	Summer Internship Report	Code : DO-PFE-01
		Indice de révision : 00
		Edition : 07/2022

Summer Internship Report

Specialization : Computer Networks And Telecommunications
Level: 4th Year

<u>Subject</u>:

Data-Driven Models for Rainfall and River Water Level Forecasting

Prepared by : **Oussama ZIADA**

Host Company:

Karunya Institute Of Technology And Sciences

Host Company Supervisor: **M. Vinodh EWARDS**	***Decision of Internship committee***

Academic Year : 2023/2024

1 Introduction

Water is a vital resource that supports life, ecosystems, and human activities. Managing water resources effectively is crucial, especially in the face of climate change, which has led to increasingly erratic weather patterns. Accurate rainfall and water-level prediction is essential for effective water resource management, agricultural planning, disaster prevention, and policy-making. Traditional methods of forecasting often fall short due to the complex, non-linear, and dynamic nature of environmental systems.

This project aims to harness the power of artificial intelligence (AI) techniques to analyze and predict rainfall patterns and water levels in reservoirs, rivers, and dams. By leveraging state-of-the-art machine learning models, such as Long Short-Term Memory (LSTM) networks and convolutional neural networks (CNNs), we seek to capture complex temporal and spatial dependencies inherent in meteorological data. The project integrates diverse data sources, including historical rainfall records, inflow measurements, and seasonal indicators, to build robust predictive models that can improve the accuracy of forecasts.

Through this work, we hope to provide actionable insights for stakeholders involved in water management, agriculture, and disaster mitigation. By advancing AI-based techniques for environmental prediction, the project contributes to sustainable development efforts, enhancing resilience against the adverse effects of climate variability and change.

2 Host Company presentation

Karunya Institute of Technology and Sciences, formerly Karunya University, is a private deemed university in Coimbatore, Tamil Nadu, India.. The founders aimed to create an institution that combined technical education with strong moral values and a commitment to social service.

Karunya is recognized for its commitment to providing quality higher education with a focus on technical and engineering disciplines.

Karunya University offers a diverse range of undergraduate, postgraduate, and doctoral programs in engineering, science, technology, management, and arts. With a mission to raise leaders in various fields through holistic education, Karunya places a strong emphasis on research and innovation. It has several dedicated research centers and labs that focus on areas such as water, food, healthcare, and energy — aligning with its vision of finding solutions to humanity's pressing problems.

3 Objectives

- **Data Collection and Preprocessing:** Gather and preprocess historical rainfall and river water level data, including handling missing values, normalization, and feature engineering to prepare the data for modeling.

- **Trend and Seasonal Pattern Analysis:** Perform statistical analysis to identify trends, seasonal patterns, and anomalies in rainfall and inflow data, using techniques such as seasonal decomposition, and autocorrelation analysis.

- **Feature Engineering and Selection:** Develop and implement advanced feature engineering techniques, including cyclical encoding for time-based features and the selection of relevant variables to improve model performance.

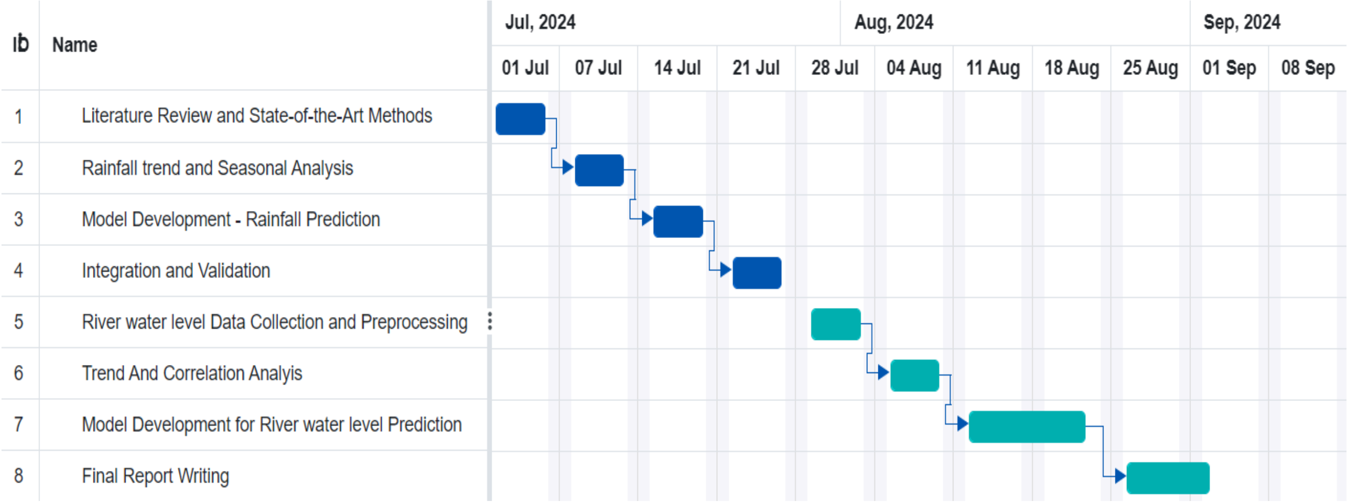
- **Model Development for Prediction:** Build and train machine learning models (e.g., LSTM, ConvLSTM, and models with attention mechanisms) to predict future rainfall and river water level based on historical data.

- **Model Optimization and Hyperparameter Tuning:** Conduct hyperparameter tuning and model optimization using techniques such as grid search or random search to improve model accuracy and robustness.

- **Model Evaluation and Validation:** Evaluate model performance using appropriate metrics (e.g., R-squared, RMSE, MAE) and validate the model with unseen test data to ensure its generalizability.

- **Time Series Forecasting:** Implement and deploy forecasting models to generate short-term and long-term predictions of rainfall and river inflow, leveraging sequential data modeling techniques.

4 Internship Diary



ID	Name
1	Literature Review and State-of-the-Art Methods
2	Rainfall trend and Seasonal Analysis
3	Model Development - Rainfall Prediction
4	Integration and Validation
5	River water level Data Collection and Preprocessing
6	Trend And Correlation Analyis
7	Model Development for River water level Prediction
8	Final Report Writing

5 Conducted work

5.1 Literature Review and State-of-the-Art Methods

To understand the current methods and techniques used in the prediction of rainfall and river water levels, including trend analysis, We focused on recent advancements in time series analysis and forecasting methods, particularly in hydrology and meteorology, covering trend analysis, machine learning (ML) and deep learning (DL) models.

5.1.1 Trend Analysis

Trend analysis is crucial for understanding long-term changes in time series data. Identifying trends helps in adjusting models to account for underlying shifts and ensures that predictive models do not overfit to short-term fluctuations.In the context of rainfall and river water level predictions, trend analysis can reveal long-term climatic changes, such as increasing rainfall or rising river levels due to environmental factors or human interventions. Change point detection helps to identify significant shifts that may correlate with policy changes, infrastructure developments, or extreme weather events.

