MAUSAM, **77**, 3 (July 2026),1079-1092

MAUSAM

DOI: https://doi.org/10.54302/mausam.v77i3.6766
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No.551.577.2:004.85:551.509.3

Prediction of rainfall through a ML/DL approach over major metro cities of Uttar Pradesh, India

MAYANK PANDEY [1,*] AND SHAILENDRA RAI[2]

[1]*K. Banerjee Centre of Atmospheric and Ocean Studies, University of Allahabad, Prayagraj-211002, UP, INDIA*

[2]*M. N. Saha Centre of Space Studies, University of Allahabad, Prayagraj-211002, UP, INDIA*

(*Received 25 July 2024, Accepted 7 August 2025*)

***Corresponding author'semail: mayankau@allduniv.ac.in**

सार – मौसम का पूर्वानुमान लगाना एक महत्वपूर्ण और चुनौतीपूर्ण कार्य है क्योंकि अधिकांश वायुमंडलीय और कृषि क्षेत्र दैनिक मौसम के उतार-चढ़ाव पर निर्भर करते हैं। वर्षा विभिन्न जलवायु स्थितियों पर निर्भर सबसे महत्वपूर्ण मापदंडों में से एक है। हमने 1901 से 2020 तक जून, जुलाई, अगस्त और सितंबर (JJAS) के महीनों में उत्तर प्रदेश के प्रमुख शहरों के लिए वर्षा का पूर्वानुमान लगाने के लिए रैंडम फ़ॉरेस्ट (RF) और लॉन्ग शॉर्ट-टर्म मेमोरी (LSTM) न्यूरल नेटवर्क मॉडलों का उपयोग किया। हमने पूर्वानुमान की गुणवत्ता का आकलन करने के लिए सहसंबंध गुणांक (CC), रूट मीन स्क्वायर त्रुटि (RMSE), और माध्य निरपेक्ष त्रुटि (MAE) जैसे विभिन्न सांख्यिकीय सूचकांकों का उपयोग किया। जैसा कि ऊपर उल्लेख किया गया है, वर्षा का पूर्वानुमान लगाने के लिए कई जलवायु सूचकांकों का उपयोग भविष्यवक्ता (Predictors) के रूप में किया गया था। इन सूचकांकों में उत्तरी अटलांटिक समुद्री सतह का तापमान, नीनो 3.4, भूमध्यरेखीय दक्षिण-पूर्वी हिंद महासागर और 12 के अंतराल मूल्य (Lag value) पर वर्षा शामिल हैं। वर्षा का पूर्वानुमान एक प्रेडिक्टिव न्यूरल नेटवर्क मॉडल का उपयोग करता है, और आउटपुट की तुलना पूर्वानुमानित अवधि के लिए वास्तविक समय में देखे गए वर्षा के आंकड़ों से की जाती है। जांच से पता चला कि LSTM ने आम तौर पर RF की तुलना में बेहतर प्रदर्शन किया। उत्पन्न आउटपुट आशाजनक है और इस प्रकार के अनुप्रयोगों में व्यापक रूप से विस्तारित किया जा सकता है।

ABSTRACT. Weather forecasting is an important and challenging attribute to predict because most atmospheric and agricultural fields depend on day-to-day weather fluctuations. Rainfall is one of the most important parameters dependent on various climatic conditions. We used the Random Forest (RF) and Long Short-Term Memory (LSTM) Neural Network models to predict rainfall for the major cities of Uttar Pradesh in the months of June, July, August, and September (JJAS) from 1901 to 2020. We used various statistical indices like the correlation coefficient (CC), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) to assess the quality of the forecast. Several climate indices were used as predictors to forecast rainfall as mentioned above. These indices include the North Atlantic Sea Surface Temperature, Nino 3.4, Equatorial south-east Indian Sea, and rainfall at a lag value of 12. The prediction of rainfall utilizes a predictive neural network model, and the output is compared to real-time observed rainfall data for the forecasted period. The investigation revealed that the LSTM generally performed better compared to the RF. The generated output is promising and can be widely extended in this type of application.

Key words – Machine Learning, Deep Learning, ANN, Prediction of Rainfall.

1. Introduction

Rainfall poses a challenge within the climatic system and is difficult to predict. However, it directly influences Earth's ecosystems, water resources, and agricultural production. India, with its agrarian economy, heavily relies on the behavior of Indian summer monsoon rainfall (ISMR) from June to September. It is a well-known fact that approximately 65% of India's cultivated land depends on agriculture (Swaminathan, 1998). Therefore, timely and accurate prediction is crucial not only for the scientific community but also for society. For over a century, the India Meteorological Department (IMD) has provided long-range forecasts of seasonal mean rainfall over India, with varying degrees of success (Rajeevan 2001; Gadgil*et al*., 2005; DelSole and Shukla, 2012).

IMD also utilizes Coupled General Circulation Models (CGCMs) for dynamical prediction of ISMR. However, leading forecasting centers in India and abroad have shown systematic biases in the prediction of the Asian summer monsoon on various time scales (Kim *et al.*, 2013; Liu *et al.*, 2012; Sharmila et al., 2013; Sahai*et al.*, 2013; Goswami*et al.*, 2014; Krishnamurthy, 2017, 2018; Shahi*et al.*, 2018, 2021), including CMIP runs (Ramesh and Goswami, 2014). The forecast skill for South Asian summer monsoon prediction has further decreased in recent years due to the models' inability to capture recent teleconnection patterns in the Pacific Ocean (Wang *et al.*, 2015). Hence, accurately predicting the monsoon on various time scales remains a challenge for the scientific community.

Artificial Neural Networks (ANNs) have been extensively used in recent years for non-linear function approximation and modeling of complex, dynamical phenomena (Hertz *et al.*, 1991; Masters, 1993; Hung et al., 2008; Hossain*et al.*, 2020). ANNs are particularly useful in disciplines where intrinsic non-linearity in dynamics hinders the development of solvable models. Elsner and Tsonis (1992, 1993) demonstrated that multi-layer feed-forward (FF) neural networks outperformed linear statistical models when dealing with chaotic and random noise systems. In the context of ISMR prediction, Goswami and Srividya (1996) used a time series approach, while Venkatesan*et al.* (1997) utilized a predictors approach using teleconnection parameters. Navone and Ceccatto (1994) employed both approaches and proposed a hybrid model.

In recent years, Machine Learning (ML) and Deep Learning (DL) approaches have been widely used by researchers for ISMR prediction (Sahai*et al.*, 2000; Aksoy and Dahamshe, 2008). Venkatesan*et al.* (1997) demonstrated that neural networks outperformed a linear model in predicting ISMR. The ML has also been successfully employed in creating early warning systems for predicting short-term heavy rainfall (Moon *et al.*, 2019). Comparisons between linear regression, neural networks, and support vector machines for predicting surface wind predictors have shown improvements with non-linear prediction methods (Mao and Monahan, 2018). The DL approaches have also shown potential in reducing the uncertainty of weather forecasts compared to traditional artificial neural network methods (Scher&Messori, 2018).

The DL approaches are commonly used for accurate reproduction of nonlinear systems, capturing noise complexity in datasets, thus enhancing the prediction of non-linear systems. Long Short-Term Memory (LSTM) and Random Forest (RF) are widely used learning-based methodologies for rainfall prediction (Zhang *et al.*, 2018). LSTM, an improved Recurrent Neural Network (RNN) architecture (Poornima et al., 2019), excels at modeling time-series data and predicting future values (Hochreiter*et al.*, 2001).

