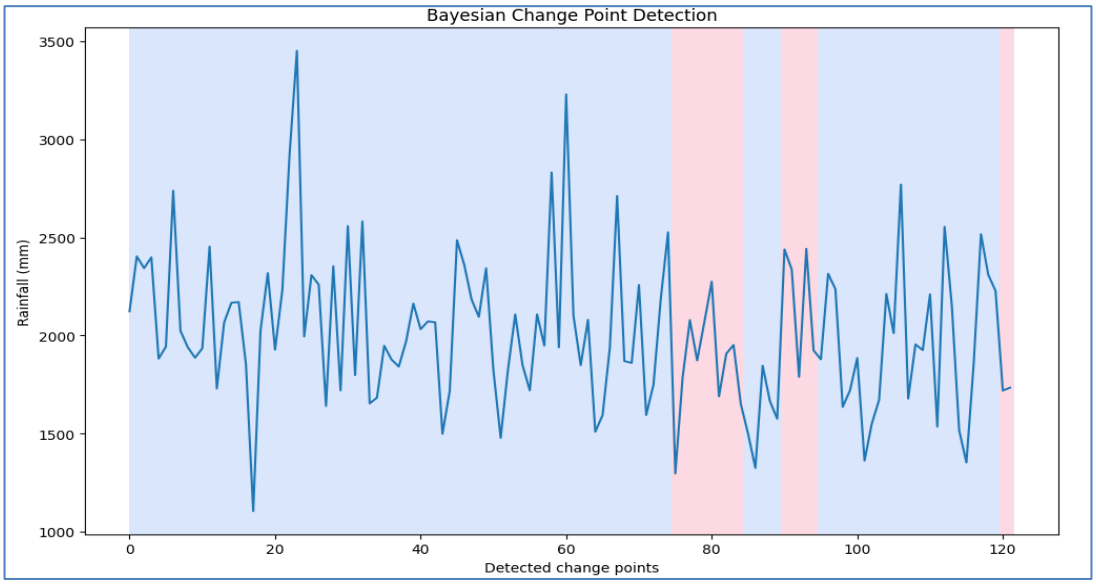


**Fig. 7**: Detection of Multiple change points in KSMR Bayesian change point analysis

meteorological reasons (Goyal *et al*., 2024). That the KSMR has a weak decreasing but statistically significant trend whereas the ISMR has statistically non-significant trend can be explained by the fact that Kerala is located on the windward side of the Western Ghats and is directly influenced by the Arabian Sea branch (Ramesh et. Al., 2009), leading to early, intense rainfall. Oceanic drivers such as the Indian Ocean Dipole, ENSO, and regional Indian Ocean warming episodes also exert a disproportionate influence on the early monsoon rainfall over Kerala. By contrast, ISMR represents an area-weighted average over 36 meteorological subdivisions, where variability in individual regions tends to offset each other, resulting in no significant long-term trend.

### 3.2 *Bayesian break points analysis*

Bayesian test (Harlé *et al*., 2016) has been applied on ISMR and KSMR. Fig. 6 represents seasonal time series of ISMR for the entire span of 122 years. The change points are observed at time steps *T*:

$$T = \{5, 10, 15, 65, 120\} \quad (5)$$

**TABLE 2**

**Parameters of normal distribution in the intervals obtained (ISMR)**

| Range | Years | Mean (μ) | Standard Deviation (σ) |
|---|---|---|---|
| 0-5 | 1901-1906 | 800.55 | 71.84 |
| 6-10 | 1907-1911 | 859.08 | 86.94 |
| 11-15 | 1912-1916 | 871.48 | 104.88 |
| 16-65 | 1917-1966 | 921.41 | 104.88 |
| 66-120 | 1967-2021 | 870.09 | 82.68 |

**TABLE 3**

**Parameters of normal distribution in the intervals obtained (KSMR)**

| Range | Years | Mean (μ) | Standard Deviation (σ) |
|---|---|---|---|
| 0-75 | 1901-1976 | 2068.62 | 401.28 |
| 76-85 | 1977-1986 | 1877.68 | 231.09 |
| 86-90 | 1987-1991 | 1769.82 | 418.54 |
| 91-95 | 1992-1996 | 2073.76 | 294.53 |
| 96-120 | 1997-2021 | 1954.90 | 384.63 |

The parameters (mean and standard deviation) of the time series obtained in these intervals are depicted in table 2.

The numbers indicate swift change in parameters from 1901 to 1915. Although the mean does not vary much, the standard deviation shows a significant change in during 1912-1916. This represents relatively stable monsoon conditions. The high standard deviation of about 105 in the mid-20$^{th}$ century indicate abrupt monsoon rain. These may be attributed to ENSO variability (Krishnamurthy *et al*., 2000; Kumar *et al*., 2006), Pacific Decadal Oscillations (Krishnan, R., & Sugi, M., 2003; Krishnamurthy, & Krishnamurthy, 2014), Post-War aerosol emissions (Ashok *et al*., 2001) and Indian Ocean warming (Ashok *et al*., 2001; Joseph, & Zeng, 2011). An outcome of these were the 1965 floods. A long span of 54 years (1967-2021) did not witness much variability although there may be intermittent years that saw fluctuations. These are small variations and did not affect the climatology on large scale.

Fig. 7 represents seasonal time series of KSMR for the entire span of 122 years. The change points are observed at time steps:

$$T = \{75, 85, 90, 95, 120\} \quad (6)$$

The parameters (mean and standard deviation) of the time series obtained in these intervals are depicted in table 3.

Results suggest shifts in the rainfall regime, potentially linked to changes in regional climate patterns or local factors. The rain patterns demonstrate high variability in the initial long spells 1901-1976 and a small period of 1987-1991. The increase in the mean rainfall after the third breakpoint (from 1769.82 to 2073.76 in the fourth interval) suggests a return to higher rainfall levels and further followed by a small decline in the last interval 1996-2020.

