World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Time series forecasting of precipitation patterns over Lucknow region using LSTM

Devasheesh Krishan * and Amrendra Singh

Department of Civil Engineering, Institute of Engineering and Technology, Lucknow -226021, Uttar Pradesh, India.

World Journal of Advanced Research and Reviews, 2023, 20(01), 577–586

Publication history: Received on 31 August 2023; revised on 09 October 2023; accepted on 12 October 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.20.1.2069

Abstract

Rainfall forecasting has assumed an important role in recent times due to uncertainities emanating from climate change as a result of environmental phenomenons like El Nino, La Nina, global warming, etc. Agriculture in India is still pretty much dependent upon rains, more so in a state like Uttar Pradesh. So it is imperative that forecasting systems are developed that can analyse the previous trends of rainfall and predict accordingly the future values of rain. The existing statistical models that forecast rain are too complex and also not cost effective. Hence we take the approach of a machine learning model, or to further specify, a deep learning model called Long Short Term Memory (LSTM) to try and predict with some accuracy. This study examines the precipitation patterns over the Lucknow region for a period of 20 years, with the dates ranging from 1st January, 2000 to 31st December, 2019. The accuracy of the LSTM model developed is judged on the basis of Mean Absolute Percentage Error (MAPE), R-square (R^2) and Root Mean Squared Error (RMSE) values.

Keywords: Precipitation; Forecasting; Global Warming; Deep Learning

1. Introduction

India has a wide range of weather, varying from hot and humid in the south to cold and mountainous in the north, particularly when you talk of the Himalayas. The desert in the west (Thar) plays an important part in bringing moisture laden winds from the Arabian sea that bring rainfall to the mainland in the months of June-October [1].

This is called Monsoon in India, and this rain is a major driver of the agricultural activities in the north region. But this rain is not evenly distributed, and hence some regions suffer recurring droughts while others are affected by severe floods. There's also been a rise in extreme weather events in the years from 1951-2000 [2]. With effects of heavy rainfall varying from losses to infrastructure in case of a flood to stoppages in the road-rail network, the social and economic effects of precipitation cannot be ignored [3]. This leads to a decline in food production, and for a growing economy, we simply can't afford to rely on the unpredictability of rainfall. Therefore it is important that proper predictions are made which can prepare the farmers for the worst. While predicting extreme weather events is out of scope of this paper, prediction of rainfall based on previous trends is tried to be achieved which might be of some benefit in planning for the next crop.

Therefore, as a safety first approach, many studies have tried to investigate and have put forth precipitation forecasting techniques in readiness for any event, whether drought or floods. But in order to increase human mobility activities [4, 5] and increase industrial development and agriculture [6, 7, 8, 9, 10], these approaches must provide timely and efficient forecasts.

This study focuses on leveraging LSTM neural networks to forecast monthly precipitation for the period from 2020 to 2024 using historical data spanning two decades.

* Corresponding author: Devasheesh Krishan

Two dominant approaches are there in rainfall forecasting- 1) Conceptual Modelling and 2) System Theoretical Modelling [11, 12]. The conceptual models are used in hydrologic forecasting as it takes into account the features of the basin. But this approach may not be suitable for prediction of precipitation due to important calibration data of precipitation being not efficiently gathered and that the process of rainfall computations needs advanced numerical methods [13].

The system theoretical methods, on the other hand, are applied by mapping models to explain the relationships between input and output variables without taking into account the physical structure processes. ARMAX is the most popular approach in time series forecasting however it is not able to predict modificationss that aren't based on previous data, especially for the non-linearity of the rainfall correlated variables[14].Recently due to progress in technology, Machine Learning has emerged as a viable tool that can be used in prediction of rainfall. The benefit of using machine learning is that it doesn't even need the factors affecting rainfall to make predictions. It can directly be used to analyse and recognise the previous trends in the historical data and forecast accordingly.

With readily available data and calculating prowess in past few years, machine learning has transformed into an important element of the next generation of time series forecasting models. New machine learning techniques offer a way to understand time and space dynamics in a solely data-driven manner [15]. Deep learning, a part of machine learning, has attained popularity recently due to its ability to recognise images[16], natural language processing[17] and learning patterns[18]. LSTM is a part of deep learning and thus has been used in this paper to forecast precipitation for the period 2020-24.

