Long-term Precipitation Prediction using Machine Learning

B Tech Project Report

Submitted

in partial fulfilment of the requirements of the

Degree of

Bachelor of Technology in Mechanical Engineering

by

Jonnalagadda Naveen Reddy

Roll no: 20ME31021

Under the supervision of

Prof. Rajib Maity

Department of Civil Engineering

Department of Mechanical Engineering

Indian Institute of Technology Kharagpur

November 2023

DEPARTMENT OF MECHANICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
KHARAGPUR-721302, INDIA

CERTIFICATE

This is to certify that the Report entitled **“Long-term Precipitation Prediction using Machine Learning”** is a bona fide record of work carried out by Jonnalagadda Naveen Reddy under my supervision and guidance for the award of degree of Bachelor of Technology in the department of Mechanical Engineering at the Indian Institute of Technology Kharagpur.

Prof. Rajib Maity
Department of Civil Engineering
Indian Institute of Technology
Kharagpur
West Bengal 721302

Date: 28-11-2023
Place: IIT Kharagpur

ACKNOWLEDGEMENT

I want to express my sincere gratitude to Prof. Rajib Maity for guiding me throughout this project and giving me permission to undertake it. His constant support and guidance have been crucial in navigating through the complexities of the research. A special thanks to my mentor, Ms. Subhasmita Dash. Her insights and assistance were invaluable throughout the project, and this endeavour would not have been possible without her expertise and encouragement.

Undertaking the project titled " **Long-term Precipitation Prediction using Machine Learning** " has been a fulfilling journey, and I appreciate the support and guidance of Prof. Rajib Maity and Ms. Subhasmita Dash.

Declaration of Generative AI and AI-assisted technologies in the writing process:

During the preparation of this report, I have used the ‘OpenAI’ platform in order to improve language. After using the service, I received and edited the content as needed. I take full responsibility for the content of the report as it is in its current form.

Jonnalagadda Naveen Reddy

Date: 28-11-2023

Roll no: 20ME31021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

The Indian Summer Monsoon (ISM) is a climatic phenomenon of paramount importance for India, with far-reaching implications for various aspects of society, economy, and ecosystems. Constituting a significant 75% of the total annual rainfall in the country, the ISM plays a central role in shaping the Indian climate and influencing numerous facets of life.

Agricultural Productivity and Economy:

- The Indian economy, especially its agricultural sector, is deeply tied to the Indian monsoon. The timely onset and adequate rainfall during the monsoon season are crucial for crop cultivation and yield.
- Fluctuations in the Indian Summer Monsoon Rainfall (ISMR) can have profound effects on agricultural productivity. A deficient monsoon may lead to drought conditions, crop failures, and economic hardships for farmers.

Ecological Balance and Biodiversity:

- The ISM is a key driver of ecological processes and biodiversity across diverse ecosystems in India. Many plant and animal species have adapted their life cycles to the seasonal rhythm of the monsoon.
- The timing and intensity of the monsoon influence vegetation growth, breeding patterns of wildlife, and overall ecosystem health. Changes in the monsoon patterns can disrupt these ecological processes, impacting biodiversity.

Water Resources:

- The health of rivers, lakes, and other water bodies is intricately linked to the Indian monsoon. Adequate rainfall replenishes water sources and ensures the availability of water for various needs.
- Variability in the monsoon can lead to water scarcity or excess, affecting water quality and availability for both human and natural systems.

Human Livelihoods:

- Beyond agriculture, the monsoon also significantly influences other sectors of the economy. Industries such as water-dependent manufacturing and hydroelectric power generation are directly impacted by monsoon patterns.
- Additionally, the livelihoods of millions of people engaged in sectors like fishing and forestry are closely tied to the health and predictability of the monsoon.

Climate Resilience and Adaptation:

- Given the crucial role of the ISM, its variability poses challenges for climate resilience and adaptation. Unpredictable monsoon patterns can necessitate adaptive measures in agriculture, water management, and disaster preparedness.

In summary, the Indian Summer Monsoon is not merely a climatic event; it is a lifeline for India, influencing its economic activities, ecological health, and the well-being of its people. Monitoring and understanding the dynamics of the ISM are essential for sustainable development, climate adaptation, and the preservation of India's rich natural heritage.