The Models have been developed by combining wavelet packet decomposition, ML, ANN, and SVM models to predict daily river stage and analyze their performance metrics (Seo*et al.*, 2016). RF and Support Vector Machine predictive machine learning models have incorporated hyperspectral reflectance vegetation indices and day of the year predictors to estimate predawn leaf water potential for determining water stress in grapevines (Pôças*et al.*, 2017). Regression analysis, a statistical method used to establish relationships between dependent and independent variables, has been employed in ML algorithms for groundwater quality parameter prediction (Ewaid*et al.*, 2018). Spectral indices, multivariate approaches, and neural network techniques have been used to predict the relative water content, a water stress indicator, under water deficit stress conditions of rice genotypes (Krishna *et al.*, 2019). ML techniques, including sea surface pressure and North Central Pacific zonal wind, have shown better results than ANN models for predicting ISMR (Acharya*et al.*, 2019). Previous values of the same variable have been utilized in ANN models to predict ISMR with a reasonable level of accuracy (Sahai*et al.*, 2000). For forecasting rainfall in Gorakhpur, Lucknow, Varanasi, Prayagraj, and Kanpur cities in Uttar Pradesh, India, various DL techniques have been employed. These cities have dense populations, and the region's economy depends on agriculture, which relies entirely on summer monsoon rainfall. Previous studies have used statistical & dynamic techniques, including neural networks built using ML techniques, to predict precipitation. However, other ML techniques can also be utilized for precipitation prediction, potentially yielding better results. This paper proposes an improved ML model using LSTM for prediction & analysis, comparing the results with a testing dataset of rainfall. The accuracy of both ML approaches is analyzed by comparing the proposed LSTM model with the RF model. Predictors such as sea surface temperature from appropriate regions of the Indian, Pacific & Atlantic Oceans are considered which will be explained in the section 2.1 of the present paper.

2. Data and methodology

2.1. *Dataset*

The objective of the present study is to estimate the accumulated rainfall during June - September at five citiesof Uttar Pradesh. We have taken the rainfall gridded data

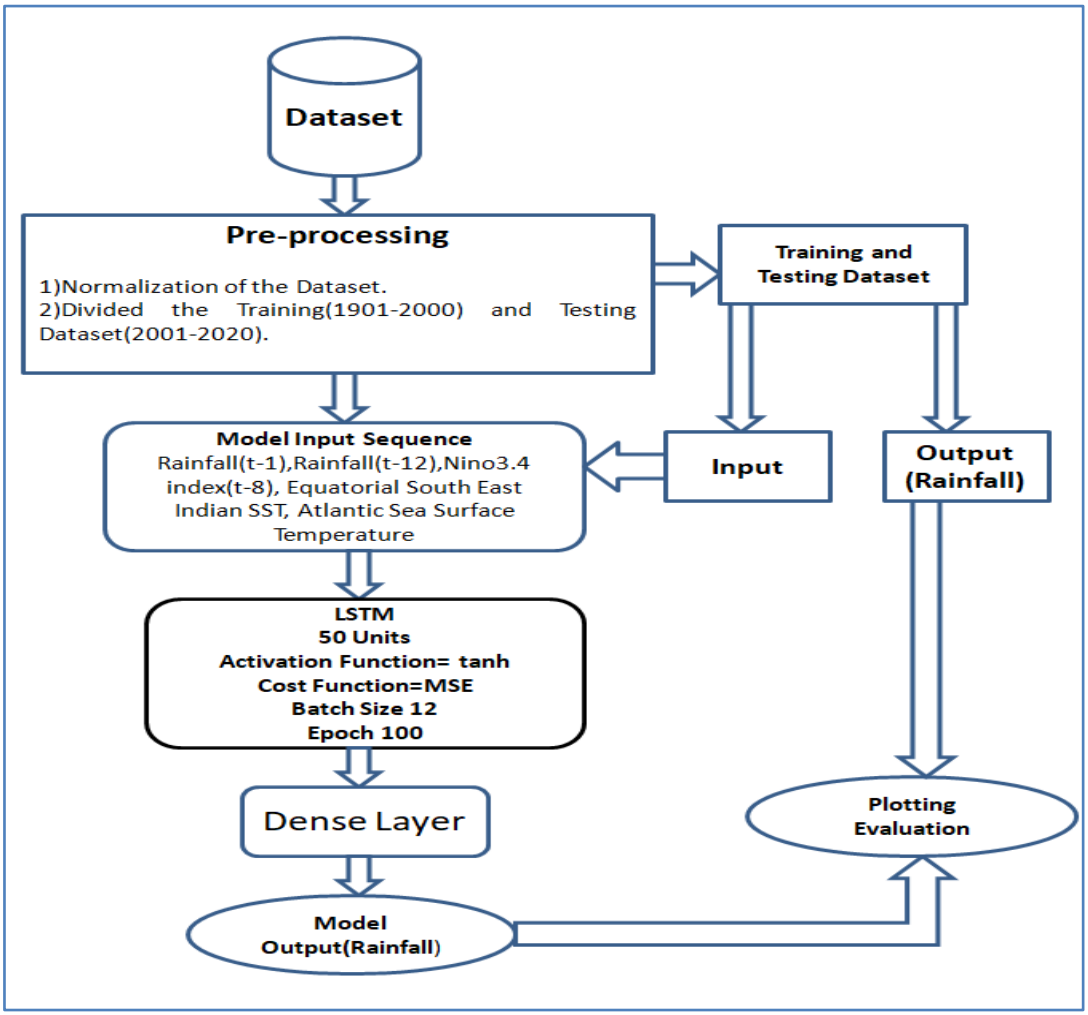


Fig. 1. Flow Chart of the Long Short Term Memory (LSTM) Machine Learning Model

of IMD (Pai*et al.*, 2014). These are daily datasets on a spatial resolution of 0.25° x 0.25°, which were converted to monthly data during the time domain of 1901–2020. Thereafter, we interpolated the data at the grid points mentioned within brackets for the five cities of Uttar Pradesh state in India, namely Lucknow (26.8°N; 80.9°E), Allahabad (25.4°N; 81.8°E), Kanpur (26.4°N; 80.3°E), Gorakhpur (26.7°N; 83.3°E), and Varanasi (25.3°N; 82.9°E). The dataset is separated into two parts for applying the ML technique: (a) a training dataset for the period 1901–2000, and (b) a test dataset for the period 2001–2020.The monthly sea surface temperature (SST) data on a 1° x 1° resolution was used from the Hadley Centre SST dataset (HadSST2) (Rayner et al., 2006) to compute various indices. We computed anomalies of the SST data to compute the North Atlantic Sea Surface Temperature index by taking the area average over the region 20°N-30°N, 100°W-80°W (Rayner*et al.*, 2006). Similarly, the Nino 3.4 index (5°S-5°N, 170°W-120°W) (Rayner*et al.*, 2006) and Equatorial south-east Indian SST (20°S-10°S, 100°E-120°E) (Rayner*et al.*, 2006) were computed by taking area-averaged values over the regions mentioned in the brackets.

2.2. *Long Short Term Memory (LSTM)*

ANN is composed of nonlinear computational components that run in parallel and are structured similarly to real neurons. To complete the learning process in the neural network, a large number of neurons are connected together (Lee *et al.*, 2006). The work performed by hidden units or neurons is not visible to the users. A fully connected network is one in which every node of ANN architecture is connected to every node of the other layers (Haykin, 2009; Aggarwal, 2018). We have used an RNN model, which is similar to the ANN model, but the RNN consists of a Multi-Layer Perceptron (MLP) with an added loop (Hochreiter*et al.*, 1997). The RNN networkhas the power of the nonlinear mapping capabilities of the multilayer perceptron, and it serves as a form of memory