2. Material and methods

2.1. Study Area

Lucknow, the capital of the second largest state of India (by area) i.e. Uttar Pradesh was chosen as the study region for this particular study. The city is located between latitudes 26^0 44' 08" and 26^0 57' 57" north and longitudes 80^0 49' 50" and 81^0 03' 14" east. Lucknow is the most populous city of Uttar Pradesh, with the last census in 2011 pegging the numbers at around 2.8 million. Bordered in the north by Sitapur and Hardoi, on the east by Barabanki, on the south by Raebareli and on the west by Unnao, Lucknow sits on the north-western shore of the Gomti River.

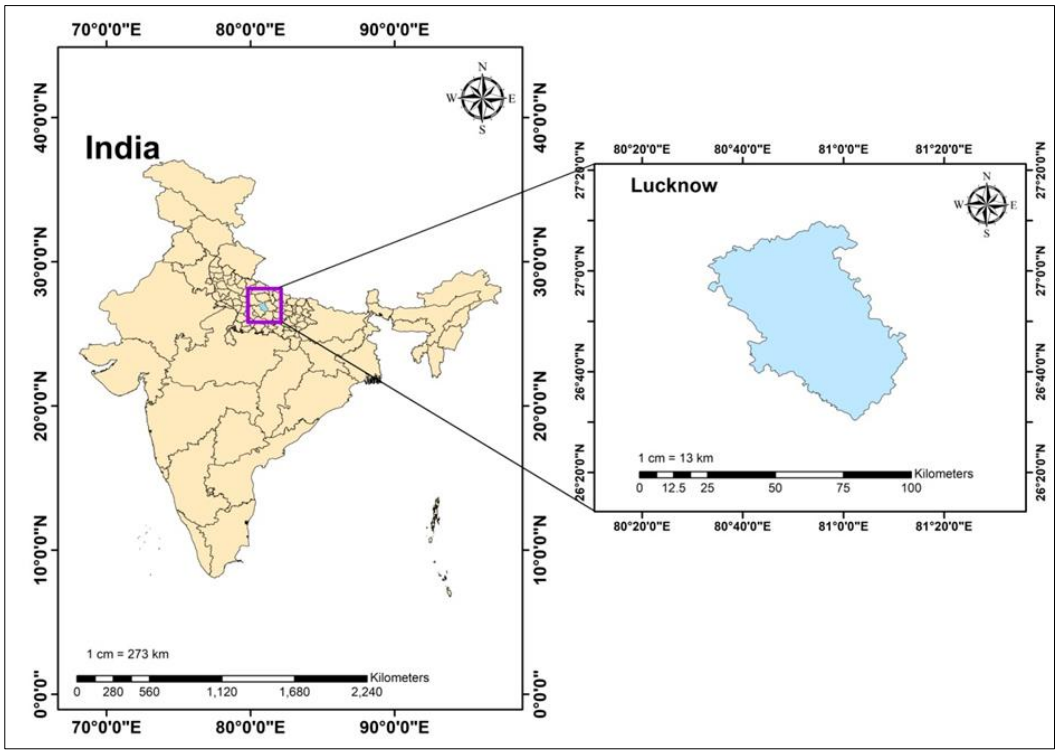

Figure 1 Map area of Lucknow

Lucknow has around 6% of forest cover, which is lower than the state average of 7%.The city is situated at around 123m height from the Mean Sea Level (MSL). Being in a landlocked state and located far from sea, Lucknow has an extreme type of continental climate with the presence of continental air during majority of the year. Only in the four months duration of monsoon (from June to September) does the air from Indian ocean origin is able to reach this region and causes increased humidity, cloudiness and rain. About three-fourths of the total rainfall in Lucknow falls during these four months.

2.2. Data Collection

For this study, data for average monthly precipitation for the period 1stJanuary, 2000 to 31st December 2019 was taken for Lucknow city. It was extracted from Nasa's GIOVANNI. Giovanni is an online instrument that shows Earth science data from NASA straighforwardly on the Internet, without the problems of traditional data procurement and evaluation techniques. Giovanni is a short-form for the Goddard Earth Sciences Data and Information Services Center (GES DISC) Interactive Online Visualization and Analysis Infrastructure.