Motivation

The advancement of forecasting capabilities for the Indian Summer Monsoon Rainfall (ISMR) at extended lead times is a critical goal with far-reaching implications for various sectors in India. Such advancements are essential for several reasons, and they hold the potential to significantly enhance the country's ability to manage water resources, respond proactively to extreme weather events, and build resilience against the impacts of the Indian summer monsoon.

Effective Water Management:

- Accurate and extended lead-time forecasts of ISMR provide valuable information for water resource management. Water reservoirs, crucial for irrigation, drinking water supply, and hydropower generation, can be managed more efficiently with advance knowledge of the expected rainfall patterns.

Proactive Responses to Flood and Drought Scenarios:

- Timely and reliable forecasts allow for proactive responses to potential flood or drought scenarios. Authorities can implement measures such as flood defenses, water storage, and drought mitigation strategies well in advance, reducing the impact on communities and infrastructure.

Humanitarian Crisis Prevention:

- Improved forecasting is a key tool in preventing or mitigating humanitarian crises associated with extreme monsoon events. Early warnings enable communities and emergency responders to prepare for, respond to, and recover from disasters, ultimately saving lives and minimizing the socio-economic impact.

Agricultural Planning and Risk Management:

- Farmers heavily depend on the monsoon for crop cultivation. Enhanced forecasting allows for better agricultural planning, enabling farmers to choose suitable crops, adjust planting schedules, and implement water-saving techniques based on anticipated rainfall conditions.

Infrastructure Resilience:

- Infrastructure planning and design, including roads, bridges, and buildings, can be informed by more accurate ISMR forecasts. This contributes to the resilience of infrastructure against the varying intensity and duration of the monsoon.

Climate Adaptation Strategies:

- Given the increasing variability in climate patterns, improved ISMR forecasts are essential for the development and implementation of climate adaptation strategies. This includes sustainable land use planning, watershed management, and ecosystem conservation.

Economic Stability:

- The stability of various economic sectors, from agriculture to energy, is closely tied to the predictability of the monsoon. Accurate forecasts allow businesses and policymakers to make informed decisions, promoting economic stability and growth.

In essence, a more nuanced understanding and accurate forecasting of ISMR are foundational elements for sustainable planning and resilience-building against the unpredictable nature of the Indian summer monsoon. The ability to extend lead times in forecasting provides a valuable window for proactive decision-making, enabling India to adapt to the challenges and opportunities presented by this crucial climatic phenomenon.

Objective

To develop machine learning models for accurately predicting long-term precipitation patterns, extending beyond traditional weather forecasting timelines, holds great significance in the realm of climate science and water resource management. Additionally, the classification of predicted total precipitation values into categories like drought, floods, and normal conditions adds a layer of actionable insights for proactive decision-making in various sectors. Here's an elaborate discussion on both aspects:

1. Long-Term Precipitation Prediction:

Importance:

- Climate Resilience: Long-term precipitation prediction is crucial for building climate resilience. It provides insights into potential water availability, enabling proactive planning for water resource management, agriculture, and ecosystem sustainability.
- Impact on Infrastructure: Many infrastructural projects, such as dams and reservoirs, are designed based on long-term precipitation patterns. Accurate predictions contribute to better infrastructure planning and resilience against extreme weather events.
- Economic Planning: Various economic sectors, including agriculture, energy, and water-dependent industries, benefit from knowing long-term precipitation trends. Businesses and policymakers can make informed decisions, promoting economic stability.

Machine Learning Approach:

- Feature Selection: Machine learning models for long-term precipitation prediction would involve selecting relevant features such as historical precipitation data, temperature, geographical features, and atmospheric pressure.
- Time Series Analysis: Time series models, like recurrent neural networks (RNNs) or Long Short-Term Memory networks (LSTMs), can capture temporal dependencies in precipitation patterns, enabling accurate long-term predictions.
- Ensemble Methods: Combining predictions from multiple models using ensemble methods can improve accuracy and robustness.

2. Classification into Drought, Floods, and Normal Conditions:

Importance:

- Risk Assessment: Classifying predicted precipitation into categories aids in assessing the risk of extreme weather events, such as droughts or floods.
- Agricultural Planning: Farmers can adjust their agricultural practices based on the expected precipitation category, optimizing irrigation and planting schedules.
- Emergency Preparedness: Governments and emergency response teams can use the classification to prepare for potential impacts on communities, infrastructure, and water resources.

