A Comparative Analysis of Data-Driven Machine Learning Models for Rainfall Forecasting in Bangladesh

Mir Mahmid Sarker[1], Arish Morshed Zobeyer[2], Tasnuva Rouf[3], and S M Mahbubur Rahman[4]

[1]Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh
[2]Institute of Water Modeling (IWM), Dhaka, Bangladesh
[3]National Oceanic and Atmospheric Administration (NOAA), Maryland, United States
[4]Institute of Water Modelling (IWM), Dhaka, Bangladesh

28 April 2025

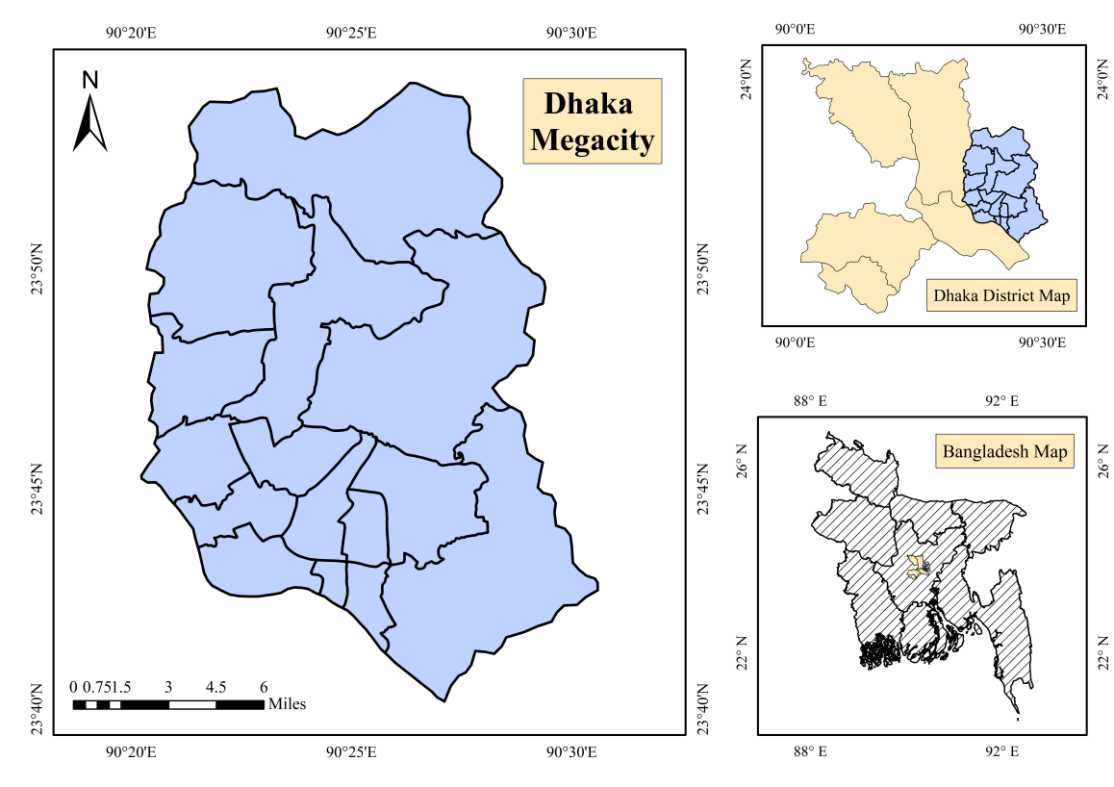
Outline of Presentation

- Background of the Study
- Study Area
- Objectives
- Methodology
- Data Collection
- Model Description
- Result & Discussion
- Conclusion
- Limitations & Recommendations

Background of the Study

- Dhaka, with a population of approximately 23.9 million
- Experiences an average annual rainfall of 2,048 millimeters
- Waterlogging in Dhaka, caused by unplanned urbanization and drainage congestion
- Poor accuracy of existing models

Study Area


90°20'E
90°25'E
90°30'E
N
Dhaka Megacity
23°50'N
23°45'N
23°40'N
0 0.751.5 3 4.5 6
Miles
90°0'E
90°30'E
24°0'N
Dhaka District Map
88° E
92° E
26° N
22° N
Bangladesh Map