Giovanni provides access to various satellite data sets, concentrated mainly in the fields of atmospheric composition, atmospheric dynamics, world precipitation, hydrology, and solar irradiance. The satellite data used in this study was provided by Tropical Rainfall Measuring Mission, or TRMM. It was a research satellite launched by JAXA, the Japanese space agency in collaboration with NASA. Originally meant to operate from 1997 to 2015, it's life was later extended. It was designed to better our understanding of the distribution and deviation of precipitation within the tropics as part of the water cycle in the current climatic system. It gives both daily and monthly average precipitation values but we have taken the monthly average precipitation values in this study.

The data is openly accessible and fosters transparency which in turn encourages the aspiring researchers to use this information for their purpose.

2.3. LSTM

In 1997, Long Short Term Memory (or LSTM) was introduced as an alternative to Recurrent Neural Networks (or RNNs) to overcome the problem of vanishing gradient and exploding gradient. It was first proposed by Hochreiter and later by Schmidhuber (Sepp Hochreiter et al.1997;Jurgen Schmidhuber et al.,2000). Since then, LSTM has found applications in different fields like Language modelling, Machine translation, Handwriting recognition, Image generation using attention models, Speech synthesis, Polymorphic music modelling, Question answering, Video-to-text conversion, Image captioning and prediction of protein's secondary structure. Recently, it has also been used in forecasting even the Covid transmission rates in Canada [19].

The LSTM structure consists of 4 neural networks and n number of memory chunks known as cells. Generally, an LSTM unit consists of a cell, an output gate, an input gate, and a forget gate. The three gates manage the flow of info into and out of the cell, and the cell retains the values over arbitrary time intervals.

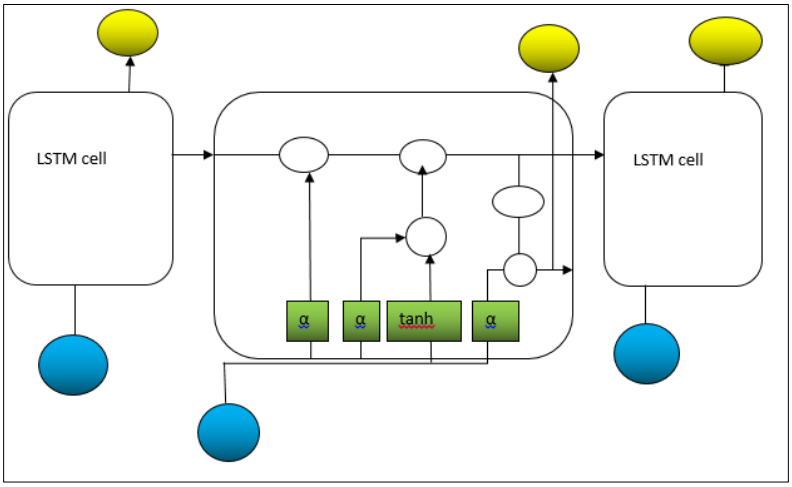

Figure 2 LSTM network

The cells store info while the Gates influence memory. The green boxes are neural networks, blue circles denote input, white circles denote are point-wise operators while the yellow circles represent the cell states. There are 3 Gates:

Input Gate- It confirms the value that was input can be used to alter the memory. The sigmoid function decides if 0 or 1 should be allowed and the tanh function gives the weight to be provided on a scale from -1 to 1.

$$i_m = \sigma(W_s[h_m - 1, x_m] + b_s) \quad \text{.....1}$$

$$C_m = \tanh(W_c\ [h_m - 1, x_m] + \ b_c) \text{..... 2}$$

Forget Gate- It searches for the details that have to be cleared from the block. It is done by a sigmoid function. For every number in the cell state C_m-1, it sees the previous state (h_m-1) and the content input (x_m) and gives a number between 0 (remove this) and 1 (keep it as it is).

$$f_m = \sigma\big(W_f[h_m - 1, x_m] + b_f\big) \text{.....3}$$

Output Gate- The blocks' input and memory are used to determine the output.

$$O_m = \sigma(W_o[h_m - 1, x_m] + b_o) \text{.....4}$$

$$h_m = O_m * \tanh(C_m) \text{ ..... 5}$$

The data we extracted was in time series format. The first step after extracting the data from the site is looking for missing values as the code needs all the values to function properly. We can do that manually but since it is a tedious task, we assign values by backward or forward interpolation to empty places with the help of code. Since no missing values were found in the dataset that we had extracted, we could skip this process.