Statistical Classification:

- Standard Deviation-Based Approach: Classifying new precipitation values can be done statistically by considering standard deviation from historical data.
    - Below Normal (Standard Deviation < -1): Predicted values falling significantly below the historical mean suggest conditions below normal, indicative of potential drought risks.
    - Normal (1 > Standard Deviation > -1): Values within one standard deviation from the mean are classified as normal, suggesting typical precipitation patterns.
    - Above Normal (Standard Deviation > 1): Precipitation values exceeding one standard deviation above the mean signify conditions above normal, hinting at potential flood risks.

In conclusion, utilizing statistical measures like standard deviation for classification offers a practical and transparent method for predicting and categorizing long-term precipitation patterns. While machine learning models enhance precision, this approach integrates statistical rigor with historical context, ensuring robust predictions for climate resilience planning and risk assessment.

Literature Review

The climate indices, Dipole Mode Index (DMI), El Niño-Southern Oscillation (NINO), North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO), and Southern Oscillation Index (SOI) are indeed crucial determinants of atmospheric conditions, and they play pivotal roles in shaping climatic dynamics that influence precipitation patterns. Let's elaborate on each of these indices:

1. Dipole Mode Index (DMI):

- Definition: The DMI measures the difference in sea surface temperature anomalies between the western and eastern equatorial Indian Ocean. It is particularly associated with the Indian Ocean Dipole (IOD) phenomenon.
- Impact on Precipitation: The IOD affects the monsoon patterns over the Indian subcontinent. Positive IOD events are associated with increased rainfall in some regions and droughts in others.

2. El Niño-Southern Oscillation (NINO):

- Definition: NINO indices, such as NINO3 and NINO4, represent sea surface temperature anomalies in the central and eastern equatorial Pacific. El Niño and La Niña events are phases of the ENSO phenomenon.
- Impact on Precipitation: El Niño tends to bring drier-than-average conditions to some regions and increased rainfall to others, while La Niña often has the opposite effect, influencing global climate patterns.

3. North Atlantic Oscillation (NAO):

- Definition: NAO measures the difference in atmospheric pressure between the Icelandic Low and the Azores High in the North Atlantic.
- Impact on Precipitation: NAO influences the strength and position of the jet stream, affecting storm tracks and precipitation patterns in North America and Europe. Positive NAO phases are associated with wetter conditions in northern Europe and drier conditions in southern Europe.

4. Pacific Decadal Oscillation (PDO):

- Definition: PDO is a long-term climate pattern in the North Pacific that involves variations in sea surface temperatures.

- Impact on Precipitation: PDO influences the position and intensity of the Aleutian Low, affecting storm tracks and precipitation patterns in the Pacific Northwest of North America. It has long-term effects on regional climate.

5. Southern Oscillation Index (SOI):

- Definition: SOI measures the difference in atmospheric pressure between Tahiti and Darwin, Australia, and is associated with the ENSO phenomenon.
- Impact on Precipitation: SOI is an indicator of El Niño and La Niña events. El Niño tends to result in drier conditions in the western Pacific, while La Niña often brings wetter conditions.

In summary, these climate indices are integral components of climate systems, and their interactions shape the climatic dynamics that impact precipitation patterns on regional and global scales. Studying and monitoring these indices contribute significantly to improving our ability to predict long-term precipitation and enhance climate resilience.

Study Area and Data

Study Area:

Geographical Focus: The study is centered on the Hyderabad region, which is situated at approximately 18.25° N latitude and 78.75° E longitude. This geographic area includes the city of Hyderabad and its surrounding regions.

Importance of Study Area Selection:

- Local Relevance: Focusing on a specific geographic region allows for a detailed examination of climatic conditions in a local context, which is particularly relevant for understanding the impacts on the local population, ecosystems, and infrastructure.
- Hyderabad's Significance: Hyderabad, being a major urban center in India, experiences a diverse range of climatic conditions. Studying this area provides insights into the unique challenges and opportunities associated with urban climate dynamics.

Data Collection:

Primary Dataset: The primary dataset utilized for this study is sourced from the ERA-5 (European Centre for Medium-Range Weather Forecasts Reanalysis 5) database. ERA-5 is a state-of-the-art reanalysis dataset that combines a vast array of observational data with a numerical weather prediction model to create a comprehensive and high-resolution representation of past weather conditions.

Temporal Coverage: The data collection spans a substantial period, from January 1, 1951, to October 1, 2023. This extensive timeframe ensures a comprehensive representation of climatic conditions over the study period, allowing for the examination of long-term trends, seasonal variations, and potential shifts in climate patterns.