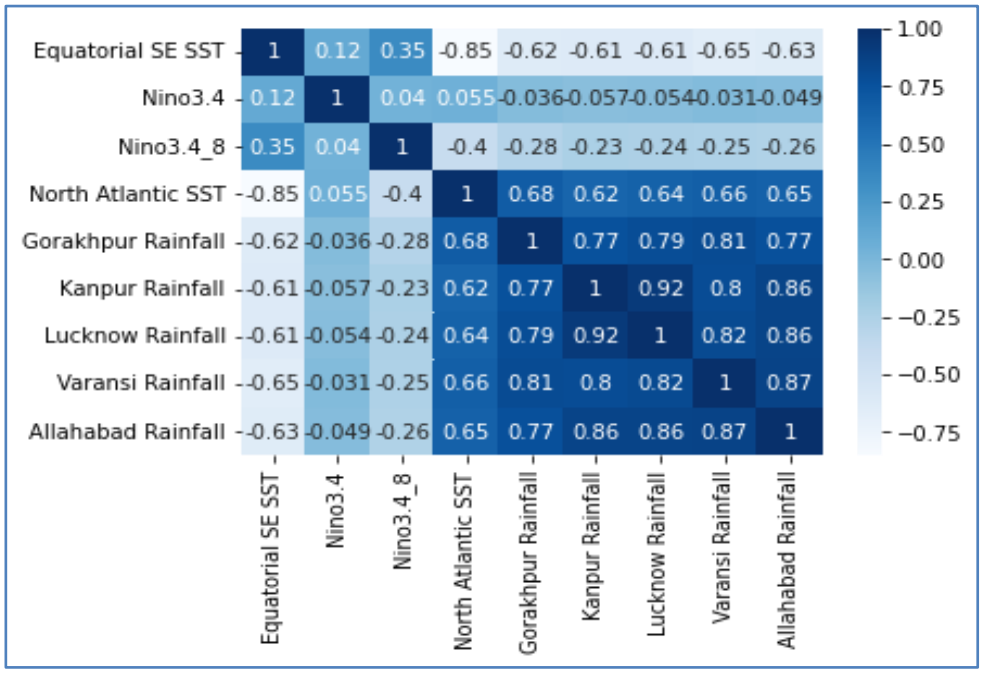


Fig. 2. Heat map of correlations between Predictors and Rainfall

(Mandic*et al.*, 2001a, 2001b). This design replaces the traditional hidden layer mode with memory cells (Jozefowicz*et al.*, 2015). The memory cells can store, compose, and read information through logical gates, similar to how information is stored in system memory (Zhao *et al.*, 2017).LSTM is a nonlinear model that has been used to forecast sequence patterns in music and text. Additionally, LSTM can be trained to recognize patterns in sequential data with careful handling and anticipation of what comes next (Graves *et al.*, 2009). Unlike RNN, LSTM contains special units called memory blocks in the recurrent hidden layer. The memory blocks include unique multiplicative units called gates to regulate the information flow, as well as memory cells with self-connections that store the network's temporal state. The mathematical equation of the LSTM model is given as (Hochreiter*et al.*, 1997):

$$i_t = \sigma\left(W_{xt}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (2)$$

$$c_t = \mathrm{i}_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) + f_t c_{t-1} \quad (3)$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \quad (4)$$

$$h_t = \mathrm{o}_t \tanh\left(c_t\right) \quad (5)$$

In equations 1 to 5, i, o, f, and c represent the input, output, forget, and cell gates, respectively, and σ is the tanh activation function, which has the same size as the hidden vector. The weight matrices are denoted by W, while the cell input gate matrix is denoted by Wci. In the memory cell, the flow of input activation is controlled by the input gate, and in the rest of the network, the output gate handles the flow of cell activations. The internal state of the cell is scaled by the forget gate before being added as input through the cell's self-recurrent link, which adaptively forgets or clears the cell's memory. LSTM architecture also includes peephole connections from the interior cells to the gates in the current implementation.

2.3. *Random forest (RF)*

Random Forest (RF) is a machine learning technique that fits a number of decision trees on various subsets of the dataset to improve predictive accuracy and control over-fitting (Breiman. 2001). A tree can be "computed" by dividing the training dataset into subsets based on attribute value tests, and each internal node represents a test on a feature resulting from the division of the current sample (Breiman, 2001; Liaw*et al.*, 2002). At each step, the method selects the feature and a threshold value that maximizes a given metric. Different metrics exist for regression trees, which are quantitative in nature, and classification trees, where the output is qualitative (Abdel-Rahman*et al.*, 2013). The recursion is completed when the subset at a node has the same value as the output variable or when no further division improves the predictions. This

general method is employed by many recursive partitioning tree algorithms (James *et al.*, 2014).

2.4. *Measure the performance of Model*

This study uses different statistical indices (such as root mean square error; RMSE, correlation coefficient; CC and mean absolute error; MAE to examine model potential.

$$RMSE = \sum_{i=1}^{n} \frac{(predicted - actual)^2}{n} \quad (6)$$

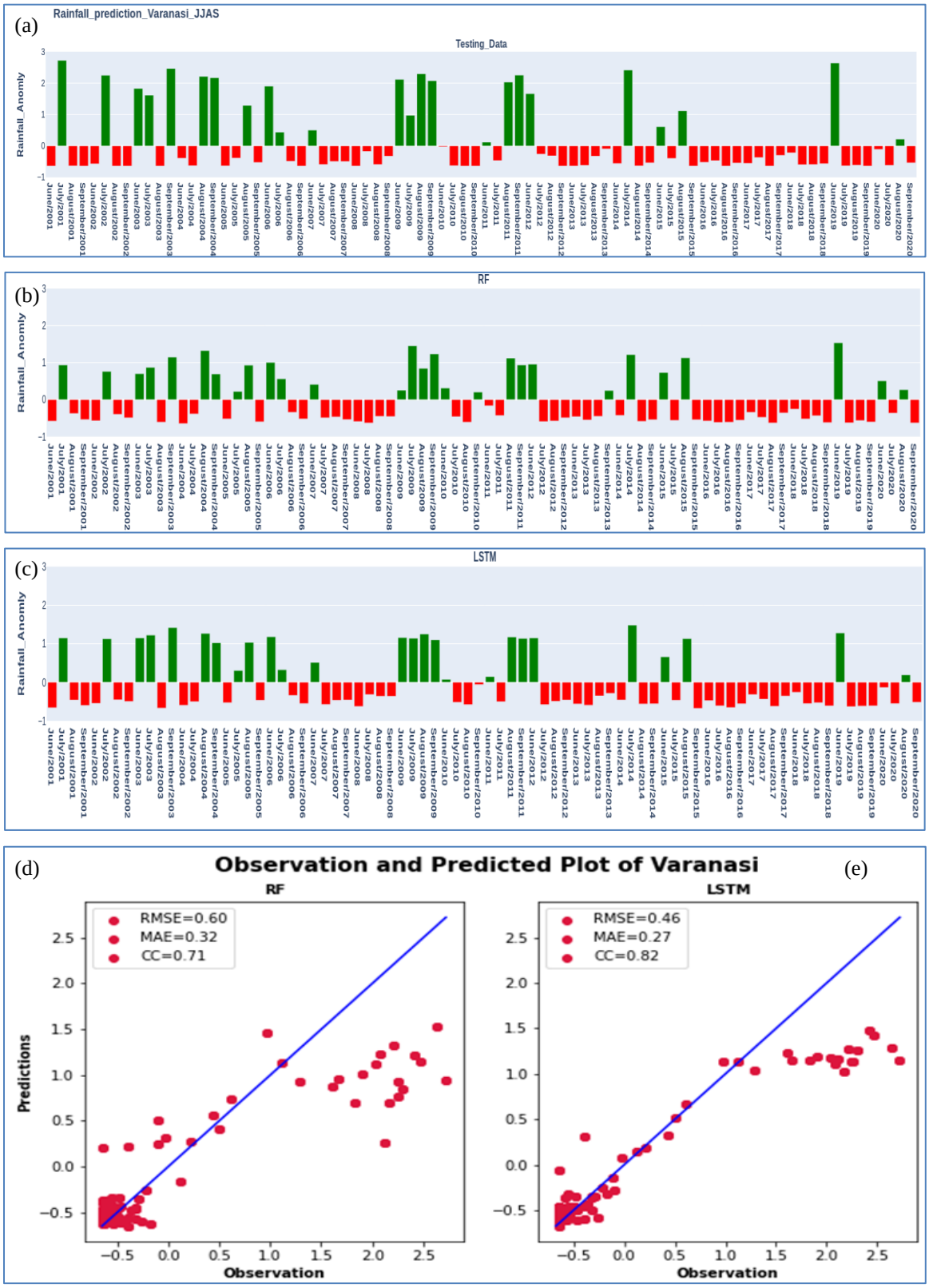
$$MAE = \frac{\sum_{t=1}^{n} | predicted_t - actual_t |}{n} \quad (7)$$

$$CC = \frac{\sum_{t=1}^{n} (actual_t - actual\_mean_t)(predicted_t - predicted\_mean_t)}{\sqrt{\sum_{t=1}^{n} (actual_t - actual\_mean_t)^2 \sum_{t=1}^{n} (predicted_t - predicted\_mean_t)^2}} \quad (8)$$