The recent observations indicate return to usual variability. It may be noted that while the ISMR is most variable during 1912-1966 (104.88) the KSMR is more variable during 1901-1976 (401.28). This can be understood as the KSMR drives the ISMR on most occasions. However, there is a stark difference. More break points in the ISMR are observed during the initial periods of observations (1901-1916) while for the KSMR maximum breakpoints are observed during the last periods (1987-2021). This may be attributed to local factors like western disturbances influencing the KSMR. These tend to die down while progressing northwards or eastwards.

### 3.3. *Development of machine learning models for long term modelling of the time series*

As discussed in section 4.5, autocorrelation analysis was done to select the predictors of the time series ahead. The result of autocorrelation analysis of the ISMR time series is presented in fig. 8.

A significant correlation of 0.78, 0.77, 0.88, 0.77, 0.76 and 0.87 are observed at lag values of 1, 11, 12, 13, 23 and 24 respectively. Lag zero is not considered because of the obvious trivial case. Machine learning models SVM, RF, RNN and LSTM were developed using these lag values for predictions: 1, 11, 12, 13, 23, and 24 time steps ahead. The results are presented in Tables 4 (a-f).

It can be observed that best RMSE on training (147) and test (167) cases are obtained by the LSTM model at lag = 12. These RMSE values are significantly less than the standard deviation of the corresponding observed data. The model using lag 12 as the correlated index, $LSTM_{12}$, has been trained as 64-unit LSTM layer to capture sequential patterns and a dense layer for continuous rainfall prediction. To enable LSTM, learn complex patterns, the activation function that is employed is '*tanh*'.

A custom callback monitors RMSE and MAE, saving the optimal weights during training. With a batch size of 1, the model is trained across 1250 epochs, and performance is tracked using validation data. Lastly, the

**TABLE 4**

**Performance of machine learning models for time series modelling of the KSMR at various lag values**

**(a): Performance at $L_1$ (lag=1)**

**Training data at lag 1: coefficient of correlation= .777 and standard deviation=257.2**

| | $SVM_1$ | $RF_1$ | $RNN_1$ | $LSTM_1$ |
|---|---|---|---|---|
| MAE | 151.124 | 170.063 | 164.958 | 156.857 |
| RMSE | 231.584 | 239.995 | 224.141 | 218.799 |

**Testing data at lag 11: coefficient of correlation= .777 and standard deviation=231.963**

| | $SVM_1$ | $RF_1$ | $RNN_1$ | $LSTM_1$ |
|---|---|---|---|---|
| MAE | 142.836 | 161.145 | 146.173 | 144.147 |
| RMSE | 200.526 | 225.218 | 193.660 | 189.542 |

**(b): Performance at $L_{11}$ (lag=11)**

**Training data at lag 11: coefficient of correlation= .772 and standard deviation=257.2**

| | $SVM_{11}$ | $RF_{11}$ | $RNN_{11}$ | $LSTM_{11}$ |
|---|---|---|---|---|
| MAE | 167.275 | 176.774 | 165.513 | 159.258 |
| RMSE | 243.562 | 246.791 | 228.455 | 218.724 |

**Testing data at lag 11: coefficient of correlation= .772 and standard deviation=231.963**

| | $SVM_{11}$ | $RF_{11}$ | $RNN_{11}$ | $LSTM_{11}$ |
|---|---|---|---|---|
| MAE | 154.890 | 177.802 | 155.721 | 149.653 |
| RMSE | 220.431 | 242.831 | 210.738 | 201.947 |

**(c): Performance at $L_{12}$ (lag=12)**

**Training data at lag 12: coefficient of correlation= .88 and standard deviation=257.2**

| | $SVM_{12}$ | $RF_{12}$ | $RNN_{12}$ | $LSTM_{12}$ |
|---|---|---|---|---|
| MAE | 132.359 | 122.236 | 138.033 | 95.5628, |
| RMSE | 214.667 | 184.938 | 209.513 | 147.9189 |

**Testing data at lag 12: coefficient of correlation= .88and standard deviation=231.963**

| | $SVM_{12}$ | $RF_{12}$ | $RNN_{12}$ | $LSTM_{12}$ |
|---|---|---|---|---|
| MAE | 129.624 | 125.167 | 126.916 | 105.4112 |
| RMSE | 186.914 | 183.446 | 177.387 | 167.9367 |

**(d): Performance at $L_{13}$ (lag=13)**

**Training data at lag 13: coefficient of correlation= .772 and standard deviation=257.2**

| | $SVM_{13}$ | $RF_{13}$ | $RNN_{13}$ | $LSTM_{13}$ |
|---|---|---|---|---|
| MAE | 162.481 | 168.462 | 167.307 | 158.370 |
| RMSE | 241.050 | 139.307 | 226.224 | 220.787 |

**Testing data at lag 13: coefficient of correlation= .772 and standard deviation=231.963**

| | $SVM_{13}$ | $RF_{13}$ | $RNN_{13}$ | $LSTM_{13}$ |
|---|---|---|---|---|
| MAE | 155.943 | 177.786 | 155.503 | 150.342 |
| RMSE | 214.887 | 241.635 | 203.016 | 197.466 |

**(e): Performance at $L_{23}$ (lag=23)**

**Training data at lag 23: coefficient of correlation= .756 and standard deviation=257.2**

| | $SVM_{23}$ | $RF_{23}$ | $RNN_{23}$ | $LSTM_{23}$ |
|---|---|---|---|---|
| MAE | 175.900 | 174.102 | 174.864 | 167.594 |
| RMSE | 266.530 | 246.955 | 248.955 | 223.750 |

**Testing data at lag 23: coefficient of correlation= .756 and standard deviation=231.963**

| | $SVM_{23}$ | $RF_{23}$ | $RNN_{23}$ | $LSTM_{23}$ |
|---|---|---|---|---|
| MAE | 162.927 | 148.985 | 155.393 | 143.645 |
| RMSE | 236.266 | 213.849 | 214.673 | 201.675 |