The trend analysis techniques employed in this study encompass a range of methods to detect and interpret long-term changes in time series data. The **Mann-Kendall test**[1], a non-parametric approach, is used to determine the presence of trends without assuming any specific distribution of the data, enhancing the robustness of the analysis. **Sen's Slope**[1], a robust estimator that is less sensitive to outliers and non-normal data, provides a reliable measure of the trend's magnitude. Additionally, **change point detection**[2] methods, such as the CUSUM method and Bayesian change point analysis, are employed to identify significant shifts in data patterns, indicating possible changes in underlying processes or external factors. Together, these techniques offer a comprehensive framework for understanding both gradual trends and abrupt changes in the context of rainfall and river water level predictions.

5.1.2 Future Forecasting

The literature review reveals a diverse range of approaches for predicting rainfall and river water levels[3], highlighting both the strengths and limitations of traditional and modern methods. Traditional statistical methods, such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA), have been extensively used for time series forecasting. However, these methods often fall short in capturing complex, non-linear patterns inherent in hydrological data, limiting their applicability for more dynamic and unpredictable phenomena.

Machine learning approaches, including Random Forests, Gradient Boosting Machines, and Support Vector Machines, have demonstrated considerable success in predictive tasks, especially when used with well-engineered features. Despite their predictive power, these methods often require extensive manual feature engineering, which can be time-consuming and may not always capture the full complexity of the underlying data relationships.

Deep learning models, particularly Long Short-Term Memory (LSTM) networks[4] and Convolutional LSTM (ConvLSTM) networks, represent the state-of-the-art in modeling both short-term and long-term dependencies in sequential data[4][5]. LSTM models are particularly effective for capturing time-dependent variables due to their ability to retain information over long sequences, while ConvLSTM models add a spatial context to time-series predictions, which is especially useful for predicting river inflow where spatial relationships are important. Attention mechanisms, such as those employed in Transformer-based architectures, further enhance these models by allowing them to focus on the most relevant parts of the input sequence, significantly improving performance on longer sequences or when multiple influencing factors are present.

Additionally, the review underscores the importance of effective data handling techniques, such as dealing with missing data, outliers, normalization, and seasonal decomposition. Proper data preprocessing is crucial for improving model accuracy and ensuring that the models learn meaningful patterns from the data. Overall, these findings provide a comprehensive overview of the current methodologies in use and inform the selection of appropriate models and techniques for the project's subsequent phases.

5.2 Rainfall Trends and Seasonal Analysis

The analysis of rainfall trends and seasonal patterns is a critical component of understanding long-term climatic behavior and its impact on water resources. By examining historical rainfall data, we aim to identify both consistent patterns and significant variations that occur across different seasons. This analysis provides valuable insights into the seasonality of rainfall, highlighting periods of increased or decreased precipitation and their potential impact on river water levels and agricultural activities.

5.2.1 Study Area

The first study was conducted using data from the Malampuzha Dam station, located in the Palakkad district of Kerala, India. This region is a critical area for water management due to its reliance on rainfall and river inflow to support agricultural activities, drinking water supply, and hydropower generation.

For this analysis, we utilized daily rainfall data collected from the Malampuzha Dam station over a period spanning from 2000 to 2020. This 20-year dataset provided a comprehensive view of the historical rainfall patterns, enabling the exploration of trends, seasonal variations, and the development of predictive models to better understand and forecast rainfall and river inflow behaviors in this important watershed.

5.2.2 Yearly Analysis

General Trend: There is no consistent upward or downward trend visible in the data, suggesting that the daily average rainfall has remained relatively stable over the two decades, with fluctuations around a certain mean level.

Periods of Low Rainfall: Several years show lower daily average rainfall values, such as 2002 (110 mm), 2003 (105 mm), and 2008 (109 mm). These years represent periods of below-average rainfall, which may correspond to drought conditions or reduced precipitation due to climatic anomalies.

Periods of Higher Rainfall: Conversely, years like 2004 (139 mm), 2005 (137 mm), 2006 (138 mm), 2010 (138 mm), 2017 (146 mm), and 2019 (141 mm) show higher daily average rainfall.

These years suggest periods of above-average rainfall, which could be associated with favorable weather patterns, such as monsoons, or specific climatic events that brought more precipitation.

Yearly Fluctuations: The fluctuations in average rainfall from year to year highlight the variability in rainfall patterns, which is characteristic of many regions influenced by dynamic weather systems. For example, rainfall increased significantly from 2003 (105 mm) to 2004 (139 mm), and from 2017 (146 mm) to 2018 (130 mm), showing notable interannual variability.

Outliers and Anomalies: The lowest daily average rainfall is observed in 2003 (105 mm), while the highest is recorded in 2017 (146 mm). These could be indicative of extreme weather conditions or unusual climatic events. The sharp drop in rainfall from 2017 to 2018 may also reflect a significant change in weather patterns.

Absence of Data for Certain Years: The absence of data for the year 2014 may impact the continuity of analysis. Understanding why this data is missing could provide insights into possible measurement gaps or data recording issues.

Recent Trends: The data from the last few years (2017-2020) shows some variability, with 2017 (146 mm) having the highest average daily rainfall, followed by a decrease in subsequent years, indicating possible changes in precipitation patterns or climatic influences.

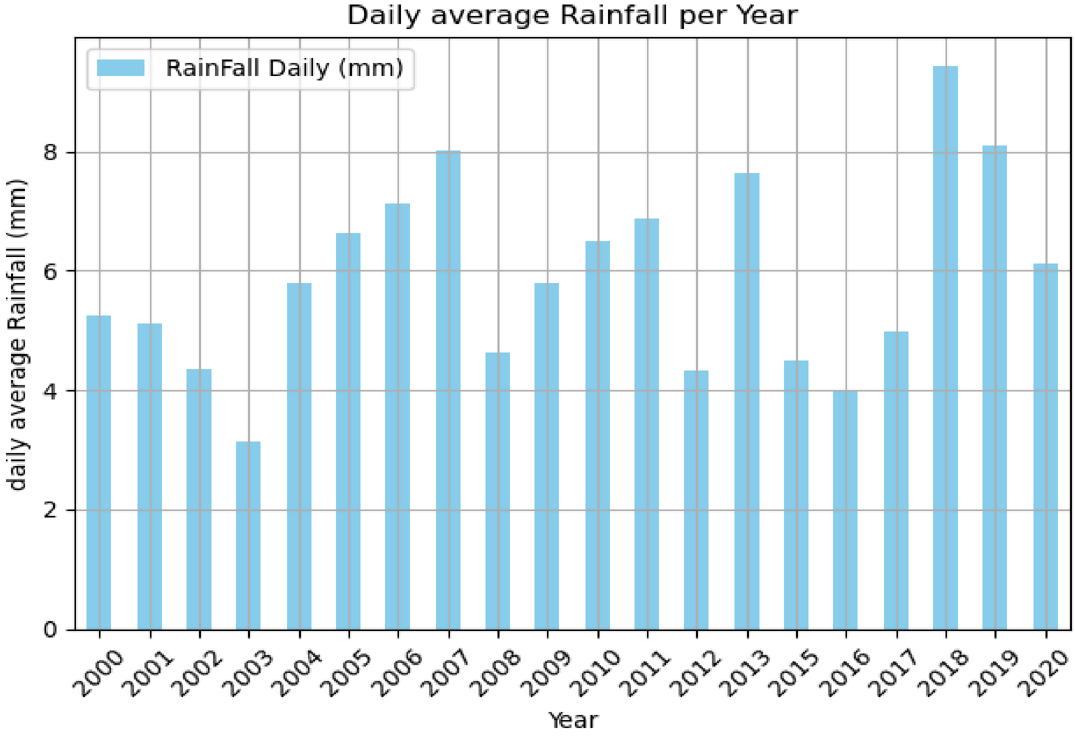


Fig1: Daily Average Rainfall Per Year (in millimeters)

Overall, the data suggests that while there is variability in the annual daily average rainfall, there isn't a clear increasing or decreasing trend over the 20-year period. The data reflects typical interannual variability in rainfall, which could be influenced by regional climatic conditions, weather patterns, or broader climatic phenomena like El Niño or La Niña. To draw more definitive conclusions, further analysis, including trend analysis, and incorporating other meteorological variables, would be beneficial.