Methodology

Pearson correlation coefficients:

The Pearson correlation coefficient, often denoted as "r", is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship. The coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. In essence, the Pearson correlation provides a numerical value that helps assess how closely two variables move together in a linear fashion. A positive correlation indicates that as one variable increases, the other tends to increase, and vice versa for a negative correlation. This coefficient is widely used in statistical analysis, providing insights into associations between variables in fields such as economics, biology, psychology, and more.

$$\text{Pearson coefficient factor } (r) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \quad \ldots\ldots\ldots\ldots \; (eq.\ 1)$$

Where x_i, y_i are values of ith point of feature X, Y and $\bar{x}$, $\bar{y}$ are mean values for features X, Y.

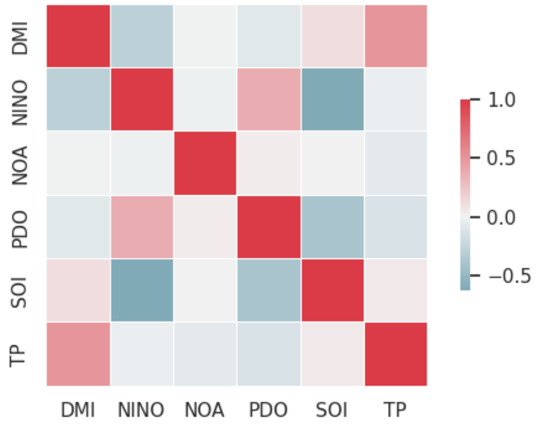
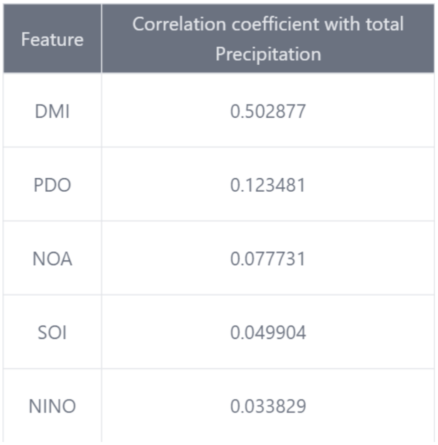


Figure 1: Correlation matrix



Feature	Correlation coefficient with total Precipitation
DMI	0.502877
PDO	0.123481
NOA	0.077731
SOI	0.049904
NINO	0.033829

Table 1: Correlation coefficients

The correlation coefficients above indicate varying degrees of linear relationships between multiple input variables and a specific output. Given the presence of correlated variables, a multivariate approach is warranted. A multivariate LSTM (Long Short-Term Memory) neural network is a suitable choice for modelling these relationships, especially in sequential data.

Multivariate Long Short-Term Memory (LSTM):