Then we performed ADF testing. The Augmented Dickey-Fuller (ADF) test is a statistical tool used to identify if a time series is stationary or non-stationary. The interpretation of the test results is based on the ADF Statistic and the p-value. The ADF Statistic is a numerical value that represents the strength of facts against a null hypothesis. In the context of the ADF test, the null hypothesis assumes that if the time series has a unit as a root, then it is non-stationary. A unit root suggests that the series has a stochastic (random) trend, making it non-stationary. The p-value is a probability value associated with the ADF Statistic. It tells us the likelihood of observing an ADF Statistic as extreme as the one computed from our data, assuming that the null hypothesis is true (i.e., assuming the time series has a unit root).

Then, we divided the values into a ratio of 4:1, wherein values of the first 16 years were utilized in training the LSTM model while the last four years' values were used as a benchmark to check the accuracy of the predicted values. The R-square test was used to check if the model is predicting values with good accuracy or not. After that, we proceeded to predict and plot the precipitation values for the next five years based on the previous trends.

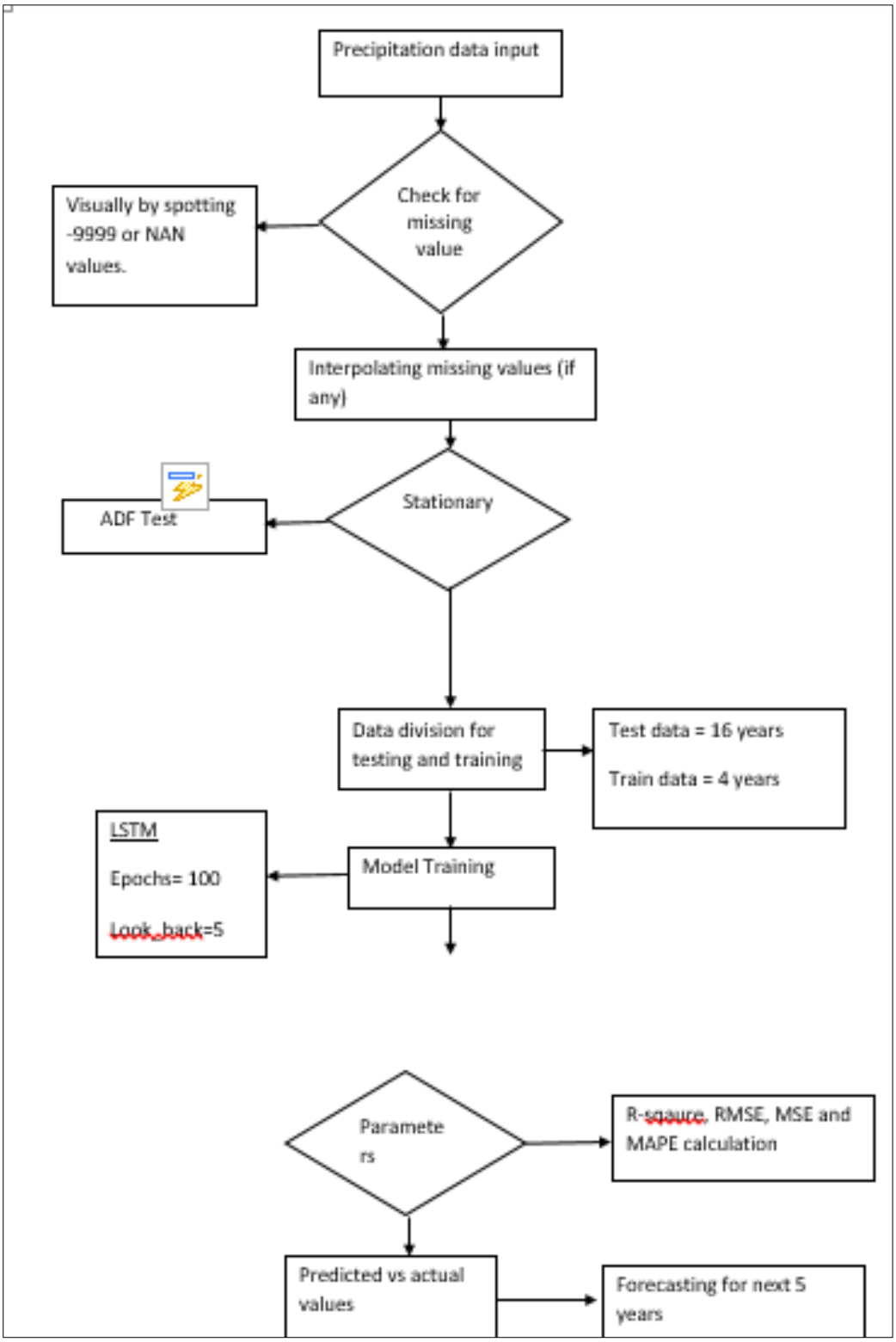

Figure 3 Structure of LSTM

The architecture of the LSTM model is discussed below:

The LSTM model architecture was made of a single LSTM layer with fifty units, followed by a Dense layer with only one output unit. To ensure the model could understand the sequential nature of the data, a 'look_back' parameter was employed, which defined the number of preceding months the model considered when making predictions. This allowed the model to capture temporal dependencies and patterns in the historical precipitation data.