5.2.3 Monthly Analysis

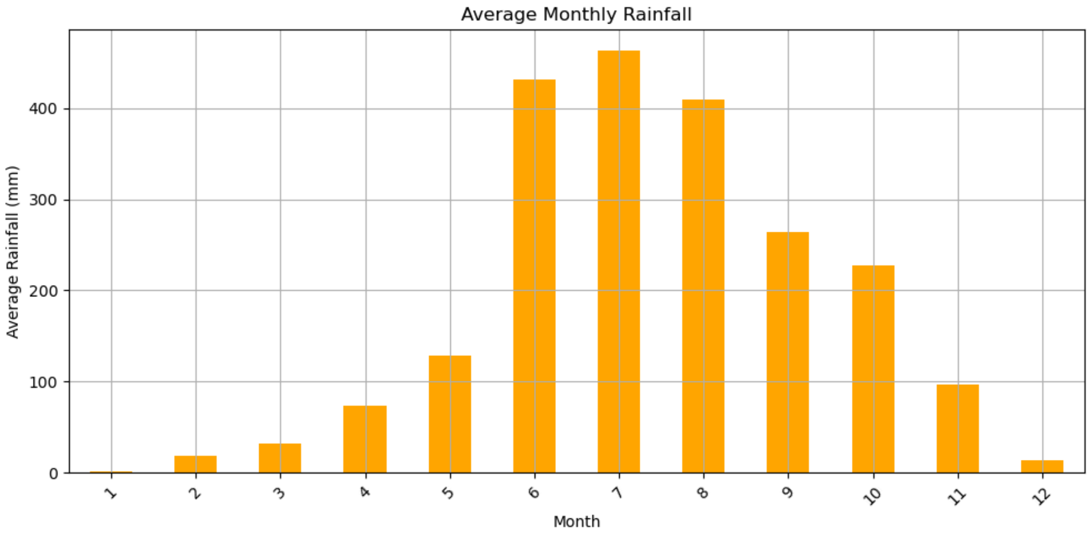


Fig2: Daily Average Rainfall Per Year (in millimeters)

Seasonal Distribution: The data shows a clear seasonal pattern, with lower average rainfall in the early months of the year (January: 1.835 mm, February: 18.160 mm) and progressively increasing amounts in the subsequent months.

Peak rainfall occurs in the middle of the year, particularly in June (431.635 mm) and July (462.695 mm), indicating a monsoon or rainy season characterized by heavy precipitation.

Rainy Season: The highest rainfall averages are observed from June to September (June: 431.635 mm, July: 462.695 mm, August: 408.955 mm, September: 264.495 mm). This period represents the peak rainy or monsoon season, where the region receives the majority of its annual rainfall. The data suggests a pronounced wet season, possibly driven by monsoon winds or other seasonal weather patterns.

Dry Season: Conversely, the lowest rainfall averages are observed from November to February (November: 97.085 mm, December: 14.065 mm, January: 1.835 mm, February: 18.160 mm), indicating a distinct dry season. During these months, rainfall is minimal, and the region may experience dry conditions, which could impact water availability, agriculture, and other water-dependent activities.

Transition Periods: There are noticeable transitions between the dry and wet seasons. March (31.870 mm) and April (73.860 mm) show a gradual increase in rainfall as the region moves toward the wet season. Similarly, October (226.895 mm) represents a transition from the wet to the dry season, with a significant reduction in average rainfall compared to the preceding months.

The monthly average rainfall data reveals a distinct seasonal pattern, with a pronounced wet season from June to September and a dry season from November to February. These results provide critical insights into the region's climate, aiding in planning for water management, agriculture, and disaster mitigation strategies. Understanding these patterns is key to anticipating and managing the impacts of both excess rainfall (flooding) and scarcity (drought) throughout the year.

5.2.4 Seasonal Analysis

Based on the provided data, a comprehensive seasonal analysis has been conducted to understand the rainfall distribution and patterns across the four seasons: Pre-Monsoon (March to May), Southwest (SW) Monsoon (June to September), Northeast (NE) Monsoon (October to November), and Non-Monsoon (December to February)

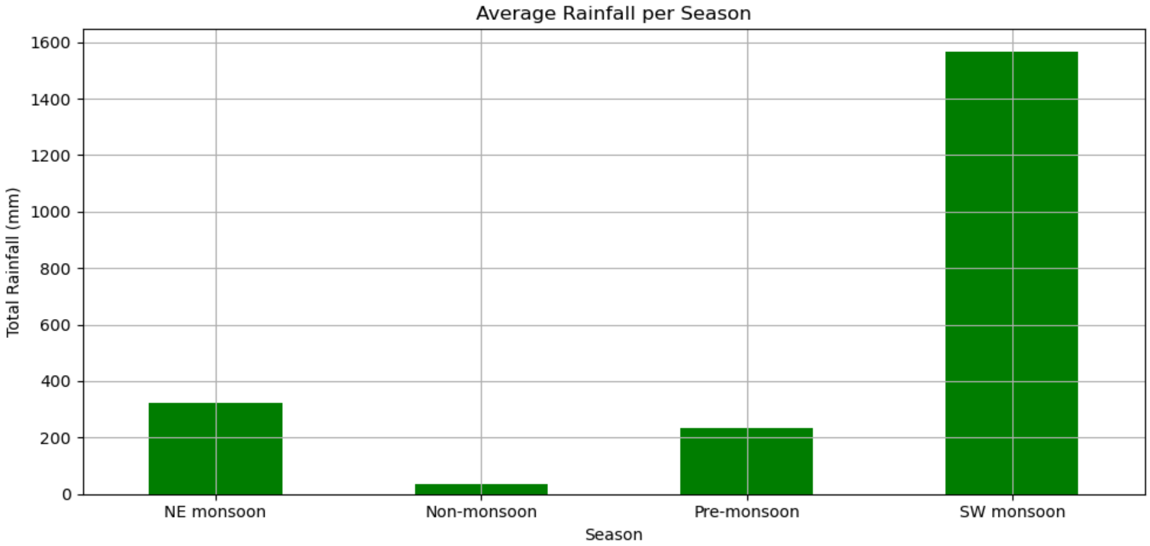


Fig. 3 : Average Rainfall Per Season(in millimeters)

Seasonal Average Rainfall Over All Years:

SW Monsoon (June to September):

The SW Monsoon season has the highest average rainfall as seen in Fig 3, totaling 1,567.78 mm. This season is characterized by intense and prolonged rainfall, making it the primary contributor to the region's total annual precipitation. The high average indicates the significant impact of the SW Monsoon, which is crucial for agriculture, water resource replenishment, and maintaining the region's hydrological balance.