**(f): Performance at $L_{24}$ (lag=24)**

**Training data at lag 24: coefficient of correlation= .88 and standard deviation=257.2**

| | $SVM_{24}$ | $RF_{24}$ | $RNN_{24}$ | $LSTM_{24}$ |
|---|---|---|---|---|
| MAE | 136.118 | 123.757 | 148.896 | 127.601 |
| RMSE | 210.411 | 183.799 | 198.142 | 189.524 |

**Testing data at lag 24: coefficient of correlation= .88and standard deviation=231.963**

| | $SVM_{24}$ | $RF_{24}$ | $RNN_{24}$ | $LSTM_{24}$ |
|---|---|---|---|---|
| MAE | 131.566 | 124.892 | 134.976 | 114.323 |
| RMSE | 200.794 | 204.284 | 178.721 | 173.497 |

model with the optimal weights is saved after RMSE and MAE trends are stabilised which happens at epoch 1016. The results are shown in Table 4(c).

In order the analyse the models for sustainability, we proceed to do the MK and Bayesian tests on the results obtained by the best model (LSTM at lag = 12). The result of break point analysis is presented in fig. 9.

As seen from the fig., the change points are observed at

$$T1 = \{20, 75, 90, 100, 105\} \quad (7)$$

Comparison with (6) is presented in table 5.

The second, third, fourth and fifth break points are observed almost at same values in both the time series. We have compared the values of the parameters in the predicted time series in the intervals of the observed time series. The results are presented in Table 6.

It is evident that there is a persistent bias, of around 300-400 in the mean. This is systemic bias of the model and can be removed using methods like z-score normalization, or statistical bias correction. However, the range of mean in the predicted series is less variable. This means that the variance in the mean is not captured. The

**TABLE 5**

**Comparison between observed and predicted break points for Kerala and Mahe**

| Break point no. | Observed data | Predicted data |
|---|---|---|
| 1 | 75 | 20 |
| 2 | 85 | 75 |
| 3 | 90 | 90 |
| 4 | 95 | 100 |
| 5 | 120 | 105 |

**TABLE 6**

**Observed and predicted parameters in the break intervals**

| Range | Mean (μ) observed | Mean (μ) predicted | Standard Deviation (σ) observed | Standard Deviation (σ) Predicted |
|---|---|---|---|---|
| 1-75 | 2068.62 | 1602.3420 | 401.28 | 166.4324 |
| 76-85 | 1877.68 | 1538.2822 | 231.09 | 156.1677 |
| 86-90 | 1769.82 | 1440.8948 | 418.54 | 122.1056 |
| 91-95 | 2073.76 | 1615.0712 | 294.53 | 118.2733 |
| 96-120 | 1954.90 | 1628.5736 | 384.63 | 195.3989 |