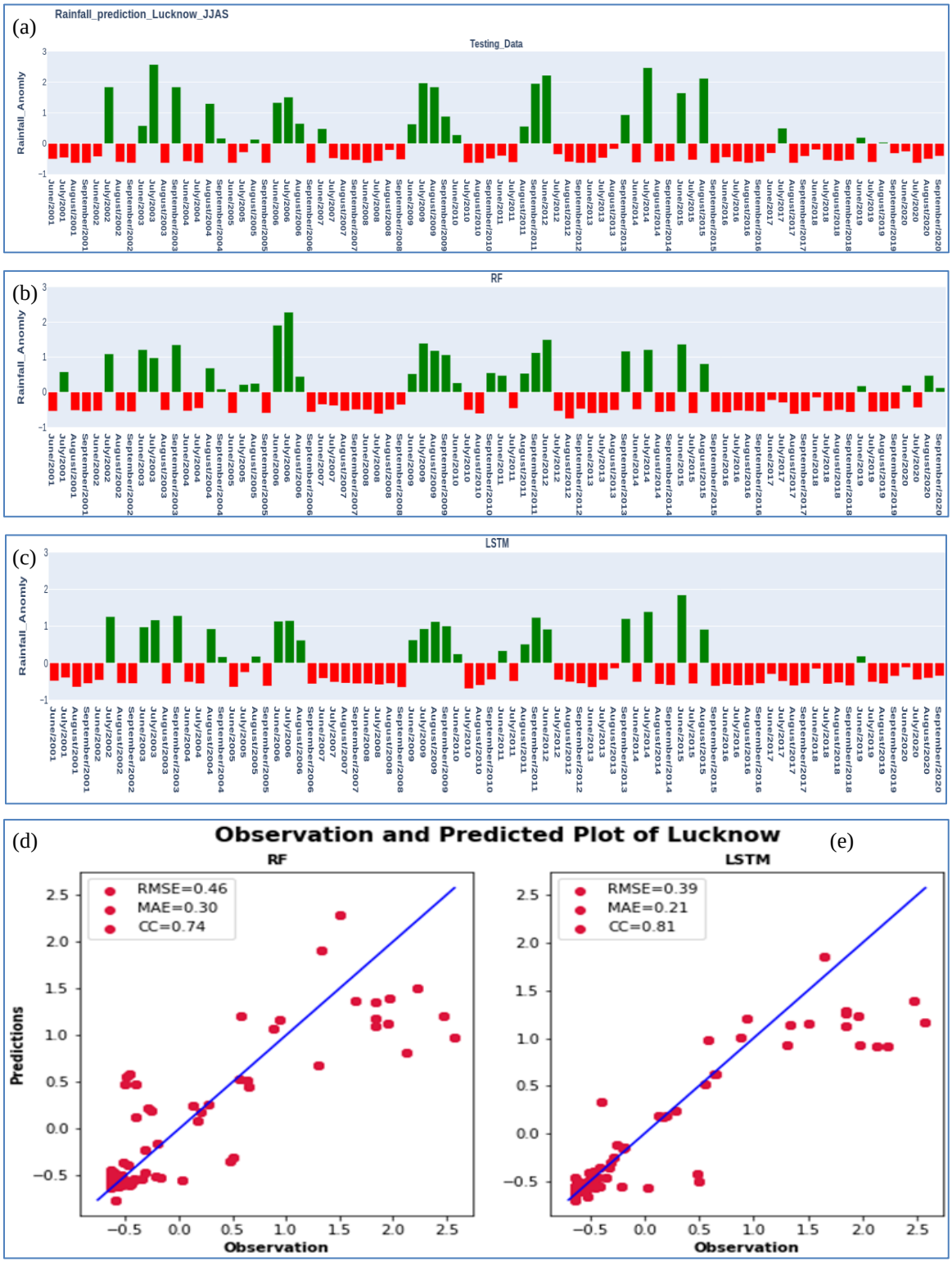
where n is a number of dataset.

3. Results and discussion

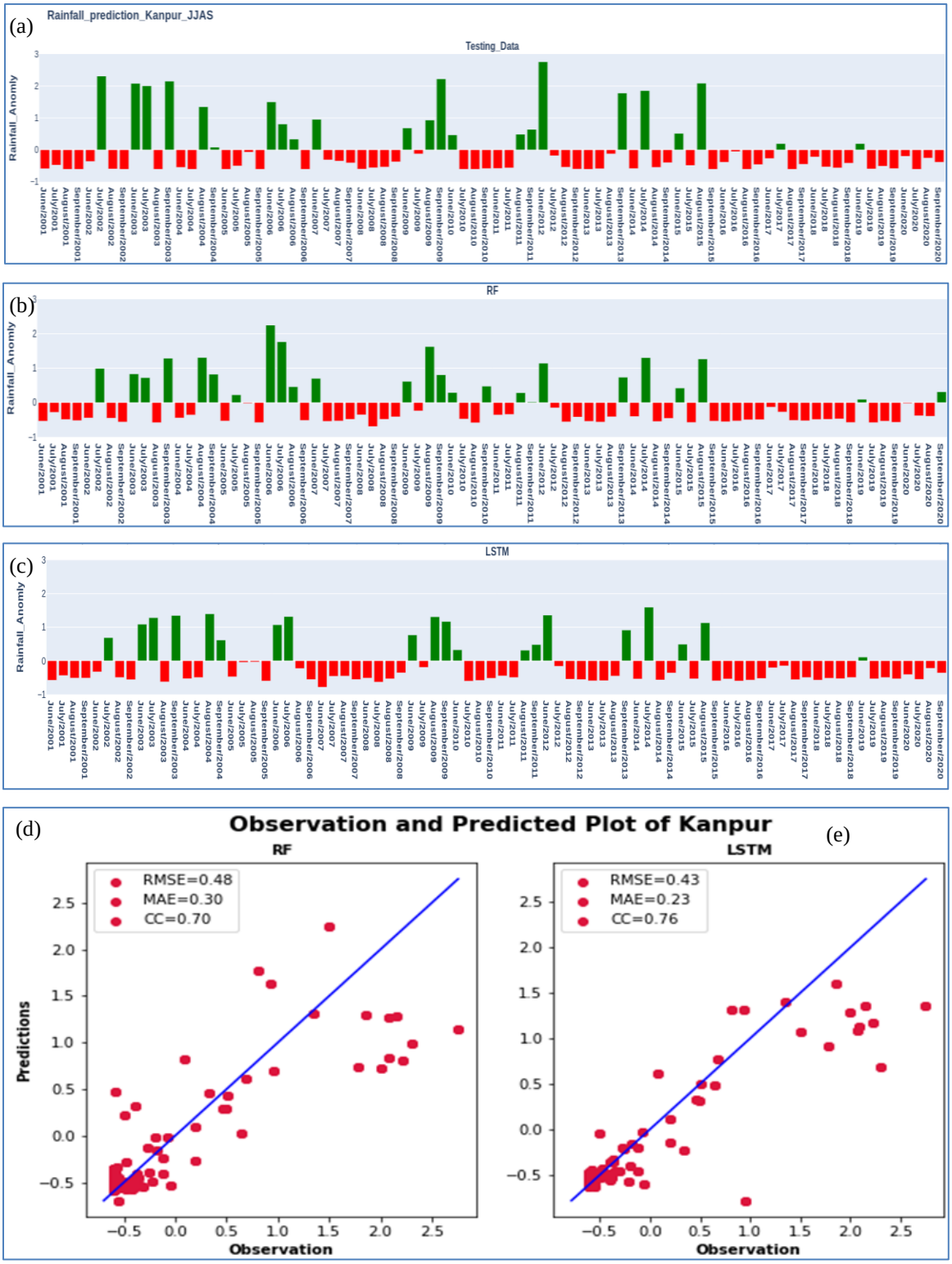
In this paper, we used a rainfall dataset from Allahabad, Kanpur, Varanasi, Lucknow, and Gorakhpur, which are five major cities in central India. We employed Random Forest (RF) and Long Short-Term Memory (LSTM) machine learning techniques to predict rainfall in these cities during the period from 1901 to 2020. The entire dataset has been divided into training and testing datasets for each of the mentioned cities. We used the training dataset from 1901 to 2000 and the test dataset from 2001 to 2020, on a monthly time period, for the proposed RF and LSTM models. We included input features, namely the North Atlantic SST, NINO3.4 index, and Equatorial south-east SST, which are defined in section 2.1 above. These five meteorological stations' rainfall variability is predicted using a variation of climatic indices in the SST parameters of the Indian, Pacific, and Atlantic Oceans. The lagged values of these indices have been used as potential predictors to create the dataset. The lagged rainfall data from t-1 to t-12 is added an input features, where t represents the current month value. To increase the correlation between the Nino index and rainfall, we used a time lag value of eight months for the Nino 3.4 index with respect to rainfall. The same input variable and its associated output variable from the rainfall dataset are used in both the RF and LSTM models. In Fig. 1 illustrates the workflow of the LSTM-based rainfall prediction model developed using monthly dataset. The dataset includes rainfall observations from five major cities in Uttar Pradesh along with key oceanic predictors such as the lagged NINO3.4 index (8-month lag), Equatorial South-East SST, and North Atlantic SST. In the preprocessing stage, the data is normalized and split into training (1901–2000) and testing (2001–2020) periods. The model input consists of a sequential time series that includes 12 months of lagged rainfall data and the selected climatic indices. These sequences are fed into an LSTM network config.d with 50 units, using the tanh activation function, mean squared error as the loss function, a batch size of 12, and 100 training epochs. The LSTM's output is passed through a dense layer to produce rainfall predictions, which are then compared with actual observations for evaluation. This framework leverages the LSTM model's ability to capture long-term dependencies in sequential data, making it suitable for modeling the delayed and nonlinear influences of ocean-atmosphere interactions on regional rainfall. The interrelation between predictors and rainfall was initially assessed using linear correlation analysis, as shown in Fig. 2, to detect statistically significant associations between lagged sea surface temperature (SST) indices and monthly rainfall across five key cities in Uttar Pradesh. This correlation matrix served as a preliminary feature screening tool, allowing us to identify predictors with meaningful linear associations that are also physically interpretable within the monsoonal context—such as the positive link between North Atlantic SST and rainfall, and the negative correlation with the lagged NINO3.4 index. While linear correlation does not capture nonlinear or complex temporal dynamics, its use at this stage helps in dimensionality reduction and strengthens the physical justification for feature inclusion. The correlation table in Fig. 2 highlights key physical relationships between oceanic SST indices and rainfall over five major cities in Uttar Pradesh, reflecting how large-scale ocean-atmosphere interactions influence regional precipitation. The strong negative correlation between the Equatorial SE SST and the North Atlantic SST ($r = -0.85$) suggests a compensatory thermal relationship between these ocean basins, which can influence global circulation patterns such as the Walker and Hadley cells. The positive correlation between the North Atlantic SST and rainfall ($r \approx 0.62–0.68$) indicates that warmer Atlantic SSTs enhance evaporation and contribute to increased atmospheric moisture, which is transported toward the Indian subcontinent by monsoonal low-level jets, enhancing rainfall (Saha et al., 1981). Conversely, the NINO3.4 index, particularly with an eight month lag, exhibits a weak to moderate negative correlation with rainfall ($r \approx -0.23$ to -0.28), consistent with the delayed suppressive effects of El Niño events. These events weaken the Indian monsoon through anomalous subsidence and altered convection patterns. The negative correlation between the