NE Monsoon (October to November):

The NE Monsoon contributes 323.98 mm to the total rainfall, marking it as the second wettest season. The rainfall during this period is significant but much lower compared to the SW Monsoon. This season's contribution is vital, especially for certain regions that rely on the NE Monsoon for water supply.

Pre-Monsoon (March to May):

The Pre-Monsoon season receives an average of 233.705 mm of rainfall. This period is characterized by short, intense rainfall events often associated with thunderstorms. Although it contributes less than the monsoon seasons, the Pre-Monsoon rainfall is important for preparing the land and replenishing moisture before the main monsoon arrives.

Non-Monsoon (December to February):

The Non-Monsoon season has the least rainfall, with an average of only 34.06 mm. This period is generally dry, with minimal precipitation, which may impact water availability and increase the demand for irrigation and water management strategies.

SW Monsoon Rainfall Trends:

The SW Monsoon shows significant year-to-year variability in total rainfall. For instance, the highest total was recorded in 2018 (2759.9 mm), while the lowest was in 2003 (653.0 mm). This variability indicates a high degree of inter-annual fluctuation, which may be influenced by atmospheric circulation patterns, and regional topography.

Periods of exceptionally high rainfall, such as in 2007 (2448.9 mm), 2013 (2138.6 mm), 2018 (2759.9 mm), and 2019 (2279.4 mm), suggest the occurrence of intense monsoon seasons that could lead to flooding and other hydrological extremes. Conversely, years like 2003 (653.0 mm) and 2015 (918.5 mm) show reduced rainfall, potentially indicating drought conditions or weaker monsoon influences.

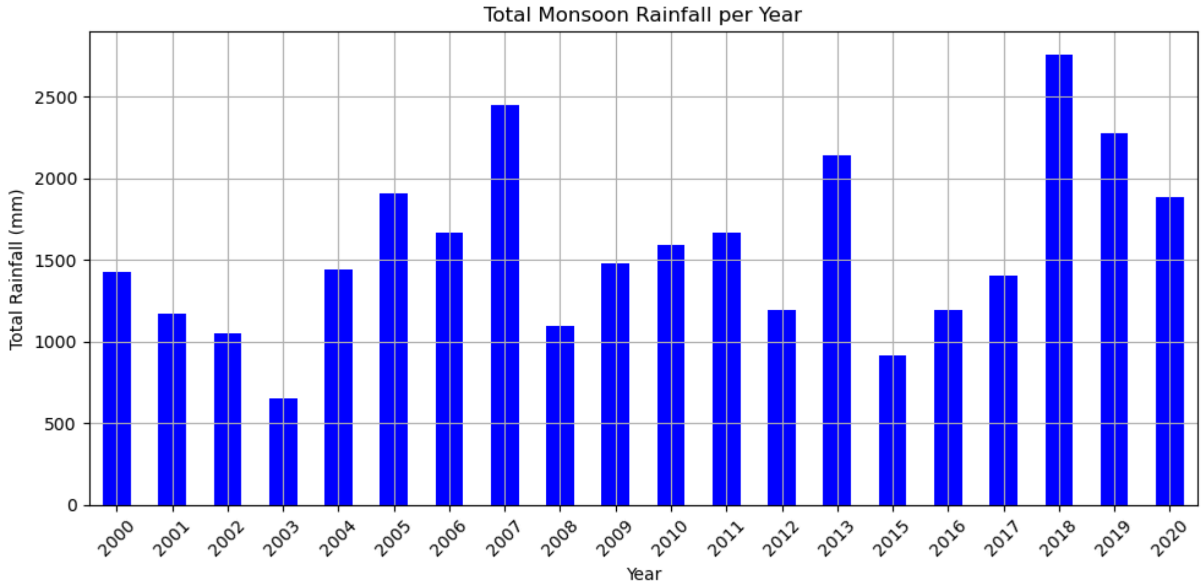


Fig. 4 : Total Monsoon Rainfall Per Year (in millimeters)

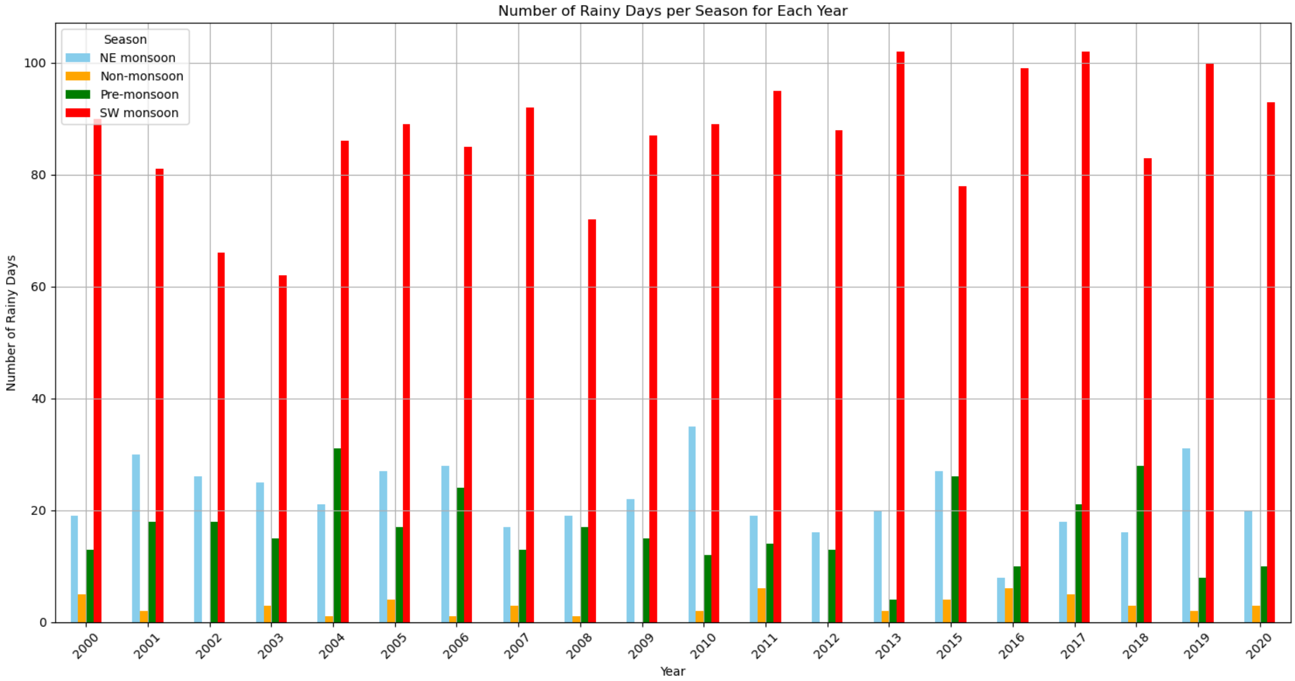


Fig. 5 : Number of rainy days per season per year (in millimeters)

Number of Rainy Days per Season:

SW Monsoon:

The number of rainy days during the SW Monsoon season is the highest, with an average of about 85-100 days of rain annually. This aligns with the high rainfall totals observed, reinforcing the importance of this season for water availability. The maximum number of rainy days is seen in 2013 (102 days) and 2017 (102 days), corresponding to years with substantial rainfall, while the minimum is observed in 2003 (62 days), correlating with a year of low rainfall.