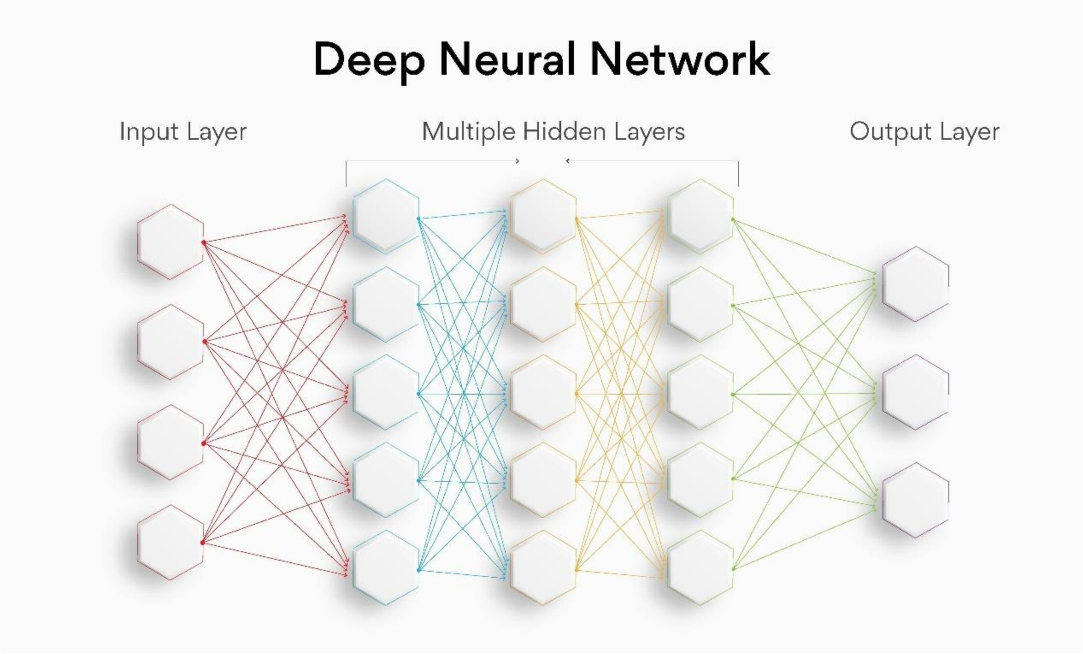


Figure 2: Deep Neural Network Architecture

Multivariate LSTM (Long Short-Term Memory) refers to the use of LSTM neural networks in a multivariate setting, where multiple input variables are considered simultaneously. LSTMs are a type of recurrent neural network (RNN) designed to capture long-term dependencies and sequential patterns in data. In a multivariate LSTM:

Multiple Input Variables:

- Instead of having a single input variable, there are multiple input features or time series that are considered together. Each input variable can be a separate sequence, and the model learns to capture dependencies and relationships among them.

Sequential Dependencies:

- LSTMs excel at capturing sequential dependencies in data. They are effective in scenarios where the order of observations matters, making them suitable for time series analysis or any data with a temporal component.

Handling Correlation:

- If input variables are correlated, the multivariate LSTM can learn to exploit these correlations to better model complex relationships in the data. This makes it particularly useful when dealing with datasets where different features influence each other.

Architecture:

- The architecture of a multivariate LSTM involves multiple input channels feeding into the LSTM layer. The network learns to weight the contributions of each input variable over time, allowing it to capture intricate patterns and dependencies.

Parameters affecting the model performance:

Certainly, understanding how various parameters affect the error and model performance is crucial for effectively training and fine-tuning your model. Let's elaborate on each parameter:

1. Batch Size:

- Batch size refers to the number of training examples utilized in one iteration. It impacts how often the model's weights are updated during training.
- Smaller batch size can lead to more frequent weight updates, potentially helping the model converge faster but might increase computational overhead. Larger batch size reduces the frequency of weight updates, but computations are more efficient. However, convergence may be slower.

2. Window Length:

- Window length is the size of the input sequence or the number of time steps considered as input to the model.
- Smaller window length captures short-term dependencies, useful for rapid changes, but may miss long-term patterns. Larger window length captures long-term dependencies but might overlook short-term fluctuations.

3. Number of Neurons:

- The number of neurons, or units, in the LSTM layer determines the complexity of the model's internal representations.
- Fewer the neurons, simpler the model, may struggle to capture intricate patterns. More the neurons, more complex the model, might overfit the training data if not controlled.

4. Dropout Rate:

- Dropout is a regularization technique where randomly selected neurons are ignored during training, reducing overfitting.
- Lower the dropout rate, less is the regularization, may lead to overfitting. Higher the dropout rate, strong is the regularization, may result in underfitting if too high.

5. Epochs:

- An epoch is one complete pass through the entire training dataset.
- Fewer epochs may underfit the data, as the model doesn't see the entire dataset enough times. More epochs may overfit the data, as the model memorizes the training set.

Parameter Tuning for Multivariate LSTM Model:

In the pursuit of optimizing the performance of the Multivariate LSTM model for long-term precipitation prediction, a systematic parameter tuning process was undertaken. The key parameters explored include Batch Size, Window Length, Number of Neurons, Dropout Rate, and Epochs. A manual tuning approach was adopted to iteratively assess the impact of these parameters on the Root Mean Squared Error (RMSE), a metric indicative of prediction accuracy.

Summary of Parameter Configurations and RMSE:

Batch Size Exploration:

- Batch sizes of 32, 64, and 96 were tested.
- Optimal performance was observed with a batch size of 32, yielding an RMSE of 0.220794291.

Window Length Variation:

- Window lengths of 12, 24, 36, 48, 60, and 72 were explored.
- The best results were achieved with a window length of 72, resulting in an RMSE of 0.217568394.

Number of Neurons Experimentation:

- Different configurations, ranging from 128 to 320 neurons, were tested.
- The model exhibited improved performance with 128 neurons, minimizing the RMSE to 0.217568394.

Dropout Rate Analysis:

- Dropout rates of 0.2 and 0.3 were examined.
- A dropout rate of 0.2 demonstrated optimal performance, yielding an RMSE of 0.220794291.