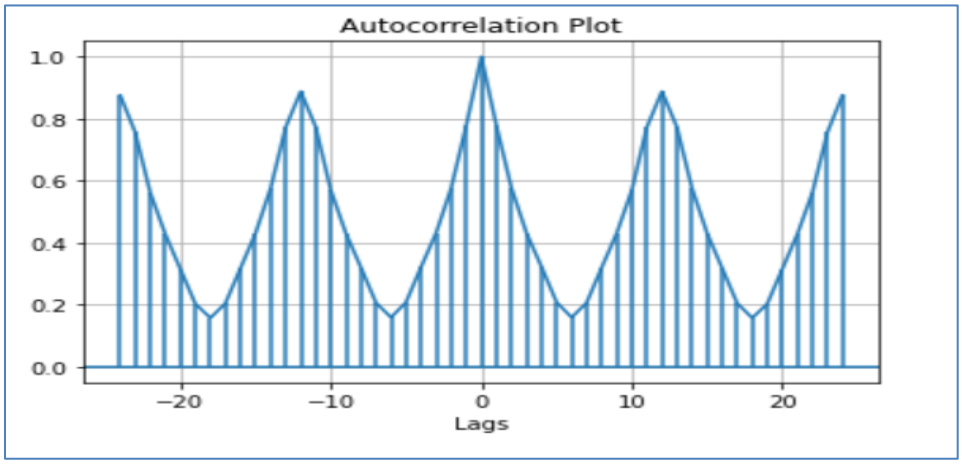


**Fig. 8.** Autocorrelation plot at different interval of time in training dataset of Kerala

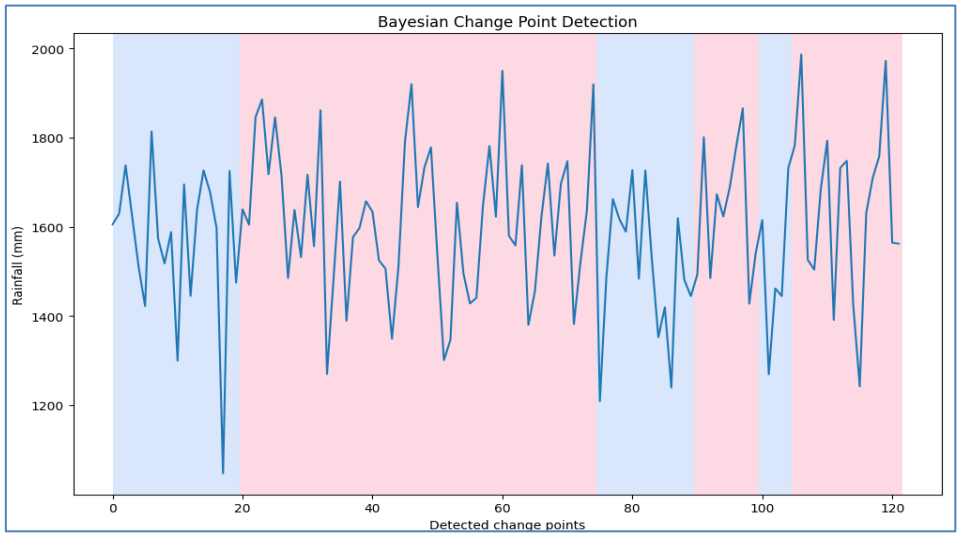


**Fig. 9.** Detection of Multiple change points in total JJAS rainfall of Kerala and Mahe in predicted data

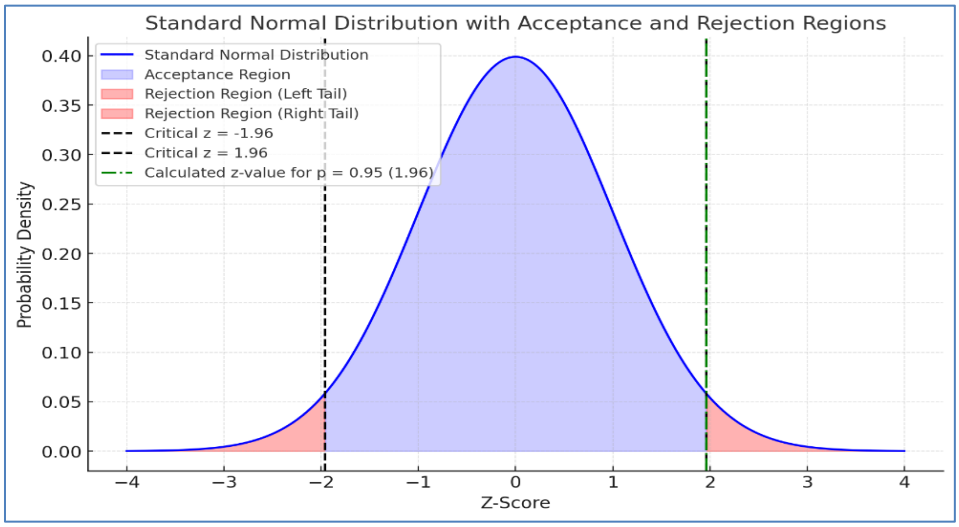


**Fig. 10**. Two tailed test result on KSMR prediction data set. The p-value (0.95) lies exactly on the critical line

same is the case with standard deviations. The standard deviations are less varied in the predicted time series. although the standard deviations do not vary abruptly and is in the scope of the observed series, the predicted series is less variable. The persistent bias of around 300–400 mm in the predicted means likely reflects the absence of explicit bias-correction in the modelling pipeline. While the present study prioritised the sustainability of long-term statistical characteristics such as trends and break points, standard post-processing methods (e.g., normalization, mean adjustment, or variance scaling) could help to reduce this offset without altering the statistical structure of the predictions. This aspect will be taken up in future work to refine absolute rainfall estimates.

This result is remarkable seeing the time series of KSMR was modelled and we are predicting with lag 12. This means we are predicting 12 "seasons" ahead or 12 years ahead. Hence the LSTM model has significantly captured the long statistics of the time series.

The result of MK analysis on the KSMR predicted series gives p-value 0.95 and $\tau = -0.003$. The result is presented in fig. 10. The small negative value of $\tau$ indicates a very weak negative trend. This is in line with the trend in the observed data. It may be noted that this result is borderline, lying exactly on the critical threshold ($z = \pm 1.96$ at 95% confidence). While this does not strongly confirm or reject the null hypothesis, the outcome remains consistent with the weak negative trend found in the observed KSMR data. Hence, the borderline result does not alter our conclusion but rather highlights that the model is able to replicate even marginal statistical signals in the observed series. As can be seen, it is a peculiar case here as the obtained z-value aligns with the corresponding critical line (1.96 at confidence level of 95%). Thus, the hypothesis of "no trend" may be rejected. This is in line with calculations on the observe data set where the trend was weak but statistically significant. However, in the prediction case the significance is a border-line significance.

### 3.4 *Limitations*

The present study has certain limitations that should be noted. First, the Mann–Kendall test applied to the predicted KSMR series yielded a borderline result, lying exactly on the 95% confidence threshold; while this does not change our overall conclusion, it indicates marginal statistical significance. Second, the prediction models exhibit a persistent bias of around 300–400 mm in the mean, which was acknowledged but not post-processed in this work. Third, break point analysis was conducted using a Bayesian method, and results may vary if alternative change-point detection techniques are applied. Finally, the modelling framework relied solely on rainfall time series, without incorporating external predictors such as SST anomalies or atmospheric circulation indices that may further improve robustness. These limitations do not detract from the main findings but highlight areas for refinement in future studies.

## 4. Conclusions

The trend analysis indicates no significant trend in the ISMR while a statistically significant one in the KSMR. There is also a difference in the positions of maximum break points observed in the ISMR and the KSMR. While the ISMR witnessed more breakpoints in the initial phases the KSMR witnessed more in the recent years. It may thus be concluded that while the onset of monsoon in India is from the Kerala, the trend and pattern of Kerala does not decide the trends of the remaining parts. While the local factors affect Kerala, the all-India pattern is dominated by global factors. Four different machine learning models were developed to model these break points and trends using the autocorrelations in the KSMR time series. It was seen that the model was able to correlate with the mean and standard deviations in the intervals of break points. Further, the prediction results were in line with the observed values in the MK test. It can thus be concluded that the machine learning based very long term modelled series is statistically sustainable. Further research is needed to corroborate the findings based on dynamical nature of the time series and analysis with external forcings. Further research is needed to extend the present approach by incorporating large-scale dynamical forcings such as the El Niño–Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and regional Indian Ocean warming events, which are known to strongly modulate monsoon rainfall variability. Including such predictors in the modelling framework could enhance the robustness and physical interpretability of long-term rainfall forecasts.

*Authors' Contributions*

Namita Goyal: Methodology, coding, initial draft and analysis. (*e-mail - er.namitagoel@gmail.com* )
Aparna N Mahajan: Supervision, review and editing.
(*e-mail - aparnanmahajan@yahoo.co.in* )
K. C. Tripathi: Supervision, concept, methodology, experimental set up, review, interpretation and analysis. (*e-mail - kctripathi@mait.ac.in*)

*Disclaimer*: The contents and views presented in this article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

**References**


Acharya, R., Pal, J., Das, D., and Chaudhuri, S., 2019, "Long-range forecast of Indian summer monsoon rainfall using an artificial neural network model," *Meteorological Applications*, 26, 3, 347-361, https://doi.org/10.1002/met.1766.