NE Monsoon:

The number of rainy days in the NE Monsoon season is moderate, typically ranging between 16-31 days per year. The variation indicates that the NE Monsoon is less consistent in terms of precipitation frequency, but it still plays a crucial role in contributing to total rainfall.

Pre-Monsoon:

Rainy days during the Pre-Monsoon season range from 4 to 31 days annually, indicating considerable variability. For example, 2004 (31 days) and 2018 (28 days) experienced more frequent rainfall events, which can be crucial for initial soil moisture and agricultural activities.

Non-Monsoon:

The Non-Monsoon season has the least number of rainy days, typically ranging from 0 to 6 days per year. This confirms the season's dry nature, where rainfall is minimal and sporadic, potentially causing water scarcity in some areas.

Overall, the data reveals a distinct seasonal pattern of rainfall, with the SW Monsoon being the most significant period for precipitation, followed by the NE Monsoon, Pre-Monsoon, and Non-Monsoon seasons. The high variability in both total rainfall and the number of rainy days, particularly during the SW Monsoon, suggests that the region's climate is highly dynamic, influenced by various atmospheric and oceanic factors. Understanding these seasonal patterns is crucial for effective water resource management, agricultural planning, and disaster preparedness.

5.2.5 Trend Analysis

Overall Daily Data trend Analysis:

For the daily rainfall data, the Mann-Kendall test indicates a statistically significant increasing trend. The p-value of 0.0138 is below the standard significance level of 0.05, suggesting that this trend is not due to random chance. The positive Z-value of 2.46 further supports the presence of an increasing trend in daily rainfall. Although Kendall's Tau value is 0.0164, indicating a weak positive correlation, there is still evidence of a slight overall increase in daily rainfall amounts over time.

positive	Season	Test Statistic	Z-value	Trend
SW Monsoon	0.053462	0.000070	3.976375	increasing
NE Monsoon	-0.043968	0.008491	-2.631894	decreasing
Non-Monsoon	0.006089	0.188223	1.315855	no trend
Pre-Monsoon	-0.011973	0.251581	-1.146518	no trend

Table 1 : Mann-Kendall Test Results On Seasonal Data

Seasonal Data Trend Analysis:

For the SW Monsoon (June to September), the test results show a statistically significant increasing trend in rainfall. The very low p-value of 0.000070 strongly confirms the presence of this trend, far below the 0.05 threshold, indicating high confidence in this finding. The positive Z-value of 3.9764 further underscores the strength of this trend. This suggests that rainfall during the SW Monsoon season, which is the most critical period for annual precipitation, has been increasing over time. Such a trend could indicate an intensification of monsoon seasons, potentially leading to more frequent and severe weather events, such as flooding.

In contrast, the NE Monsoon (October to November) season displays a statistically significant decreasing trend in rainfall. The p-value of 0.008491 is below 0.05, indicating that this observed decline is statistically significant. A negative Z-value of -2.6319 further confirms the presence of a downward trend. This suggests that the amount of rainfall during the NE Monsoon has been decreasing over the years. Such a trend could have serious implications for regions that rely heavily on this season for their water supply, potentially leading to increased risk of water scarcity and challenges for agricultural practices.

The analysis for the Non-Monsoon (December to February) season shows no significant trend in rainfall. The p-value of 0.188223 is above 0.05, indicating that any observed changes in rainfall during this period are not statistically significant. The Z-value of 1.3159, being close to zero, further supports the lack of a strong trend. This result suggests that rainfall in the Non-Monsoon season has remained relatively stable over time, which is consistent with the typically low precipitation during these months.

Similarly, for the Pre-Monsoon (March to May) season, there is no significant trend detected in the rainfall data. The p-value of 0.251581 is well above 0.05, suggesting that any apparent changes in rainfall are not statistically significant. The Z-value of -1.1465 is relatively low and close to zero, reinforcing the absence of a detectable trend. This indicates that rainfall patterns during the Pre-Monsoon period have remained fairly constant over the years, with no significant increase or decrease observed.

Change-point Analysis:

The analysis using the Pettitt Test, Standard Normal Homogeneity Test (SNHT), and Buishand Range Test reveals no statistically significant change points in the rainfall data. Although all three tests suggest a potential change point at the 4th observation, the p-values (0.6049, 0.34945, and 0.75485, respectively) are substantially higher than the significance level of 0.05. This indicates that any apparent shifts in the data are likely due to random variability rather than a meaningful change in the underlying rainfall patterns. Furthermore, the average rainfall values before and after the identified point (118.25 and 130.625, respectively) are nearly identical, reinforcing the conclusion that there is no significant alteration in the data's mean. Overall, the rainfall data exhibits a stable pattern, with no notable shifts or disruptions detected during the period of analysis.

5.3 Rainfall Forecasting

5.3.1 Proposed architectures

LSTM and Prophet models were used to train on monthly average data, enabling them to effectively capture and predict the temporal patterns and trends in rainfall and river inflow

5.3.1.1 LSTM Model

An LSTM[6] (Long Short-Term Memory) neural network was developed for predicting daily rainfall. The model consists of a single LSTM layer with 50 units, utilizing the tanh activation function and hard_sigmoid as the recurrent activation function. The input shape is set to a window size of 6, representing six days of historical data to predict the subsequent day's rainfall.

To prevent overfitting, a Dropout layer with a rate of 0.2 is applied. The network ends with a Dense layer containing a single neuron with a relu activation function to output the prediction.

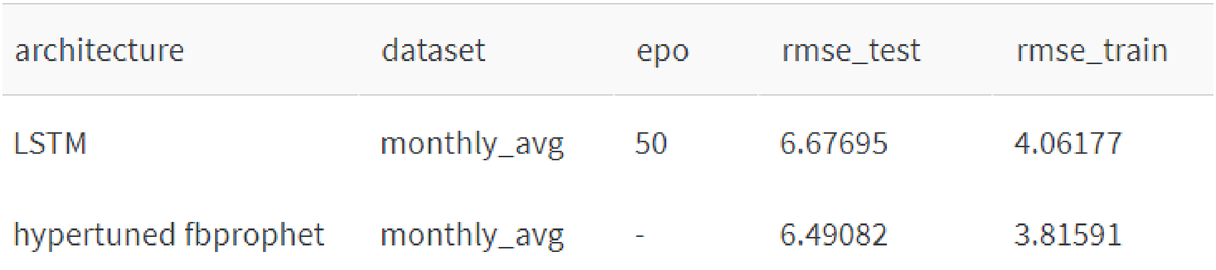
The model was compiled using the Adam optimizer and the Mean Squared Error (MSE) loss function. It was trained for 50 epochs with a batch size of 16. During training, the EarlyStopping callback monitored the loss with a patience of 5 epochs to prevent overfitting, while the ModelCheckpoint callback saved the best model.