Epochs Adjustment:

- The impact of epochs (training iterations) on performance was investigated.
- The model achieved its best performance at 75 epochs, resulting in an RMSE of 0.207427328.

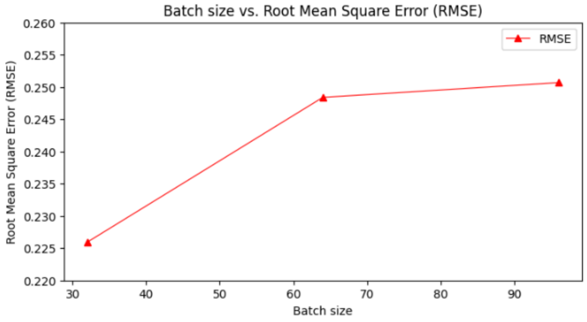


Figure 3: Batch size vs. RMSE

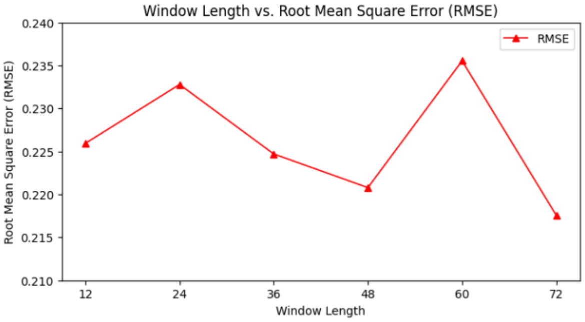


Figure 4: Window length vs. RMSE

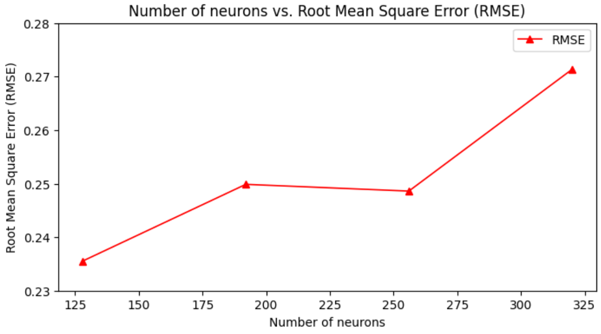


Figure 5: Number of neurons vs. RMSE

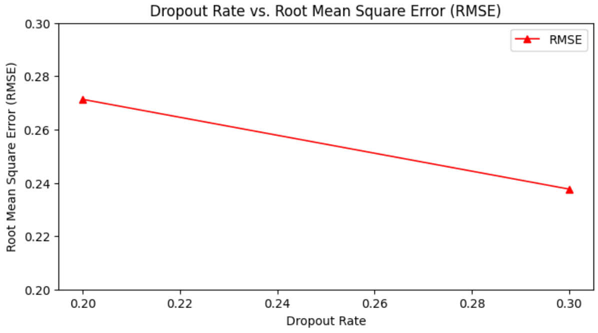


Figure 6: Dropout rate vs. RMSE

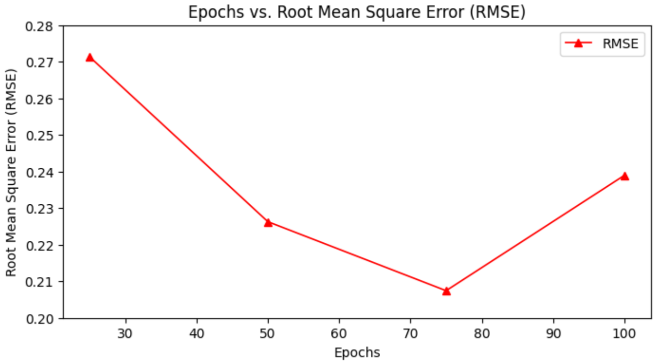


Figure 7: Epochs vs. RMSE

Results and Discussion

Optimal Model Configuration:

To achieve minimal loss and peak performance, the Multivariate LSTM model was fine-tuned with the following parameters:

- Batch Size: 32
- Window Length: 72
- Number of Neurons: 128
- Dropout Rate: 0.3
- Epochs: 75

The resulting Root Mean Square Error (RMSE) value for the model is approximately 0.19 (or ~40 for unscaled values). In the context of precipitation prediction, this low RMSE signifies a highly accurate model, indicating the close alignment of predicted values (P) with actual observed values (A). This optimal configuration ensures that the model strikes the right balance between complexity and generalization, providing robust and precise long-term precipitation predictions.

$$\sqrt{\frac{\Sigma_{i=1}^{N}(A_i - P_i)^2}{N}} \quad \text{............ } \textit{(eq. 2)}$$

Where A_i and P_i are actual value, predicted values of ith data point.