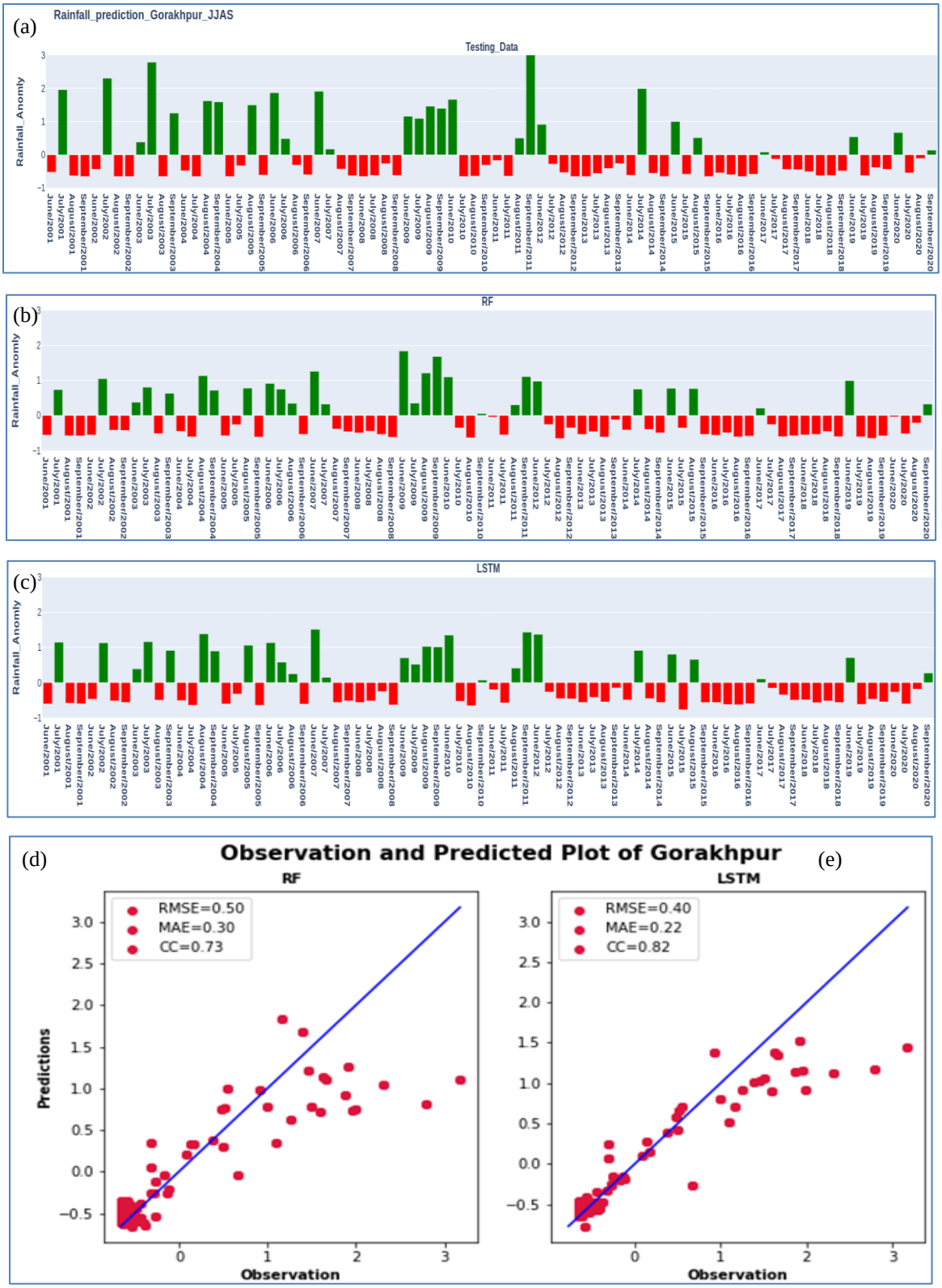
Figs. 3(a-e). Rainfall dataset of Varanasi city during 2001-2020 in JJAS months for (a) Observed (b) Predicted using RF model (c) Predicted using LSTM model (d) Observed vs. predicted scattered plot for RF model and (e) Observed vs. predicted scattered plot for the LSTM model