Data pre-processing was a crucial step, involving Min-Max scaling to normalize values within the range of 0 to 1. This normalization aided the model in effectively learning from data with varying scales. Additionally, the inverse transformation of the scaled predictions was performed to retrieve precipitation values in their original units.

The dataset was divided into testing and training subsets to evaluate the model's performance. Training occurred over 100 epochs, with a batch size of 32, facilitating the optimization process. Following training, the model was used to forecast precipitation values for the subsequent five years, encompassing 60 months. Model evaluation was accomplished using the R-squared (R2) score, which quantified the goodness of fit between the model's predictions and the actual data.

Overall, this study showcased the LSTM model's effectiveness in predicting precipitation for a five-year period, leveraging historical average monthly precipitation data from two decades. The model's proficiency in capturing temporal patterns and its ability to generate accurate forecasts position it as a valuable tool in climate and weather-related research and applications.

3. Results and discussion

3.1. Temporal Variations

Daily variations of precipitation were plotted for the period 1 January 2000 to 31st December 2019.

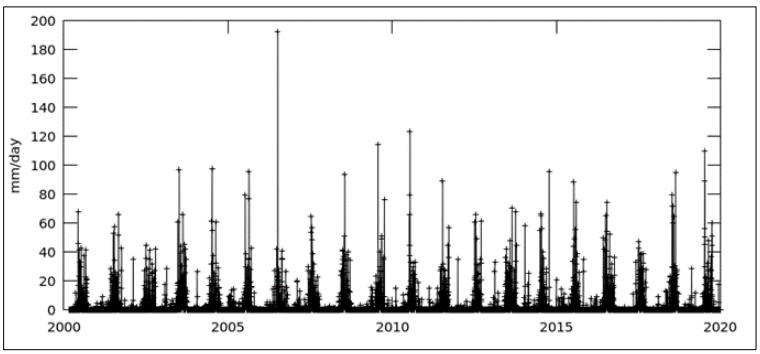

Figure 4 Daily Precipitation plot for Lucknow (2000-2019)

Observing closely, it can be seen that the peak values of rain have fallen considerably in the past few years. Also the number of rainy days seem to have decreased in the recent past. While this is an interesting observation, it is out of scope of this research study so we cannot focus on this much.