Adcroft, A., Hill, C., Campin, J.M., Marshall, J., and Heimbach, P., 2004, "Overview of the formulation and numerics of the MIT GCM," *Proc. ECMWF Seminar Series on Numerical Methods, Recent Developments in Numerical Methods for Atmosphere and Ocean Modelling*, 139-149

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

Ashok, K., Guan, Z., & Yamagata, T., 2001, "Impact of the Indian Ocean dipole on the relationship between the Indian monsoon rainfall and ENSO," *Geophysical research letters*, 28, 23, 4499-4502, https://doi.org/10.1029/2001GL013294.

Ashok, S. P., & Pekkat, S., 2022, "A systematic quantitative review on the performance of some of the recent short-term rainfall forecasting techniques," *Journal of Water and Climate Change*, 13, 8, 3004-3029, https://doi.org/10.2166/wcc.2022.302.

Bhatla, R., Pant, M., Ghosh, S., Verma, S., Pandey, N., and Bist, S., 2023, "Variations in Indian summer monsoon rainfall patterns in changing climate," *Mausam*, 74, 3, 639-650, https://doi.org/10.54302/mausam.v74i3.5940

Calvo-Olivera, C., Guerrero-Higueras, Á.M., Lorenzana, J., and García-Ortega, E., 2024, "Real-Time Evaluation of the Uncertainty in Weather Forecasts Through Machine Learning-Based Models," *Water Resources Management*, 1-16, 10.1007/s11269-024-03779-y.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y., 2018, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, 8, 1, 6085, https://doi.org/10.1038/s41598-018-24271-9.

Climate Research & Services, Pune, "Data and information on Indian meteorology," IMD Pune (imdpune.gov.in), https://www.imdpune.gov.in/

Dangeti, P., 2017, *Statistics for machine learning*, Packt Publishing Ltd.

Dash, Y., Mishra, S.K., and Panigrahi, B.K., 2019, "Predictability assessment of northeast monsoon rainfall in India using sea surface temperature anomaly through statistical and machine learning techniques," *Environmetrics*, 30, 4, e2533, https://doi.org/10.1002/env.2533.

Ferreira, L.B., and da Cunha, F.F., 2020, "Multi-step ahead forecasting of daily reference evapotranspiration using deep learning," *Computers and electronics in agriculture*, 178, 105728, https://doi.org/10.1016/j.compag.2020.105728.

Gadgil, S., 2003, "The Indian monsoon and its variability," *Annu. Rev. Earth Planet. Sci.*, 31, 429-467, https://doi.org/10.1146/annurev.earth.31.100901.141251.

Gallagher, C., Lund, R., & Robbins, M., 2013, "Changepoint detection in climate time series with long-term trends," *Journal of Climate*, 26, 14, 4994-5006, https://doi.org/10.1175/JCLI-D-12-00704.1.

Goyal, N., Mahajan, A. N., & Tripathi, K. C. 2024, "A Systematic Review on Relationship Between the Monsoon's Variability and the Variables Influencing it Through Machine Learning," *Emerging Trends in IoT and Computing Technologies*, 460-464, https://doi.org/10.1201/9781003535423.

Goyal, N., Mahajan, A.N., and Tripathi, K.C., 2023, "Conception of Indian Monsoon Prediction Methods," *International Conference on Communication and Intelligent Systems*, Springer Nature Singapore, 247-263, https://doi.org/10.1007/978-981-97-2079-8_20.

Harlé, F., Chatelain, F., Gouy-Pailler, C., & Achard, S., 2016, "Bayesian model for multiple change-points detection in multivariate time series," *IEEE Transactions on Signal Processing*, 64, 16, 4351-4362, https://doi.org/10.1109/TSP.2016.2566609.

Hayaty, N., Kurniawan, H., Rathomi, M. R., Chahyadi, F., & Bettiza, M., 2023, "Rainfall Prediction with Support Vector Machines: A Case Study in Tanjungpinang City, Indonesia," *Bio web of conferences*, 70, 01003, https://doi.org/10.1051/bioconf/20237001003.

Joseph, R., & Zeng, N., 2011, "Seasonally modulated tropical drought induced by volcanic aerosol," *Journal of Climate*, 24, 8, 2045-2060, https://doi.org/10.1175/2009JCLI3170.1.

Kelkar, R.R., and Sreejith, O.P., 2020, "Meteorological sub-divisions of India and their geopolitical evolution from 1875 to 2020", Mausam, 71, 4, pp. 571-584, https://doi.org/10.54302/mausam.v71i4.38.

Kang, J., Wang, H., Yuan, F., Wang, Z., Huang, J., & Qiu, T., 2020, "Prediction of precipitation based on recurrent neural networks in Jingdezhen, Jiangxi Province, China," *Atmosphere*, 11, 3, 246, https://doi.org/10.3390/atmos11030246.

Krishnamurthy, V., & Goswami, B. N., 2000, "Indian monsoon–ENSO relationship on interdecadal timescale," *Journal of climate*, 13, 3, 579-595, https://doi.org/10.1175/1520-0442(2000)013<0579:IMEROI>2.0.CO;2.

Krishnamurthy, L., & Krishnamurthy, V. J. C. D., 2014, "Influence of PDO on South Asian summer monsoon and monsoon–ENSO relation," *Climate dynamics*, 42, 2397-2410, https://doi.org/10.1007/s00382-013-1856-z.

Krishnan, R., & Sugi, M., 2003, "Pacific decadal oscillation and variability of the Indian summer monsoon rainfall," *Climate dynamics.*, 21, 233-242, https://doi.org/10.1007/s00382-003-0330-8.