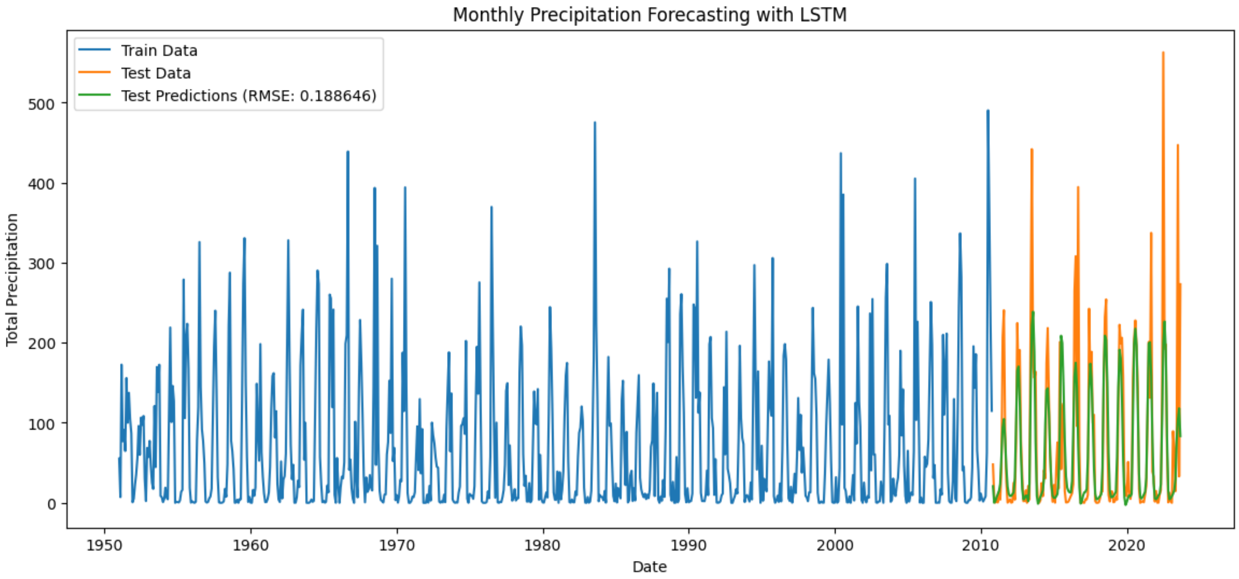


Figure 8: Training data, testing data and predictions for total precipitation

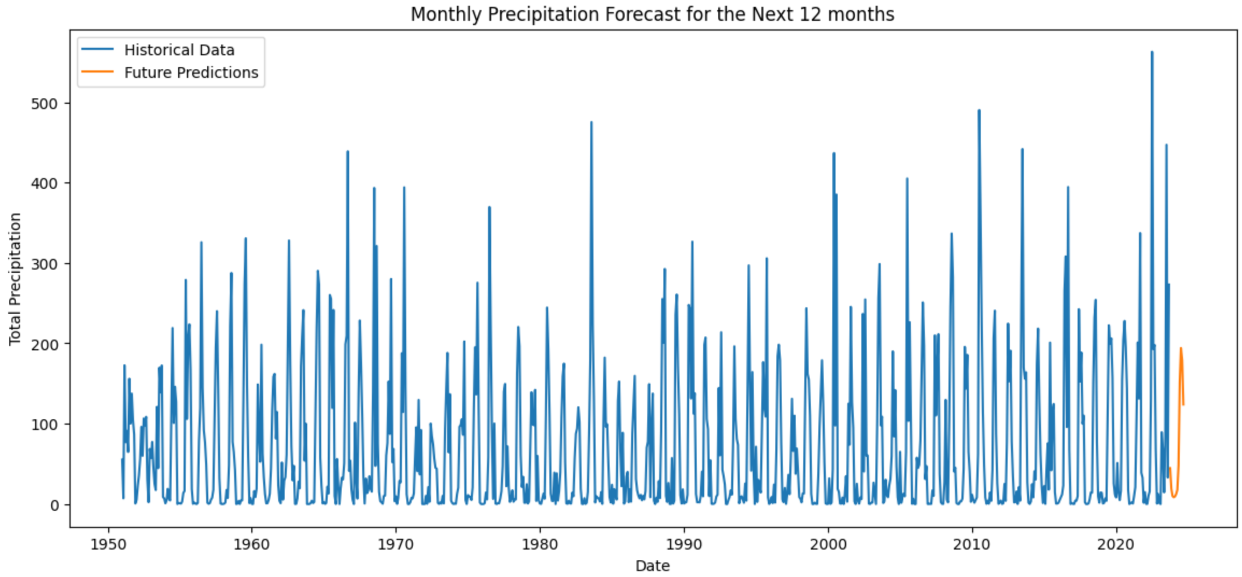
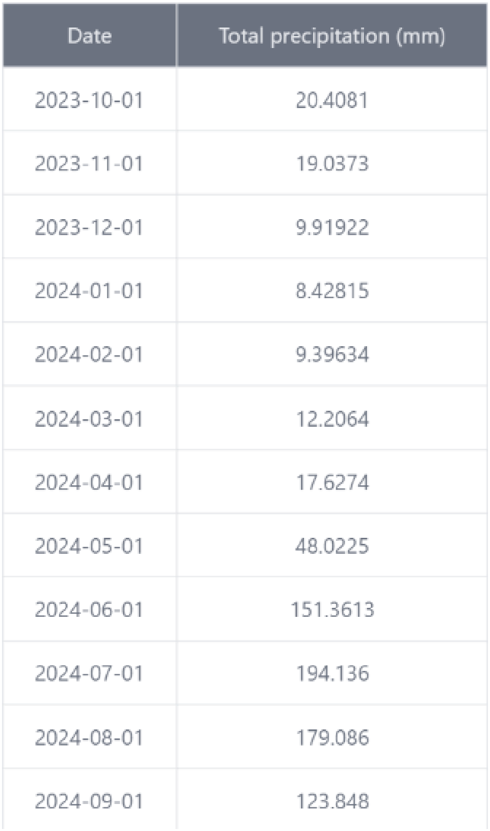


Figure 9: Historical data and Future data for total precipitation



Date	Total precipitation (mm)
2023-10-01	20.4081
2023-11-01	19.0373
2023-12-01	9.91922
2024-01-01	8.42815
2024-02-01	9.39634
2024-03-01	12.2064
2024-04-01	17.6274
2024-05-01	48.0225
2024-06-01	151.3613
2024-07-01	194.136
2024-08-01	179.086
2024-09-01	123.848

Table 2: Future predictions for total precipitation

Conclusions:

This research presents an innovative Multivariate Long Short-Term Memory (LSTM) framework tailored for the forecasting of long-term monthly precipitation patterns. By harnessing time series precipitation data and incorporating pertinent influencing indices, the model demonstrates a remarkable ability to predict "Total Precipitation" for the succeeding 12 months. The achieved Root Mean Square Error (RMSE) value of 0.19 attests to the model's high precision, showcasing its effectiveness in capturing and predicting intricate precipitation patterns.