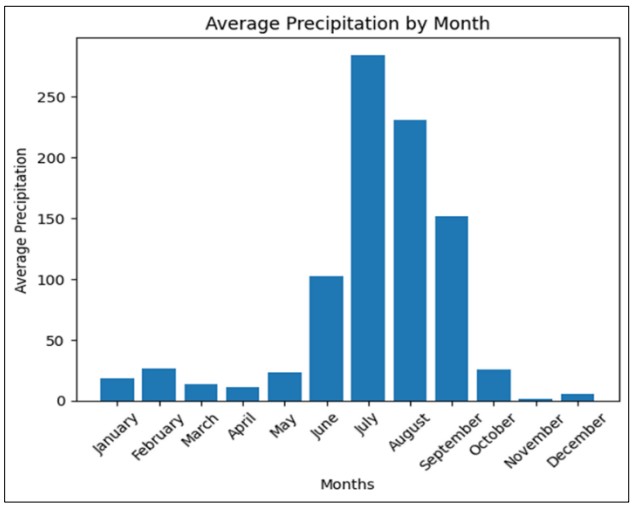

Figure 5 Average Month-wise Precipitation

Next we studied the monthly variations of precipitation. The average precipitation peaks in the monsoon months and there is nothing surprising here. The rainfall season begins in June and continues through July, August and September with July getting the maximum average amount of precipitation which is over 250 mm.

The quarterly analysis further strengthens the influence of monsoon over the precipitation received over the Lucknow region, with the three month span of July- August-September being the quarter receiving the maximum rainfall while October-November-December is the quarter with the least amount of average precipitation.

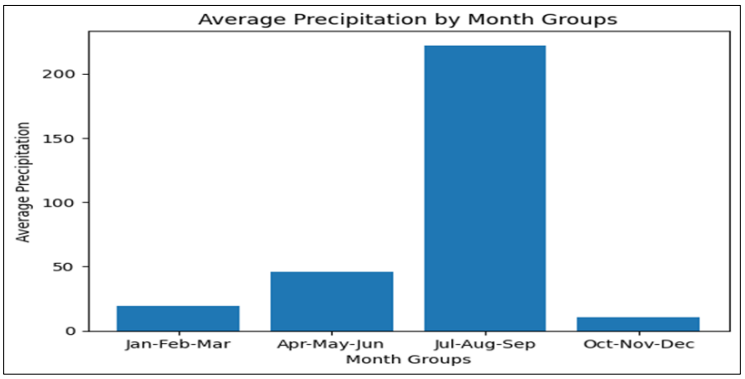

Figure 6 Average Quarterly Precipitation

3.2. ADF test

We used Python to conduct ADF test to check if the time series is stationary or not. We got the following results:

ADF Statistic = -3.25306

p-value = 0.01711

Since the ADF Statistic is a negative value it denotes a more negative statistic which indicates a stronger rejection of the null hypothesis. Hence we can say it suggests that the time series is more likely to be stationary. On the other hand, the p-value is approximately 0.0171 which is less than the common significance level of 0.05. It represents the probability of obtaining the ADF Statistic as severe as, or more severe than, the one observed in our data, given the assumption that the null hypothesis is true. Here the null hypothesis is that the time series is non-stationary.

Taking both into consideration, we can safely conclude that the time series is stationary, which means the null hypothesis stands rejected. In practical terms, this means that the series is not characterized by a significant trend or seasonality that evolves over time. It can be treated as a stationary time series for modelling and analysis purposes.

3.3. Time Series Analysis

The last four years of data from the dataset is used to plot the graph between actual data and the data predicted by code. The result obtained is shown below:

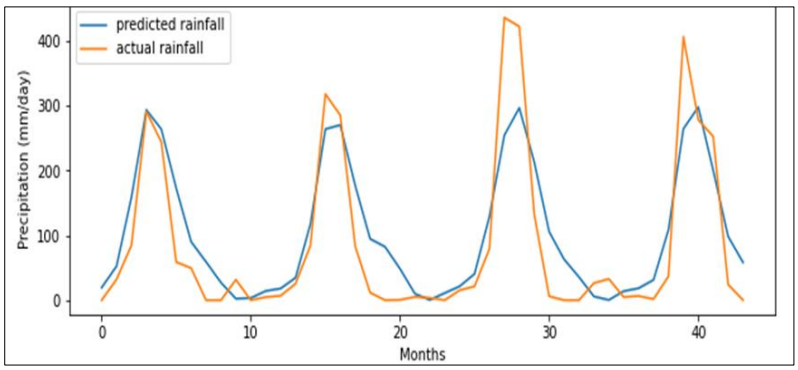

Figure 7 Predicted vs Actual values plot

Since the code performed satisfactorily in predicting the values, we proceeded further to forecast the rainfall pattern for the next five years i.e. from 2020-24. The results obtained are shown below:

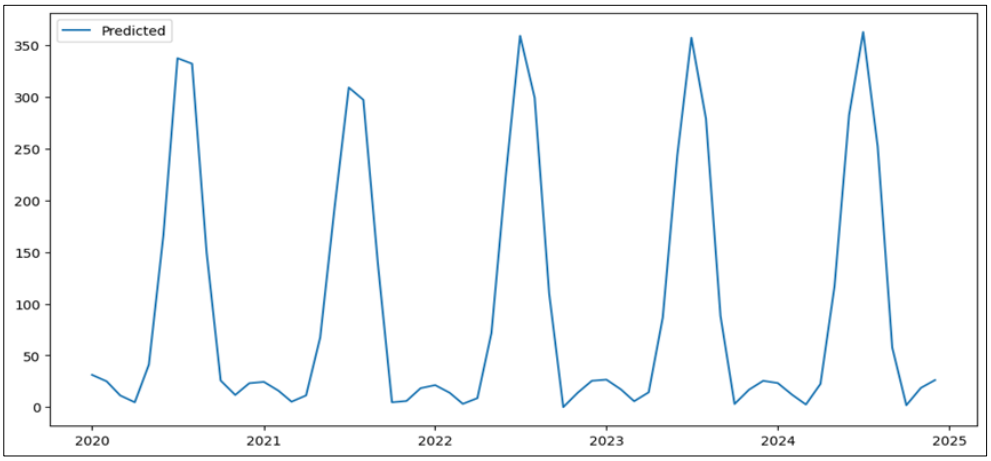

Figure 8 Forecasted Value Plot

3.4. Metrics of performance of LSTM model

The performance of the model was measured by four parameters, namely R-square, RMSE, MSE and MAPE. The values of these parameters are summarized below:

Table 1 Various Parameter values

Parameter	value
R- square	0.627
RMSE	65.808
MSE	4330.801
MAPE	60.019

The model yielded promising results, as indicated by the coefficients of determination (R-squared) value of 0.627. This metric measures the amount of variance in the observed data that is explained by the model's predictions. An R-squared value of 0.627 suggests that the LSTM model accounts for approximately 62.7% of the variability in precipitation, indicating a moderately strong predictive capability.

Additionally, we evaluated the model's performance using root mean squared error (RMSE), mean squared error (MSE), and mean absolute percentage error (MAPE) values. The RMSE value of 65.808 quantifies the average magnitude of errors between predicted and actual precipitation values, with lower values indicating better predictive accuracy. The MSE value of 4330.801 is a measure of the average squared errors, which provides some understanding of the model's precision.

Finally, the MAPE value of 60.019%, which represents the mean absolute percentage difference between predicted and observed values, offers valuable information about prediction accuracy. While a lower MAPE is desirable, a value of 60.019% suggests that the model shows a moderate level of accuracy in predicting monthly precipitation.

In summary, our LSTM model looks promising in predicting precipitation patterns for the next five years, with an R-squared value of 0.627 and acceptable values of RMSE, MSE, and MAPE. Further improvements and validations can enhance its predictive capabilities for practical applications in hydrology and climate science.

4. Conclusion

To wrap it up, this research is a big step in helping us better understand and predict rainfall patterns. We used a Long Short-Term Memory (LSTM) neural model to forecast monthly rainfall for the next 5 years based on 20 years of data from 2000 to 2019.

The LSTM model did pretty well, with an R-squared of 0. 627 and this means it explained about 62.7% of the variation in the rainfall data, so it's decent at capturing complex patterns over time.

We also looked at other metrics like root mean squared error (RMSE) mean squared error (MSE), and mean absolute percentage error (MAPE) to evaluate the model thoroughly. With an RMSE around 65, MSE around 4330, and MAPE of 60% the model seems okay at making reasonably accurate forecasts.

This could be useful for farming, water management, disaster prep and other areas that need good rainfall predictions to plan well. Our LSTM approach seems promising for tackling these problems. But more work is probably needed to improve accuracy. Overall though, this is a solid step toward better understanding and predicting rainfall over time using neural nets.

Looking ahead, this research paves the way for further investigations and refinements of predictive models in the realm of hydrology and climate science. As we continue to confront the complex and evolving challenges posed by changing weather patterns, such models will play an increasingly pivotal role in guiding sustainable and resilient practices for the benefit of society and the environment.