5.3.1.2 Prophet Model

A Prophet model[7][8] was employed independently to predict river inflow. Prophet is a time-series forecasting tool that handles various components such as trend, seasonality, and changepoints. The model was configured with specific hyperparameters: a changepoint_prior_scale of 0.08 to control trend flexibility, a seasonality_prior_scale of 5 for regularizing seasonal components, and a multiplicative seasonality mode to capture multiplicative seasonal effects.

These parameters were chosen to optimize the model's ability to capture the complex seasonal and trend dynamics inherent in river inflow data. The model was trained on historical inflow data.

5.3.2 Results And Discussion



architecture	dataset	epo	rmse_test	rmse_train
LSTM	monthly_avg	50	6.67695	4.06177
hypertuned fbprophet	monthly_avg	-	6.49082	3.81591

Fig 6 : Trained Models Results

LSTM Model Performance :

The LSTM model achieved a Root Mean Squared Error (RMSE) of 4.07 on the training dataset and 6.64 on the test dataset. These results demonstrate the model's ability to learn complex temporal dependencies within the rainfall data, capturing underlying patterns effectively during training. The relatively low RMSE on the training data suggests that the LSTM model is capable of fitting the training data well, showcasing its strength in modeling sequential data with its recurrent architecture.

Although the test RMSE of 6.64 is higher than that of the training data, it still reflects a reasonable prediction accuracy for unseen data. The increase in RMSE on the test dataset indicates that while there is room for improvement in terms of generalization, the LSTM model has performed competently given the variability and non-linear characteristics of the rainfall data. With its flexibility and capacity to model complex relationships, the LSTM model remains a promising approach, especially if further tuning or architectural adjustments are applied.

Prophet Model Performance :

The Prophet model demonstrated an RMSE of 3.82 on the training dataset and 6.49 on the test dataset. The model showed a slightly better generalization compared to the LSTM, with a lower test RMSE, suggesting that it handled the test data more effectively. This performance can be attributed to Prophet's ability to automatically detect and model changepoints, trends, and seasonal patterns, which are inherent in the river inflow and rainfall data.

Comparison and Discussion :

When comparing both models, the Prophet model slightly outperformed the LSTM model on both training and test datasets. The Prophet model's test RMSE of 6.49 was lower than the LSTM model's test RMSE of 6.64, indicating a marginally better predictive capability on unseen data. This could be due to the Prophet model's inherent strengths in handling seasonal components and changepoints, which are critical aspects of rainfall data.

However, the difference in performance between the two models is not substantial, and the choice of model may depend on specific use cases. While the LSTM model offers the advantage of learning complex temporal dependencies through deep learning techniques, the Prophet model provides an interpretable framework that can efficiently capture trends and seasonal variations.

Conclusion :

Overall, the results suggest that while both models are effective for predicting rainfall, the Prophet model shows slightly better generalization on test data. This makes it a more reliable choice for practical applications where interpretability and generalization to unseen data are crucial. Future work could focus on refining the LSTM architecture, experimenting with more layers, different activation functions, or additional regularization techniques to potentially improve its performance and narrow the gap with the Prophet model.

5.4 River Water Level Study

5.4.1 Study Area

The study area encompasses the Ohio River, a significant waterway in the United States. The Ohio River originates at Pittsburgh, Pennsylvania, from the confluence of the Monongahela and Allegheny Rivers. Spanning approximately 1,579 kilometers, it flows in a south-westerly direction and eventually merges with the Mississippi River. The Ohio River traverses six states: West Virginia, Pennsylvania, Ohio, Kentucky, Indiana, and Illinois. It drains a vast basin of about 528,358 square kilometers, making it one of the largest river basins in the USA and home to nearly 25 million people[9].

The river's extensive network includes numerous tributaries such as the Scioto, Wabash, Miami, and Muskingum rivers from the north, and the Green, Kentucky, Licking, Sandy, Big, Kanawha, Cumberland, and Tennessee rivers from the south[10]. For this analysis, river water level data were obtained from the gage station at Cincinnati. Additionally, climatic parameters were collected from six locations: Beaver, Wellsburg, Wheeling, Parkersburg, Portsmouth, and Cincinnati. The dataset spans daily records from 1993 to 2024, providing a comprehensive temporal overview essential for accurate analysis and prediction of river water levels.

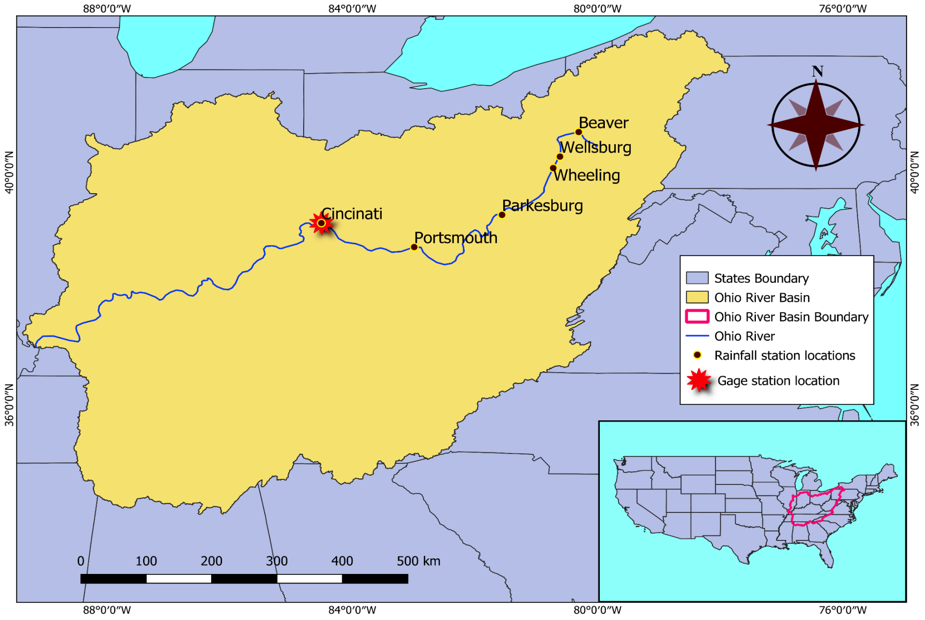


Fig 7 : Ohio River Map

5.4.2 Correlation Analysis

The correlation analysis was conducted to explore the relationships between river water level, temperature(T2M) , and relative humidity(RH2M) both on the same day and with time lags.