This success not only underscores the robustness of the proposed framework but also highlights its potential significance in providing valuable insights into long-term climatic trends. The ability to accurately forecast total precipitation over an extended timeframe holds substantial implications for climate research, aiding in proactive decision-making and planning across various sectors, including agriculture, water resource management, and disaster preparedness. The findings of this study contribute to the growing body of knowledge in the field of climatology, showcasing the practical application of advanced machine learning techniques for improved understanding and prediction of complex weather phenomena.

References

1. Spatial variation in long-lead predictability of summer monsoon rainfall using a time-varying model and global climatic indices Riya Dutta, Rajib Maity.
2. A multivariate EMD-LSTM model aided with Time Dependent Intrinsic Cross-Correlation for monthly rainfall prediction Kavya Johny, Maya L. Pai, Adarsh S.
3. Long-Lead Statistical Forecasts of the Indian Summer Monsoon Rainfall Based on Causal Precursors G. Di Capua, M. Kretschmer, J. Runge, A. Alessandri, R. V. Donner, B. Van Den Hurk, R. Vellore, R. Krishnan and D. Coumou.
4. Long-term precipitation prediction in different climate divisions of California using remotely sensed data and machine learning S. Majnooni, M. Reza Nikoo, B. Nematollahi, M. Fooladi, N. Alamdari, G. Al-Rawas and Amir H. Gandomi.