Kumar, K. K., Rajagopalan, B., Hoerling, M., Bates, G., & Cane, M., 2006, "Unraveling the mystery of Indian monsoon failure during El Niño," *Science*, 314, 5796, 115-119, doi: 10.1126/science.1131152.

Kumar, A., & Singh, S., 2021, "A review on Indian summer monsoon rainfall prediction using machine learning techniques," *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, IEEE, 524-528, doi: 10.1109/ICSCCC51823.2021.9478104.

Kumar, B., Abhishek, N., Chattopadhyay, R., George, S., Singh, B. B., Samanta, A., ... & Singh, M., 2022, "Deep learning based short-range forecasting of Indian summer monsoon rainfall using earth observation and ground station datasets," *Geocarto International*, 37, 27, 17994-18021, https://doi.org/10.1080/10106049.2022.2136262.

Kumar, S., Ahmed, S.A., and Karkala, J., 2023, "Time series data and rainfall pattern subjected to climate change using non-parametric tests over a vulnerable region of Karnataka, India," *Journal of Water and Climate Change*, 14, 5, 1532-1550, https://doi.org/10.2166/wcc.2023.441.

Lenka, S., Gouda, K.C., Devi, R., and Joseph, C.M., 2024, "Dynamics of Indian summer monsoon in different phases," *Climate Dynamics*, 62, 1, 473-495, https://doi.org/10.1007/s00382-023-06925-1.

Liyew, C.M., and Melese, H.A., 2021, "Machine learning techniques to predict daily rainfall amount," *Journal of Big Data*, 8, 1-11, https://doi.org/10.1186/s40537-021-00545-4.

Mellor, G.L., 1998, "Users guide for a three-dimensional, primitive equation, numerical ocean model," *Princeton University*, Program in Atmospheric and Oceanic Sciences, http://jes.apl.washington.edu/modsims_two/usersguide0604.pdf.

Meyer, P. L., 1965, *Introductory probability and statistical applications*, Oxford and IBH Publishing.

Mitsui, T., & Boers, N., 2021, "Seasonal prediction of Indian summer monsoon onset with echo state networks," *Environmental Research Letters*, 16, 7, 074024, DOI 10.1088/1748-9326/ac0acb.

Munot, A.A., and Kumar, K.K., 2007, "Long range prediction of Indian summer monsoon rainfall," *Journal of earth system science*, 116, 73-79, https://doi.org/10.1007/s12040-007-0008-4.

Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., and Zhang, J., 2020, "Streamflow and rainfall forecasting by two long short-term memory-based models," *Journal of Hydrology*, 583, 124296, https://doi.org/10.1016/j.jhydrol.2019.124296.

Palmer, T.N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., and Thomson, M.C., 2004, "Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER)," *Bulletin of the American Meteorological Society*, 85, 6, 853-872, https://doi.org/10.1175/BAMS-85-6-853.

Pacanowski, R.C., Dixon, K., and Rosati, A., 1993, "The GFDL modular ocean model users guide," *GFDL Ocean Group Tech. Rep.*, 2, 46, 08542-0308.

Perera, A., Ranasinghe, T., Gunathilake, M., and Rathnayake, U., 2020, "Comparison of different analyzing techniques in identifying rainfall trends for Colombo, Sri Lanka," *Advances in Meteorology*, 2020, 1-10, https://doi.org/10.1155/2020/8844052.

Poornima, S., & Pushpalatha, M., 2019, "Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units," *Atmosphere*, 10, 11, 668, https://doi.org/10.3390/atmos10110668.

Prakash, S., Mitra, A. K., Momin, I. M., Rajagopal, E. N., Milton, S. F., & Martin, G. M., 2016, "Skill of short-to medium-range monsoon rainfall forecasts from two global models over India for hydro-meteorological applications," *Meteorological Applications*, 23, 4, 574-586, https://doi.org/10.1002/met.1579.

Praveen, B., Talukdar, S., Shahfahad, Mahato, S., Mondal, J., Sharma, P., and Rahman, A., 2020, "Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches," *Scientific reports*, 10, 1, 10342, https://doi.org/10.1038/s41598-020-67228-7.

Rajan, D., and Desamsetti, S., 2021, "Prediction of Indian summer monsoon onset with high-resolution model: A case study," *SN Applied Sciences*, 3, 1-14, https://doi.org/10.1007/s42452-021-04646-w.

Rajput, J., Kushwaha, N.L., Sena, D.R., Singh, D.K., and Mani, I., 2023, "Trend assessment of rainfall, temperature and relative humidity using non-parametric tests in the national capital region, Delhi," *Mausam*, 74, 3, 593-606, https://doi.org/10.54302/mausam.v74i3.4936.

Ramesh Kumar, M.R., Sankar, S., and Reason, C., 2009, "An investigation into the conditions leading to monsoon onset over Kerala", *Theoretical and applied climatology*, 95, 1, pp. 69-82, https://doi.org/10.1007/s00704-008-0376-y.

Rao, G., Sowjanya, A., Shekhar, D., Naik, B., and Kiran, B.V.S., 2023, "Rainfall analysis over 31 years of Chintapalle, Visakhapatnam, high altitude and Tribal zone, Andhra Pradesh, India," *Mausam*, 74, 3, 685-698, https://doi.org/10.54302/mausam.v74i3.818.

Saminathan, S., Medina, H., Mitra, S., and Tian, D., 2021, "Improving short to medium range GEFS precipitation forecast in India," *Journal of hydrology*, 598, 126431, https://doi.org/10.1002/essoar.10506333.1.

Saha, M., Mitra, P., & Nanjundiah, R. S., 2017, "Deep learning for predicting the monsoon over the homogeneous regions of India," *Journal of Earth System Science*, 126, 1-18, https://doi.org/10.1007/s12040-017-0838-7.

Saha, M., Santara, A., Mitra, P., Chakraborty, A., & Nanjundiah, R. S., 2021, "Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model," *International Journal of Forecasting*, 37, 1, 58-71, https://doi.org/10.1016/j.ijforecast.2020.03.001.