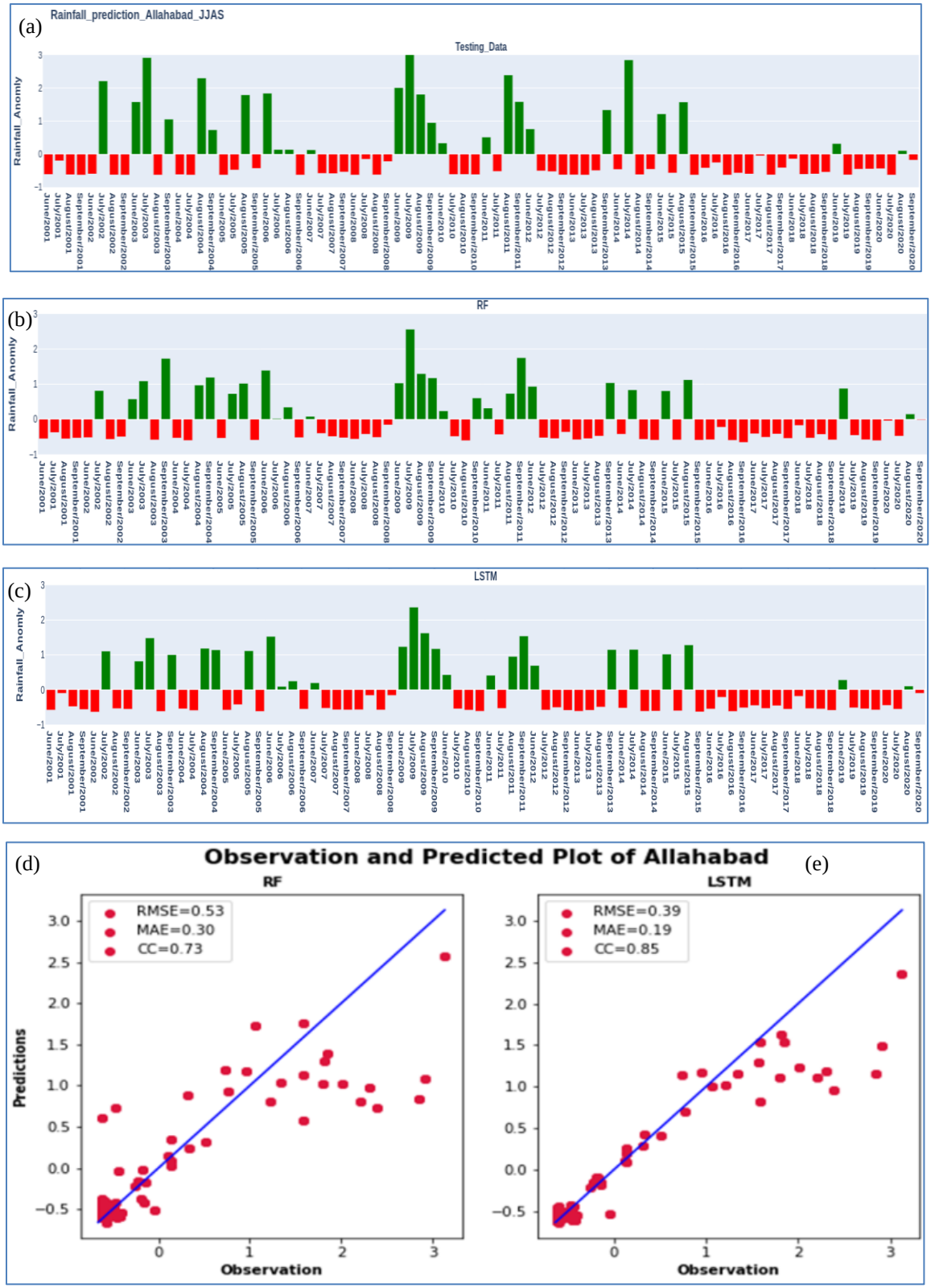
Figs. 4(a-e). Rainfall dataset of Lucknow city during 2001-2020 in JJAS months for (a) Observed (b) Predicted using RF model (c) Predicted using LSTM model (d) Observed vs. predicted scattered plot for RF model and (e) Observed vs. predicted scattered plot for the LSTM model



Figs.5 (a-e). Rainfall dataset of Kanpur city during 2001-2020 in JJAS months for (a) Observed (b) Predicted using RF model (c) Predicted using LSTM model (d) Observed vs. predicted scattered plot for RF model and (e) Observed vs. predicted scattered plot for the LSTM model



Figs. 6(a-e). Rainfall dataset of Gorakhpur city during 2001-2020 in JJAS months for (a) Observed (b) Predicted using RF model (c) Predicted using LSTM model (d) Observed vs. predicted scattered plot for RF model and (e) Observed vs. predicted scattered plot for the LSTM model



Figs. 7(a-e). Rainfalldataset of Prayagraj city during 2001-2020 in JJAS months for (a) Observed (b) Predicted using RF model (c) Predicted using LSTM model (d) Observed vs. predicted scattered plot for RF model and (e) Observed vs. predicted scattered plot for the LSTM model

Equatorial SE SST and rainfall ($r \approx -0.60$) further supports this, as warm anomalies in this region may disrupt zonal wind patterns at 850 hPa and shift convective zones eastward, thereby reducing moisture convergence over India (Acharya et al., 2019). These physically consistent correlations justify the selection of these predictors and validate their relevance in driving the variability of rainfall in the study region.

The graph between the actual and predicted data for Varanasi city is shown in Fig. 3. Both the RF and LSTM models accurately predicted the results for the test dataset for the JJAS season from 2001 to 2020 (Fig. 3.b and 3.c). The performance of the RF model was inferior to the LSTM model in some months. Fig. 3b and 3d clearly show that the RF model predicts well when the rainfall value is less than one, but when the value is greater than one, the predicted rainfall values are scattered compared to the actual dataset. However, the LSTM model's predictions were close to the observed data with a reduction in error, even when the value of rainfall is greater than 1 (Fig. 3.c and 3.e). The correlation coefficient (CC), mean absolute error (MAE), and root mean squared error (RMSE) of the RF model for the Varanasi city are 0.71, 0.32, and 0.60, respectively. Additionally, the CC, MAE, and RMSE of the LSTM model are 0.82, 0.27, and 0.46, respectively.

The actual and predicted values of the Lucknow city are shown in Fig. 4 for each prediction model. In Fig. 4.b, the testing dataset for the months of July 2001, September 2010, June 2020, and September 2020 have negative actual rainfall anomaly values, but the RF model predicted positive rainfall anomaly values. However, in Figs. 4(c&e), the LSTM model has followed the same pattern as the actual rainfall anomaly dataset. The RMSE, MAE, and CC of the RF and LSTM models for the Lucknow city are 0.46 (0.39), 0.30 (0.21), and 0.74 (0.81), respectively.

The actual and predicted rainfall data from June 2001 to September 2020 for the RF and LSTM models are represented in Fig. 5. It is observed that the RF model's negative and positive rainfall trends have not followed the same pattern as the actual dataset in some months (Fig. 5b). However, the LSTM model performs better in following the trends in the months of July 2005, September 2010, and September 2020 (Figs. 5c&e). The CC, RMSE, and MAE for the RF and LSTM models are 0.70, 0.48, 0.30, and 0.76, 0.43, 0.23, respectively, for Kanpur city.

Fig. 6 represents the actual and predicted values for the Gorakhpur city. In Fig. 6(b &.c), we can see that when the actual rainfall anomaly value (in the months of July 2001, July 2003, and September 2011) is greater than two, the RF and LSTM models do not accurately predict the value. However, it is clearly visible from Fig. 6.c that the LSTM model performs better in identifying the actual rainfall dataset. The LSTM model displays the majority of its data points on a linearly fitted line (Fig. 6.e). Furthermore, the number of data points above the line does not fit neatly in the RF model (Fig. 6.d). The CC, MAE, and RMSE of the RF and LSTM models for the Gorakhpur city are 0.73, 0.30, 0.50, and 0.82, 0.22, 0.40, respectively.

The actual and predicted plots for the meteorological station of Allahabad are shown in Fig. 7. In Fig. 7.b, the RF model seems particularly less reliable when predicting high rainfall values (greater than two mm), and it also has visibly higher errors, as shown in the performance metrics. In Fig. 7.d and 7.e, we can observe that the dispersion in the test predictions obtained by the LSTM model has a better-centered regression line compared to the RF method. This implies that its predictions are more balanced across different values and generally more consistent. The RMSE, MAE, and CC of the RF and LSTM models for Allahabad are 0.53 (0.39), 0.30 (0.19), and 0.73 (0.85), respectively.