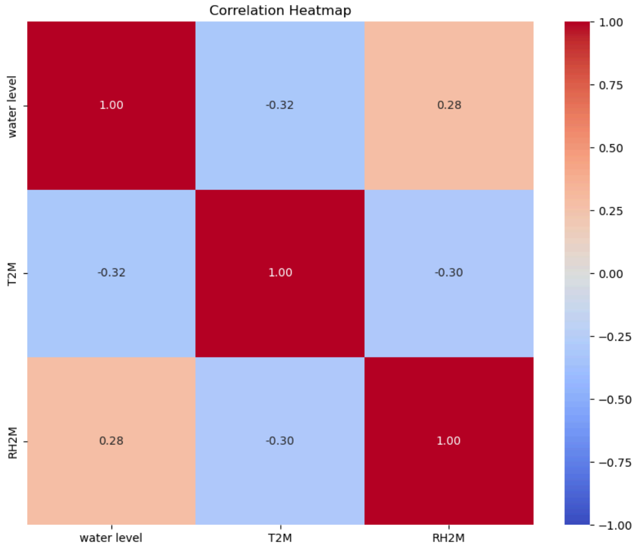


Fig 8 : Same-day correlation matrix

Same-Day Correlation:

The correlation analysis conducted on the same day highlighted important relationships between river water levels, temperature (T2M), and relative humidity (RH2M). The river water level exhibited a moderate negative correlation of -0.324 with temperature, suggesting that higher temperatures are associated with lower river water levels. Conversely, the river water level had a moderate positive correlation of 0.281 with relative humidity, indicating that increased humidity correlates with higher river water levels. Additionally, temperature and relative humidity showed a moderate negative correlation of -0.296, reflecting that higher temperatures are generally associated with lower humidity levels.

Lagged Correlation:

The lagged correlation analysis, which explores the impact of past climate conditions on river water levels, revealed significant insights. The lag that provided the highest correlation was at lag 4. At this lag, the correlation between river water levels and the previous day's temperature (T2M_lag4) was -0.288, while the correlation with the previous day's relative humidity (RH2M_lag4) was 0.411. This indicates that, with a lag of 4 days, temperature shows a notable negative correlation, meaning that higher temperatures in the past are linked to lower current river water levels. In contrast, higher past relative humidity levels are associated with higher current river water levels.

The results from the lagged correlation analysis underscore the delayed effects of climate conditions on river water levels, with temperature having a more consistent negative impact and relative humidity showing a strong positive influence when considering a lag of 4 days. These findings are crucial for refining predictive models and understanding the temporal dynamics between climate variables and river water levels.

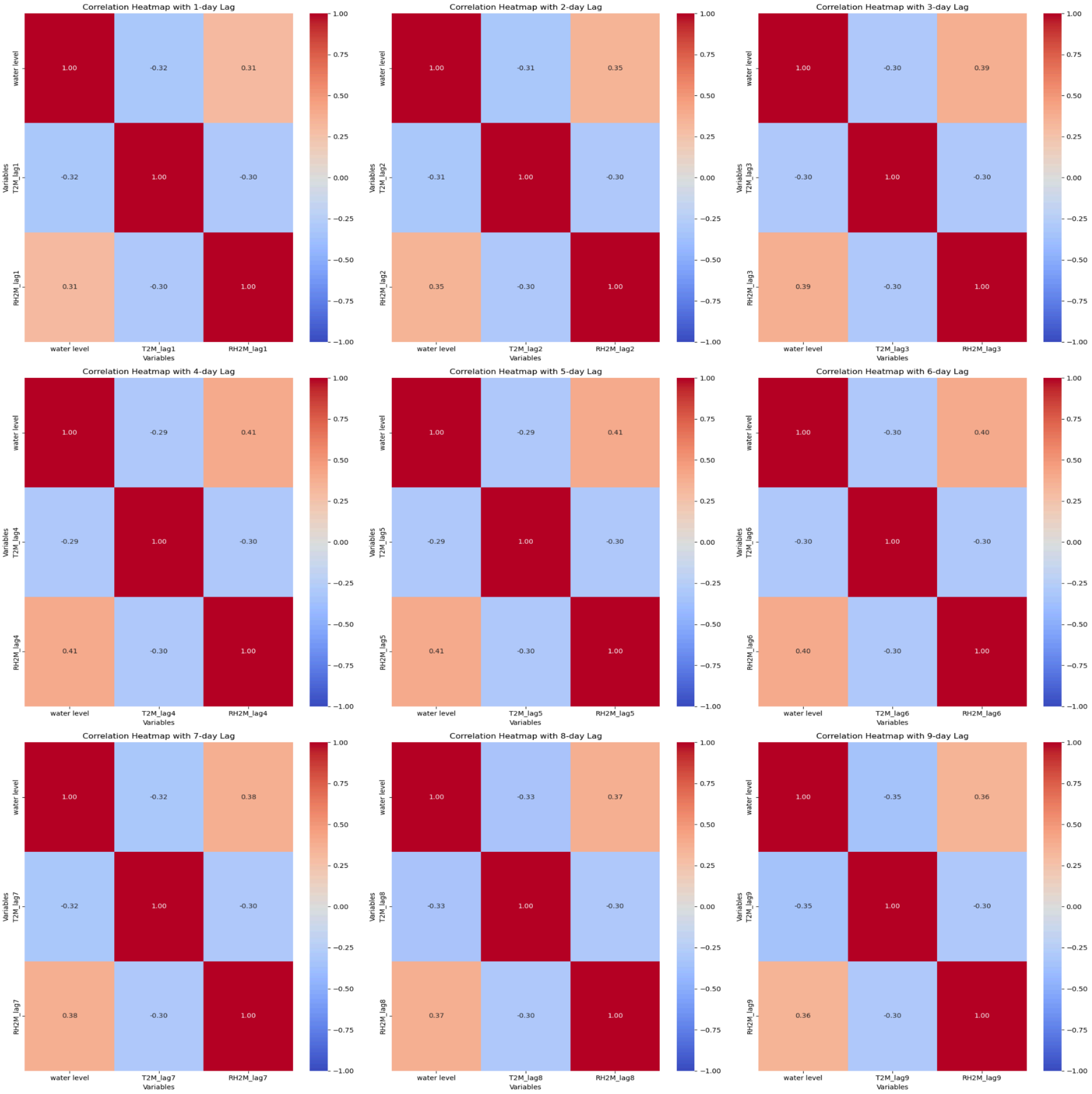


Fig 9 : Lagged-correlation matrix

5.4.3 River Water-level forecasting

5.4.3.1 Proposed Architectures

For forecasting river water levels, two distinct model architectures were employed to evaluate their effectiveness and performance.

LSTM Model

The first architecture is a Long Short-Term Memory (LSTM) network, which is well-suited for time series forecasting due to its ability to capture temporal dependencies. The LSTM model was configured with the following specifications:

- Layers: The model includes an LSTM layer with 50 units, employing the 'tanh' activation function and 'hard_sigmoid' for recurrent activation. This is followed by a Dropout layer with a rate of 0.2 to mitigate overfitting. The output layer consists of a Dense layer with a 'relu' activation function to predict the river water level.
- Compilation: The model uses Mean Squared Error (MSE) as the loss function and the Adam optimizer for training.
- Training: The model is trained over 50 epochs with a batch size of 16. Early stopping and model checkpointing are utilized to avoid overfitting and to save the best-performing model.

The LSTM model processes all features—temperature, humidity, and water level data from previous days—with a window size of 6 days, enabling it to learn from past patterns and predict future water levels effectively.