Scher, S., and Messori, G., 2018, "Predicting weather forecast uncertainty with machine learning," *Quarterly Journal of the Royal Meteorological Society*, 144, 717, 2830-2841, https://doi.org/10.1002/qj.3410.

Shukla, R. P., Tripathi, K. C., Pandey, A. C., & Das, I. M. L., 2011, "Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach," *Atmospheric Research*, 102, 1-2, 99-109, https://doi.org/10.1016/j.atmosres.2011.06.013.

Suman, M., and Maity, R., 2020, "Southward shift of precipitation extremes over south Asia: Evidences from CORDEX data," *Scientific reports.*, 10, 1, 6452, https://doi.org/10.1038/s41598-020-63571-x.

Sudarsan, G., & Lasitha, A., 2023, "Rainfall Trend analysis using Mann-Kendall and Sen's slope test estimation-A case study," *E3S Web of Conferences*, 405, 04013, https://doi.org/10.1051/e3sconf/202340504013.

Uddin, M. J., Li, Y., Tamim, M. Y., Miah, M. B., & Ahmed, S. S., 2022, "Extreme rainfall indices prediction with atmospheric parameters and ocean–atmospheric teleconnections using a random forest model," *Journal of Applied Meteorology and Climatology*, 61, 6, 651-667, https://doi.org/10.1175/JAMC-D-21-0170.1.

Tabari, H., Somee, B. S., & Zadeh, M. R., 2011, "Testing for long-term trends in climatic variables in Iran," *Atmospheric research*, 100, 1, 132-140, https://doi.org/10.1016/j.atmosres.2011.01.005.

Tripathi, K. C., Das, I. M. L., & Sahai, A. K., 2006, "Predictability of sea surface temperature anomalies in the Indian Ocean using artificial neural networks," I*ndian Journal of Marine Sciences*, 26, 1-9,

Valipour, M., Khoshkam, H., Bateni, S.M., and Jun, C., 2024, "Machine-learning-based short-term forecasting of daily precipitation in different climate regions across the contiguous United States," *Expert Systems with Applications*, 238, 121907, https://doi.org/10.1016/j.eswa.2023.121907

Wangwongchai, A., Waqas, M., Dechpichai, P., Hlaing, P.T., Ahmad, S., and Humphries, U.W., 2023, "Imputation of missing daily rainfall data: A comparison between artificial intelligence and statistical techniques," *MethodsX*, 11, 102459, https://doi.org/10.1016/j.mex.2023.102459

Wang, B., Lu, J., Yan, Z., Luo, H., Li, T., Zheng, Y., and Zhang, G., 2019, "Deep uncertainty quantification: A machine learning approach for weather forecasting," *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2087-2095, https://doi.org/10.1145/3292500.3330704.

Zaz, S.N., Romshoo, S.A., Ramkumar, T.K., and Babu, V., 2018, "Climatic and extreme weather variations over mountainous Jammu and Kashmir, India: Physical explanations based on observations and modelling," *Atmos Chem Phys Discuss*, 19, 15-37, https://doi.org/10.5194/acp-2018-201.

Zenkner, G., and Navarro-Martinez, S., 2023, "A flexible and lightweight deep learning weather forecasting model," *Applied Intelligence*, 53, 21, 24991-25002, https://doi.org/10.1007/s10489-023-04824-w

***Supplementary Material***

**TABLE S1**

**Summary of related work**

| Ref | Objective | Model/Technique Used | Dataset | Performance Metrics | Findings |
|---|---|---|---|---|---|
| Valipour *et al*., 2024 | To acquire short-term precipitation projections for each day | WPSOANFIS: Wavelet Particle Swarm Optimization Adaptive Neuro-Fuzzy Inference System. WGMDH: Wavelet Group Method of Data Handling. WLSTM: wavelet long short-term memory | 1995 -2019 | MAE (Mean Absolute Error) RMSE (Root Mean Square Error) | WGMDH and WLSTM produce better forecasts than WPSOANFIS. |
| Lenka *et al*., 2024 | To estimate onset and withdrawal date of the monsoon based on daily rainfall data | The principal component neural network model created by combining two existing models | 1901-1997, IMD | RMSE | Proposed model outperformed than other two models and currently employed by IMD for LRF of Indian summer monsoon. |
| Calvo-Olivera *et al*., 2024 | To assess weather prediction's unreliability produced by forecasting model. | Model Evaluator (MoEv), a Scikit-Learn library wrapper | Self-created dataset using inputs from WRF model | RMSE, Uncertainty index 'u' | Ensemble techniques based on decision trees achieve the lowest generalization error. Accuracy-90% |
| Wangwong chai *et al*., 2023 | To impute missing daily rainfall data in northern Thailand using statistical and AI techniques. | Statistical Techniques (STs) and Artificial Intelligence-based Techniques (AITs) | 30 years of daily rainfall data from 20 northern Thailand rainfall stations. | MAE, RMSE, R (Corelation Coefficient) $R^2$ (Coefficient of Determination) | Multiple linear regression model outperformed. |
| Kumar *et al*., 2023 | Comprehension of the semi-arid region of Karnataka's rainfall pattern. | Innovative trend analysis (ITA) and the modified Mann-Kendall test (MMK), ARIMA model. | 1952-2019, IMD | The SNHT, Pettitt, and Buishand tests for homogeneity test. | Compared to the ITA, the MMK approach demonstrated a greater occurrence of a significant growing trend. |
| Zenker *et al*., 2023 | To forecast the weather for the upcoming 24 and 72 hours based on the last 120 hours. | Two different models of Bi-LSTM recurrent neural networks | 2015- 2021, Met Office, London | RMSE | First model shows temperature prediction with ±2 °C accuracy and RMSE of 1.45 °C Second model predicted air temperature and relative humidity over a 72-hour period with RMSE of 2.26 °C and 14% . |
| Bhatla *et al*., 2023 | To assess how the Indian subcontinent's geographic variability in ISMR | Studied ISM time scale trends on the tricadal and decadal scales. | 1901-2013 from IMD. | Standard deviation and Coefficient of variation | improvement in ISMR over Western India but a significant decrease over Northeast Indian regions, indicating a shift in ISMR toward the west due to climate change. |