We carried out comparable feature engineering and selection in the RF models. We kept the RF model with five trees and five branches depth. The model assigned weights to the characteristics and displayed the corresponding performance. It is noteworthy that the above tree-based models show considerable performance even with the limited depth of five or fewer branches. The RMSE, MAE, and CC results demonstrate that the LSTM-based RNN prediction model provided better results compared to the RF model. The superior performance of the LSTM model is due to its complex architecture that allows the network to retain layer information from past inputs for long storage, which can be used in subsequent training (Ahmed *et al.*, 2010; Makridakis*et al.*, 2018). This makes the model suitable for time series applications where current values depend on past records.

All the findings, comparisons, and validation of the suggested prediction model are covered in this part. The standard parameters for prediction errors are RMSE, MAE, and CC. MAE is computed as a performance evaluation for the suggested strategies' reduced prediction error. The RMSE value represents the variance between the properties of the actual and forecasted rainfall.

4. Conclusions

In this study, meteorological time series rainfall data were used as input for the five cities (Allahabad, Kanpur,

Varanasi, Lucknow, and Gorakhpur) of Uttar Pradesh with a twelve-month lag. Additionally, input features were added, namely North Atlantic SST, NINO 3.4 index, and Equatorial south-east SST, during the period of 1901 to 2020 on a monthly basis. The selection of suitable lags was based on co-linearity and correlation analysis, as mentioned in the first paragraph of Section 3. The primary purpose of this study was to compare the capability of ML/DL methods in predicting rainfall in the major cities of Uttar Pradesh. Two ML/DL methods, RF and LSTM, were employed to predict rainfall using input variables such as rainfall with a twelve-month lag and SST over the Indian, Pacific, and Atlantic oceans with appropriate lags from 1901 to 2020. The test results demonstrated that the methods were well-trained. Three performance indicators, namely MAE, RMSE, and CC, were utilized as performance indexes, and they showed that the LSTM model outperformed the RF model. The next step involved predicting rainfall for the testing period from 2001 to 2020. As expected, the MAE, RMSE, and CC results revealed that RF had lower accuracy in predicting rainfall compared to LSTM.

From this study, it was observed that the LSTM model performed well for time series data compared to the RF model, which was not as effective in handling nonlinear data. However, it was also noted that while the LSTM model was better than the RF model in predicting high rainfall values (greater than two mm), there were still discrepancies in the LSTM models for higher rainfall values. Further investigation is needed to enhance the LSTM model and improve its accuracy for high rainfall values. Nonetheless, our proposed method serves as an efficient rainfall forecasting model that can aid the agricultural sector and contribute significantly to the nation's economy.

Acknowledgements

The authors thankfully acknowledge the respective agencies of the IMD and HadSST2 for making these datasets available. MP will like to acknowledge the funding from University Grants Commission (UGC), Ministry of Education, New Delhi for fellowship to conduct PhD study. SR. would like to thank DST for providing grant through the FIST scheme vide sanction order no. SR/FST/ES-I/2017/5.

Author's Statement

The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Data Availability

All relevant data are available in the paper.

Funding

This material was not funded by any organization. All the costs of research works are undertaken by authors itself.

Authors' Contributions

Shailendra Rai: Analysed the data and prepared the manuscript.(*email:shailendrarai@allduniv.ac.in*).

Disclaimer: The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Abdel-Rahman, E. M., Ahmed, F. B., & Ismail, R. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing*, 34(2), 712-728. https://doi.org/10.1080/01431161.2012.713142.

Acharya, R., Pal, J., Das, D., &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.1766Aggarwal, C. C. (2018). Neural networks and deep learning.Springer, 10(978), 3 https://doi.org/10.1007/978-3-031-29642-0.

Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. 2010. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews. 29(5-6):594–621. https://doi.org/10.1080/07474938.2010.481556.

Aksoy, H. and Dahamshe, A. (2009) Artificial neural network models for forecasting monthly precipitation in Jordan. Stochastic Environmental Research and Risk Assessment, 23, 917–931. https://doi.org/10.1007/s00477-008-0267-x.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

DelSole, T., &Shukla, J. (2012). Climate models produce skillful predictions of Indian summer monsoon rainfall. Geophysical Research Letters, 39,9. https://doi.org/10.1029/2012GL051279.

Elsner, J. B., &Tsonis, A. A. (1992).Nonlinear prediction, chaos, and noise. Bulletin of the American Meteorological Society, 73(1), 49-60. https://doi.org/10.1175/1520-0477(1992)073%3C0049:NPCAN%3E2.0.CO;2.

Elsner, J. B., &Tsonis, A. A. (1993). Nonlinear dynamics established in the ENSO. Geophysical research letters, 20(3), 213-216. https://doi.org/10.1029/93GL00046.

Ewaid, S. H., Abed, S. A., &Kadhum, S. A. (2018). Predicting the Tigris River a water quality within Baghdad, Iraq by using water quality index and regression analysis. Environmental Technology & Innovation, 11, 390-398. https://doi.org/10.1016/j.eti.2018.06.013.

Gadgil, S., Rajeevan, M., &Nanjundiah, R. (2005). Monsoon prediction–Why yet another failure?.Current science, 88(9), 1389-1400. https://www.jstor.org/stable/24110705.

Goswami, B. B. (2014). Study of Indian Summer Monsoon Intraseasonal Oscillation in MultiscaleModelling Framework (Doctoral dissertation. https://doi.org/10.13140/RG.2.2.24610.95685.

Goswami, P., &Srividya.(1996). A novel neural network design for long range prediction of rainfall pattern. Current Science, 447-457. https://www.jstor.org/stable/24097412.

Graves, A. &Schmidhuber, J., 2009.Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks.Advances in Neural Information Processing Systems 22, NIPS'22, pp. 545-552.https://proceedings.neurips.cc/paper_files/paper/2008/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

Haykin, S. (2009).Neural Networks and Learning Machines. New York: PrenticeHall.https://books.google.com/books?hl=en&lr=&id=ivK0DwAAQBAJ&oi=fnd&pg=PR1&dq=Haykin,+S.+(2009).Neural+Networks+and+Learning+Machines.+New+York:+Prentice+Hall.&ots=8QQ_KfW1pV&sig=KmYk8Rv4tW2Kf8iUXDOgDrpDWro

Hertz, J., & RG, P. (1991). Krogh. Introduction to the theory of neural computation.https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9780429499661&type=googlepdfHochreiter S, Schmidhuber J. Long short-term memory. Neural Comput.1997; 9:1735–80. https://doi.org/10.1162/neco.1997.9.

Hochreiter, S., Bengio, Y., Frasconi, P., &Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. https://ml.jku.at/publications/older/ch7.pdf.

Hossain, I., Rasel, H. M., Imteaz, M. A., &Mekanik, F. (2020). Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western Australia. Meteorology and Atmospheric Physics, 132, 131-141. https://doi.org/10.1007/s00703-019-00679-4.

Hung, N. Q., Babel, M. S., Weesakul, S., &Tripathi, N. K. (2008).An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology and Earth System Sciences Discussions, 5(1), 183-218. https://hal.science/hal-00298924/.

James, G., Witten, D., Hastie, T., &Tibshirani, R. (2013).An introduction to statistical learning (Vol. 112, p. 18). New York: springer. https://doi.org/10.1007/978-3-031-38747-0.

Jozefowicz, R., Zaremba, W. &Sutskever, I., 2015.An Empirical Exploration of Recurrent Network Architectures. Lille, France,, JMLR: W&CP volume 37. https://proceedings.mlr.press/v37/jozefowicz15.html.

Kim, Y., Kim, K. Y., &Jhun, J. G. (2013).The Seasonal evolution mechanism of the East Asian winter monsoon and its interannual variability.Climate dynamics, 41(5), 1213-1228. https://doi.org/10.1007/s00382-012-1491-0.

Krishna, G., Sahoo, R. N., Singh, P., Bajpai, V., Patra, H., Kumar, S., ...&Sahoo, P. M. (2019). Comparison of various modelling approaches for water deficit stress monitoring in rice crop through hyperspectral remote sensing.Agricultural water management, 213, 231-244. https://doi.org/10.1016/j.agwat.2018.08.029.