XGBoost Model

The second architecture is an XGBoost regressor, known for its high performance in regression tasks. The XGBoost model was implemented with the following parameters:

Flattening: The input data is reshaped to a flat format suitable for XGBoost, which does not handle sequential data inherently.

Parameters: The model uses 100 estimators with a learning rate of 0.1, a maximum depth of 5, and subsample and colsample_bytree rates of 0.8 to control overfitting and improve generalization.

Training: The model is trained with a train-validation split, and early stopping is applied to halt training if no improvement is observed in the validation set for 10 rounds.

Similar to the LSTM model, the XGBoost model utilizes all features (temperature, humidity, and water level from previous days) with a window size of 6 days. This approach allows it to leverage historical data for accurate river water level forecasting.

Both models were developed and trained independently to compare their effectiveness in predicting river water levels, providing insights into their relative performance and suitability for the forecasting task.

5.4.3.2 Results And Discussion

Model	**R^2 train**	**R^2 test**	**RMSE train**	**RMSE test**
LSTM	0.98	**0.97**	0.87	**0.97**
XGBoost	**0.99**	**0.97**	**0.69**	1.08

LSTM Model Performance

The Long Short-Term Memory (LSTM) model demonstrated strong performance in forecasting river water levels. The model achieved a high R^2 score of 0.98 on the training data, indicating that it explains 98% of the variance in the training set. The Root Mean Squared Error (RMSE) on the training data was 0.87, reflecting the model's ability to make accurate predictions with minimal error. For the test data, the LSTM model achieved an R^2 score of 0.97, suggesting excellent generalization to unseen data. The RMSE for the test data was slightly higher at 0.97, which is still quite low, indicating robust performance.

The LSTM model's ability to maintain a high R^2 score and a low RMSE across both training and test datasets underscores its effectiveness in capturing the temporal dependencies of the river water level data. This performance is particularly noteworthy given the model's sensitivity to sequential patterns and its successful use of past data to predict future values.

XGBoost Model Performance

The XGBoost model also performed well, though with some differences compared to the LSTM model. The training data results showed an impressive R^2 score of 0.99, meaning the model explained 99% of the variance in the training set, and a notably low RMSE of 0.69. This indicates that the XGBoost model was highly effective in fitting the training data. However, for the test data, the model achieved an R^2 score of 0.97 and an RMSE of 1.08. While these results are still strong, the test RMSE is higher compared to the LSTM model, suggesting that the XGBoost model may have slightly less predictive accuracy on new, unseen data.

The XGBoost model's superior R^2 score on training data reflects its powerful fitting capabilities. However, the increase in RMSE on test data compared to the LSTM model suggests that while

XGBoost excels at learning from historical data, it may be slightly more prone to overfitting or less effective in generalizing to new data compared to LSTM.

Comparison and Discussion

Both models exhibited strong predictive performance, with high R^2 scores and low RMSE values across the training and test datasets. The LSTM model, with its focus on sequential data and temporal patterns, showed a slightly better performance in terms of test data RMSE, indicating its effectiveness in generalizing to new data. On the other hand, the XGBoost model demonstrated exceptional training performance but faced a slightly higher test RMSE, which suggests a potential trade-off between fitting the training data and generalizing to unseen data.

In summary, while both models are effective for forecasting river water levels, the choice between LSTM and XGBoost may depend on specific project requirements and data characteristics. The LSTM model's strength lies in its sequential learning capability, making it a robust choice for time series forecasting. The XGBoost model, with its powerful regression capabilities, performs exceptionally well on training data and remains a competitive option for predictive tasks.

6 Acquired skills

Skill Area	Description
Time Series Analysis and Forecasting	Enhanced understanding of time series data, including trend and seasonal pattern identification. Acquired Proficiency in techniques such as Mann-Kendall test, Pettitt test, and lagged correlation analysis.
Advanced Modeling Techniques	Gained experience with machine learning models and designing deep learning architectures, tuning hyperparameters for predictive tasks.
Data Preprocessing and Feature Engineering	Improved skills in data cleaning, handling missing values, scaling features, and creating new variables for better model performance.
Project Management and Reporting	Enhanced skills in organizing and documenting complex analyses and results, presenting insights clearly and effectively in a structured code and report format.

7 Conclusion

The project provided a comprehensive exploration into the prediction and analysis of rainfall and river water levels, leveraging advanced time series forecasting techniques. Through a detailed literature review and analysis of seasonal and trend patterns, we gained valuable insights into the behavior of rainfall and river water levels, identifying key seasonal variations and trends.

For rainfall prediction, we compared LSTM and Prophet models, each demonstrating strong performance and providing valuable insights into their respective strengths. Similarly, in river water level prediction, both LSTM and XGBoost models were applied, showcasing the effectiveness of different forecasting approaches in handling time series data.

Overall, the project highlighted the successful application of sophisticated machine learning techniques to real-world time series data. The skills acquired in time series analysis, advanced modeling, statistical trend detection, and effective project reporting will be invaluable for future research and practical applications in environmental and hydrological forecasting.

Bibliography

[1] Praveen, B., Talukdar, S., Shahfahad et al. Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Sci Rep 10, 10342 (2020). https://doi.org/10.1038/s41598-020-67228-7

[2] Monir, M.M., Rokonuzzaman, M., Sarker, S.C. et al. Spatiotemporal analysis and predicting rainfall trends in a tropical monsoon-dominated country using MAKESENS and machine learning techniques. Sci Rep 13, 13933 (2023). https://doi.org/10.1038/s41598-023-41132-2

[3] Sarmad Dashti Latif, Nur Alyaa Binti Hazrin, Chai Hoon Koo, Jing Lin Ng, Barkha Chaplot, Yuk Feng Huang, Ahmed El-Shafie, Ali Najah Ahmed,
Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches, Alexandria Engineering Journal,
https://doi.org/10.1016/j.aej.2023.09.060

[4] Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, Michael Weyrich,
A survey on long short-term memory networks for time series prediction,
https://doi.org/10.1016/j.procir.2021.03.088.

[5] Vizi, Zsolt & Batki, Bálint & Rátki, Luca & Szalánczi, Szabolcs & Fehervary, Istvan & Kozák, Péter & Kiss, Tímea. (2023). Water level prediction using long short-term memory neural network model for a lowland river: a case study on the Tisza River, Central Europe. Environmental Sciences Europe. 35. 10.1186/s12302-023-00796-3.

[6] Poornima, S.; Pushpalatha, M. Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units. Atmosphere 2019, 10, 668. https://doi.org/10.3390/atmos10110668

[7]Taylor, Sean & Letham, Benjamin. (2017). Forecasting at scale. 10.7287/peerj.preprints.3190v2.

[8]Hossain, M.M.; Anwar, A.H.M.F.; Garg, N.; Prakash, M.; Bari, M. Monthly Rainfall Prediction at Catchment Level with the Facebook Prophet Model Using Observed and CMIP5 Decadal Data. Hydrology 2022, 9, 111. https://doi.org/10.3390/hydrology9060111

[9]https://ohioriverfdn.org/ohio-river/quick-facts/

[10]Britannica, The Editors of Encyclopaedia. "Ohio River". Encyclopedia Britannica, 24 Aug. 2024, https://www.britannica.com/place/Ohio-River. Accessed 9 September 2024.