| Ref | Objective | Model/Technique Used | Dataset | Performance Metrics | Findings |
|---|---|---|---|---|---|
| Rajput *et al*., 2023 | To track the National Capital Region's rainfall and temperature trends over a period of 31 years. | Mann-Kendall (MK) test, Theil Sen slope estimator test, and Pettitte Test | 1990-2020, IARI meteorological station, Pusa, New Delhi. | Mean, median, mode, standard deviation, variance, skewness, kurtosis, maximum, minimum, $25^{th}$, $50^{th}$, and $75^{th}$ percentile | Found a tendency of upward increase in the relative humidity data series. |
| Rao *et al*., 2023 | To examine and measure the distribution of rainfall in the Andhra Pradesh, India district of Chintapalli, Visakhapatnam. | Linear Regression, Sen's slope estimates, the Man-Kendall test, and Rainfall Anomaly Index (RAI) | 1990-2020, Regional Agricultural Research Station (RARS), Chintapalle. | Standard Precipitation Index (SPI), Mean, median, standard deviation, coefficient of variation, percentage of annual rainfall and ZC test | Positive increasing trend in rainfall, the extremely dry year was 2002. Three drought events have occurred in the last ten years; however, they have not been very large. |
| Saminathan *et al*. 2021 | To improve daily predictions of precipitation over the Indian subcontinent. | Global Ensemble Forecast System (GEFS) using Analog (AN) and Logistic Regression (LR) techniques. | IMD | Brier Skill Score (BSS) and Root Mean Square Error (RMSE) | The post-processing technique combining LR and AN significantly enhanced short- to medium-range (1–7 day) precipitation forecasts over India. |
| Ali *et al*., 2020 | To overcome the non-stationarity issues that are encountered by rainfall forecasting models | CEEMD-RF-KRR: Complete Ensemble Empirical Mode Decomposition in conjunction with Random Forest and Kernel Ridge Regression algorithm | 1962 to 2013 Pakistan Meteorological Department (PMD) | Correlation coefficient, Willmott's index, Nash-Sutcliffe coefficient, Legates-McCabe's index | The suggested model outperforms the compared models in terms of performance. |
| Ni *et al*., 2020 | Investigating LSTM's potential for rainfall and streamflow forecasting | Wavelet-LSTM (WLSTM), Convolutional LSTM (CLSTM) | National Meteorological Information Centre, China | RMSE, MARE, (NSE) Nash-Sutcliffe model efficiency coefficient | While LSTM could be used to predict time series, WLSTM and CLSTM performed better. |
| Perera *et al*., 2020 | To evaluate and contrast three trend analysis methods for Colombo, Sri Lanka | Sen's slope estimator, Spearman's rho test, Mann-Kendall test, and novel graphical technique. | 30 years of rainfall data collected for 10 gauging stations | Three resolution levels: yearly, seasonal, and monthly rainfall. | Spearman's rho and the Mann-Kendall test produced similar trends. While, the graphical approach yielded contradictory results occasionally. |
| Ferreira *et al*., 2020 | To estimate daily ETo in several steps ahead using direct, iterated, and multi-input multi output forecasting techniques. | CNN-LSTM, CNN-1D, and LSTM and traditional machine learning models including ANN and RF | 53 stations, Minas Gerais, Brazil | Mean RMSE | Compared to machine learning models, deep learning models outperformed them by a little margin. |
| Suman & Maity, 2020 | To analyze ISM rainfall records | Coupled Model Intercomparison Project Phase 5 (CMIP5) | CORDEX data | Daily mean precipitation, yearly mean precipitation, Mann-Kendall test, and precipitation thresholds-P95 and P-99 | Extraordinary precipitation in south India from 1971 to 2017 in contrast to north and central India. Indicates shift in precipitation towards the south over South Asia. |

| Ref | Objective | Model/Technique Used | Dataset | Performance Metrics | Findings |
|---|---|---|---|---|---|
| | | | | | |
| Acharya *et al*., 2019 | Long range forecasting of ISMR | ANN model with nonlinear perceptron rule with two set of predictors: Set I and Set II | IMD, 1980-2017 | Willmott's index | North Central Pacific zonal wind and North Atlantic Sea surface pressure anomaly in the input matrix makes the second-stage forecast superior than the first-stage forecast. |
| Wang *et al*., 2019 | To create a data-driven approach that uses an efficient information from NWP | New Loss Function based on negative log-likelihood error (NLE). | Public dataset, Beijing, China | MSE, MAE | Concurrently applies uncertain quantification and single-value forecasting, improving accuracy by 47.76% |
| Che *et al*., 2018 | Utilizing the missing patterns for efficient imputation to enhance prediction accuracy | GRU-D model based on Gated Recurrent Unit (GRU) | clinical real-world datasets (PhysioNet, MIMIC-III) and artificial datasets | AUC score (Area Under ROC curve) | Improved prediction results by capturing the long-term temporal interdependence of time series observations and missing patterns. |
| Scher & Messori, 2018 | To estimate future forecast uncertainty from previous forecasts using machine learning | CNN | GEFS reforecast v2 dataset, updated on daily basis. | RMSE | It determines if the predictability by assigning a scalar value of confidence to medium-range forecasts initiated from the stated atmospheric state |
| Zaz *et al*., 2018 | Examined six stations for specific long- and short-term characteristics, localized fluctuations in temperature and precipitation. | WRF model, Linear Regression, Mann-Kendall, Spearman Rho, Cumulative Standard Deviation, and Student's T-test | 1980-2016, IMD for stations of Jammu and Kashmir. | Different confidence levels of S; S=99%, S=95% and S=90%. | Increased temperature of higher altitude is correlated with minor drop in yearly precipitation. Winter NAO index and long-term variations in temperature and precipitation are well correlated. |