Krishnamurthy, L., & Krishnamurthy, V. (2017).Indian monsoon's relation with the decadal part of PDO in observations and NCAR CCSM4. International Journal of Climatology, 37(4), 1824-1833. https://doi.org/10.1002/joc.4815.

Krishnamurthy, L., Vecchi, G. A., Yang, X., van der Wiel, K., Balaji, V., Kapnick, S. B., ... & Underwood, S. (2018). Causes and probability of occurrence of extreme precipitation events like Chennai 2015. Journal of Climate, 31(10), 3831-3848. https://doi.org/10.1175/JCLI-D-17-0302.1.

Lee, S., & Evangelista, D. G. (2006).Earthquake-induced landslide-susceptibility mapping using an artificial neural network. Natural Hazards and Earth System Sciences, 6(5), 687-695. https://doi.org/10.5194/nhess-6-687-2006.

Liaw, A., & Wiener, M. (2002).Classification and regression by randomforest. R News 2 (3): 18–22. https://CRAN.Rproject.org/doc/Rnews.

Liu, Y., Wu, G., Hong, J., Dong, B., Duan, A., Bao, Q., & Zhou, L. (2012). Revisiting Asian monsoon formation and change associated with Tibetan Plateau forcing: II. Change.Climate dynamics, 39(5), 1183-1195. https://doi.org/10.1007/s00382-012-1335-y.

Makridakis S, Spiliotis E, Assimakopoulos V. 2018. Statistical and machine learning forecasting. https://doi.org/10.1371/journal.pone.0194889methods: concerns and ways forward. PLoS One. 13(3):e0194889 https://doi.org/10.1371/journal.pone.0194889.

Mandic, D. P., & Chambers, J. A. (2001a).Fundamentals. In S. Haykin (Ed.), Recurrent Neural Networks for Prediction. https://doi.org/10.1002/047084535X.ch2.

Mandic, D. P., & Chambers, J. A. (2001b).Recurrent Neural Networks Architectures.In S. Haykin (Ed.), Recurrent Neural Networks for Prediction. https://doi.org/doi:10.1002/047084535X.ch5.

Mao, Y., & Monahan, A. (2018).Linear and nonlinear regression prediction of surface wind components. Climate Dynamics, 1-19. https://doi.org/10.1007/s00382-018-4079-5.

Masters, T. (1993).Practical neural network recipes in C++. Academic Press Professional, Inc. https://books.google.com/books?hl=en&lr=&id=7Ez_Pq0sp2EC&oi=fnd&pg=PR17&dq=Masters,+T.+(1993).Practical+neural+network+recipes+in+C%2B%2B.+Academic+Press+Professional,+Inc.&ots=e36yqCOriQ&sig=qxxr8bfnzR-iGUt0B_LX3c6c_qE

Moon, S. H., Kim, Y. H., Lee, Y. H., & Moon, B. R. (2019).Application of machine learning to an early warning system for very short-term heavy rainfall. Journal of Hydrology, 568, 1042-1054. https://doi.org/10.1016/j.jhydrol.2018.11.060.

Navone, H. D., &Ceccatto, H. A. (1994).Predicting Indian monsoon rainfall: a neural network approach. Climate Dynamics, 10, 305-312. https://doi.org/10.1007/BF00228029.

Pai, D. S., Rajeevan, M., Sreejith, O. P., Mukhopadhyay, B., &Satbha, N. S. (2014).Development of a new high spatial resolution (0.25× 0.25) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region.Mausam, 65(1), 1-18. https://doi.org/10.54302/mausam.v65i1.851.

Pôças, I., Gonçalves, J., Costa, P. M., Gonçalves, I., Pereira, L. S., & Cunha, M. (2017).Hyperspectral-based predictive modelling of grapevine water status in the Portuguese Douro wine region.International journal of applied earth observation and geoinformation, 58, 177-190. https://doi.org/10.1016/j.jag.2017.02.013.

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

Rajeevan, M. (2001). Prediction of Indian summer monsoon: Status, problems and prospects. Current Science, 1451-1457. https://www.jstor.org/stable/24106570.

Ramesh, K. V., &Goswami, P. (2014).Assessing reliability of regional climate projections: the case of the Indian monsoon. Scientific reports, 4(1), 4071. https://doi.org/10.1038/srep04071.

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M. &Tett, S. F. B. (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. Journal of Climate, 19(3), 446-469. https://doi.org/10.1175/JCLI3637.1.

Saha, K., Sanders, F., &Shukla, J. (1981).Westward propagating predecessors of monsoon depres-sions. Monthly Weather Review, 109(2), 330-343. https://doi.org/10.1175/1520-0493(1981)109%3C0330:WPPOMD%3E2.0.CO;2.

Sahai, A. K., Soman, M. K., &Satyan, V. (2000).All India summer monsoon rainfall prediction using an artificial neural network.Climate dynamics, 16, 291-302. https://doi.org/10.1007/s003820050328.

Sahai, A. K., Sharmila, S., Abhilash, S., Chattopadhyay, R., Borah, N., Krishna, R. P. M., &Pillai, P. A. (2013). Simulation and extended range prediction of monsoon intraseasonal oscillations in the NCEP CFS/GFS version 2 framework. Current Science, 1394-1408. https://www.jstor.org/stable/24092513.

Scher, S., &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.

Seo, Y., Kim, S., Kisi, O., Singh, V. P., &Parasuraman, K. (2016). River stage forecasting using wavelet packet decomposition and machine learning models. Water Resources Management, 30(11), 4011-4035. https://doi.org/10.1007/s10661-018-6768-2.

Shahi, N. K., Das, S., Ghosh, S., Maharana, P., &Rai, S. (2021) Projected changes in the mean and intra-seasonal variability of the Indian summer monsoon in the RegCM CORDEX-CORE simulations under higher warming conditions, Climate Dynamics, 57(5), 1489-1506. https://doi.org/10.1007/s00382-021-05771-3.

Shahi, N. K., Rai, S., Sahai, A. K., &Abhilash, S. (2018). Intraseasonal variability of the South Asian monsoon and its relationship with the Indo–Pacific sea surface temperature in the NCEP CFSv2. International Journal of Climatology, 38, e28-e47. (wileyonlinelibrary.com) DOI: https://10.1002/joc.5349.

Sharmila, S., Pillai, P. A., Joseph, S., Roxy, M., Krishna, R. P. M., Chattopadhyay, R., ...&Goswami, B. N. (2013). Role of ocean–atmosphere interaction on northward propagation of Indian summer monsoon intra-seasonal oscillations (MISO).Climate dynamics, 41(5), 1651-1669. https://doi.org/10.1007/s00382-013-1854-1Swaminathan, M. S. (1998). Padma Bhusan Prof. P. Koteswaram First Memorial Lecture - 23rd March, 3-10. https://scholar.archive.org/work/7tc4fulpkjduzeehwjfax6x4bu/access/wayback/https://mausamjournal.imd.gov.in/index.php/MAUSAM/article/download/1671/1486.

Venkatesan, C., Raskar, S. D., Tambe, S. S., Kulkarni, B. D., &Keshavamurty, R. N. (1997). Prediction of all India summer monsoon rainfall using error-back-propagation neural networks. Meteorology and Atmospheric Physics, 62(3), 225-240. https://doi.org/10.1007/BF01029704.

Wang, B., Lee, J. Y., & Xiang, B., 2015, "Asian summer monsoon rainfall predictability: a predictable mode analysis. *Climate Dynamics*, 44, 61-74. https://doi.org/10.1007/s003820142218-1.

Zhang, D., Martinez, N., Lindholm, G., &Ratnaweera, H. (2018). Manage sewer in-line storage control using hydraulic model and recurrent neural network. Water resources management, 32(6), 2079-2098. https://doi.org/10.1007/s11269-018-1919-3.

Zhao, Z. et al., 2017. LSTM network: A deep learning approach for Short-term traffic forecast. IET Intelligent Transport Systems.11, 68-75. https://doi.org/10.1049/iet-its.2016.0208.