Compliance with ethical standards

Acknowledgments

The authors are thankful to the team at Earth Sciences Data and Information Services Center, or GES DISC, Interactive Online Visualization and Analysis Infrastructure (GIOVANNI) NASA for providing the satellite data. The authors are also thankful to the Department of Civil Engineering, Institute of Engineering & Technology, Lucknow for supporting the work.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chang, J. H. (1967), The Indian Summer Monsoon, Geographical Review, American Geographical Society, Wiley, vol. 57, no. 3, pp. 373–396, doi:10.2307/212640, JSTOR 212640. https://doi.org/10.2307/212640

[2] Goswami, B. N.; Venugopal, V.; Sengupta, D.; Madhusoodanan, M. S.; Xavier, P. K. (2006), "Increasing Trend of Extreme Rain Events over India in a Warming Environment", Science, vol. 314, no. 5804, pp. 1442–1445.https://www.science.org/doi/10.1126/science.1132027

[3] Le, T.-T., Pham, B. T., Ly, H.-B., Shirzadi, A., & Le, L. M. (2020). Development of 48-hour precipitation forecasting model using nonlinear autoregressive neural network. In CIGOS 2019, Innovation for Sustainable Infrastructure (pp. 1191–1196), Springer. https://doi.org/10.1007/978-981-15-0802-8_191.

[4] Salman, A. G., Heryadi, Y., Abdurahman, E., & Suparta, W. (2018). Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting. Procedia Computer Science, 135, 89–98. https://doi.org/10.1016/j.procs.2018.08.153

[5] Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in Neural Information Processing Systems (pp. 802–810). https://doi.org/10.48550/arXiv.1506.04214

[6] Aguasca-Colomo, R., Castellanos-Nieves, D., & Méndez, M. (2019). Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife island. Applied Sciences, 9(22), 4931. https://doi.org/10.3390/app9224931

[7] Chao, Z., Pu, F., Yin, Y., Han, B., & Chen, X. (2018). Research on real-time local rainfall prediction based on MEMS sensors. Journal of Sensors, 2018. https://doi.org/10.1155/2018/6184713

[8] Kumar, D., Singh, A., Samui, P., & Jha, R. K. (2019). Forecasting monthly precipitation using sequential modelling. Hydrological Sciences Journal, 64(6), 690–700. https://doi.org/10.1080/02626667.2019.1595624

[9] 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

[10] Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2018). Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. Journal of Hydrology, 561, 918–929. https://doi.org/10.1016/j.jhydrol.2018.04.065

[11] Duan A, Publisher Q. A global optimization strategy for efficient and effective calibration of hydrologic models. Item Type Text; Dissertation-Reproduction (Electronic). The University of Arizona. 1991. http://hdl.handle.net/10150/185655

[12] Luk KC, Ball JE, Sharma A. An application of artificial neural networks for rainfall forecasting. Math Comput Model 2001;33(6–7):683–93. doi: https://doi.org/10.1016/S0895-7177(00)00272-7

[13] Duan Q, Sorooshian S, Gupta VK. Optimal use of the SCE-UA Global optimization method for calibrating watershed models. J Hydrol 1994;158(3–4):265–84.doi: https://doi.org/10.1016/0022-1694(94)90057-4

[14] Box G, Jenkins G, Reinsel G, Ljung G. Fifth edition time series analysis forecasting and control. In: Balding D, Cressie N, Fitzmaurice G, Givens G, Goldstein H, Molenberghs G, Scott D, Smith A, Tsay R, Weisberg S, editors. John Wiley & Sons; 2016. https://doi.org/10.1111/jtsa.12194

[15] Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews. 2010; 29 (5-6):594–621. https://doi.org/10.1080/07474938.2010.481556

[16] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25 (NIPS); 2012. p. 1097–1105. https://doi.org/10.1145/3065386

[17] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[18] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016; 529:484–503. https://doi.org/10.1038/nature16961

[19] Vinay Kumar Reddy Chimmula, Lei Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, Chaos, Solitons & Fractals,Volume135,2020,109864,ISSN0960-0779. https://doi.org/10.1016/j.chaos.2020.109864