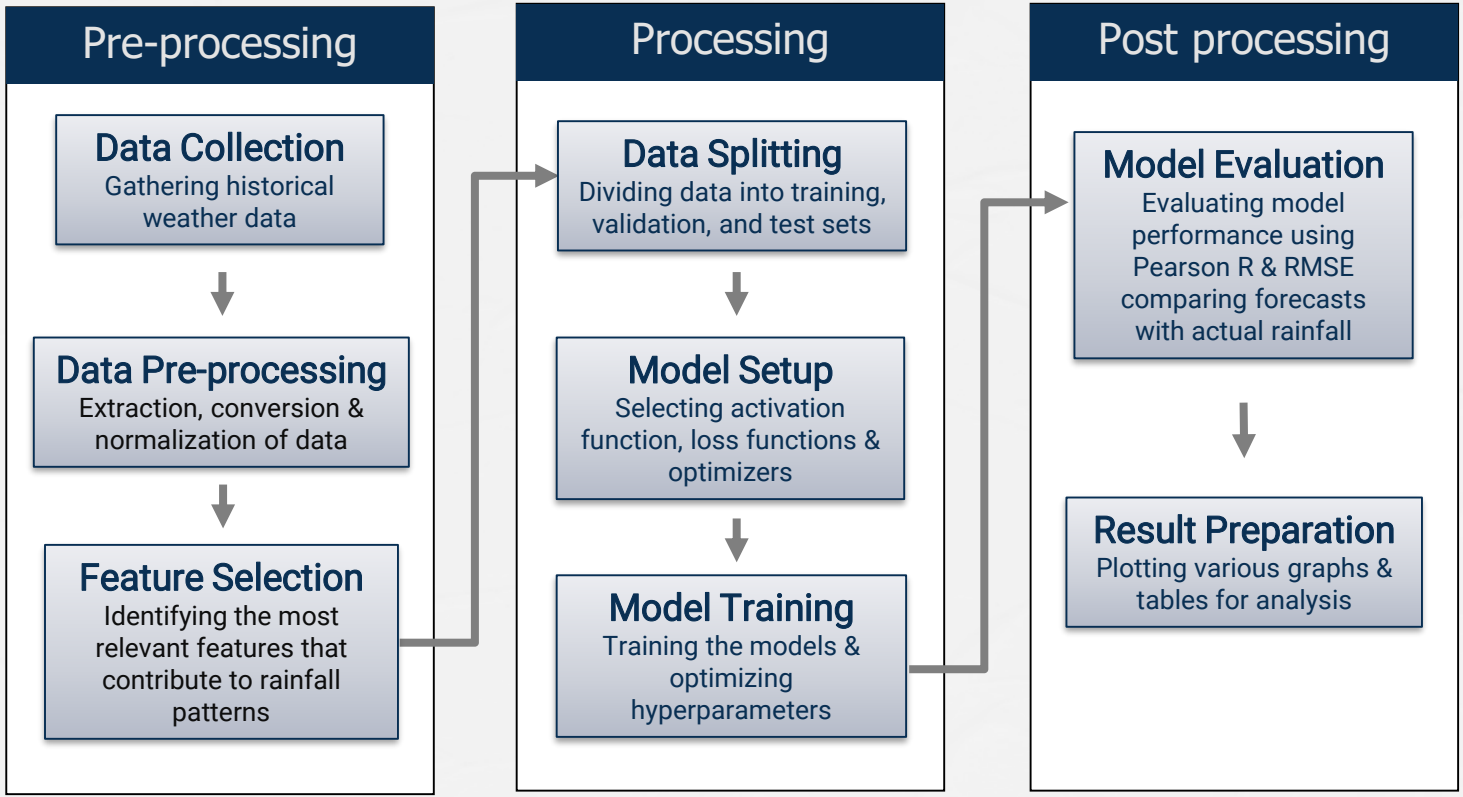
Objectives

To improve 1-5 day rainfall forecasts using Machine Learning algorithms.

&

To evaluate the performance of data driven models against traditional forecasting systems.

Methodology



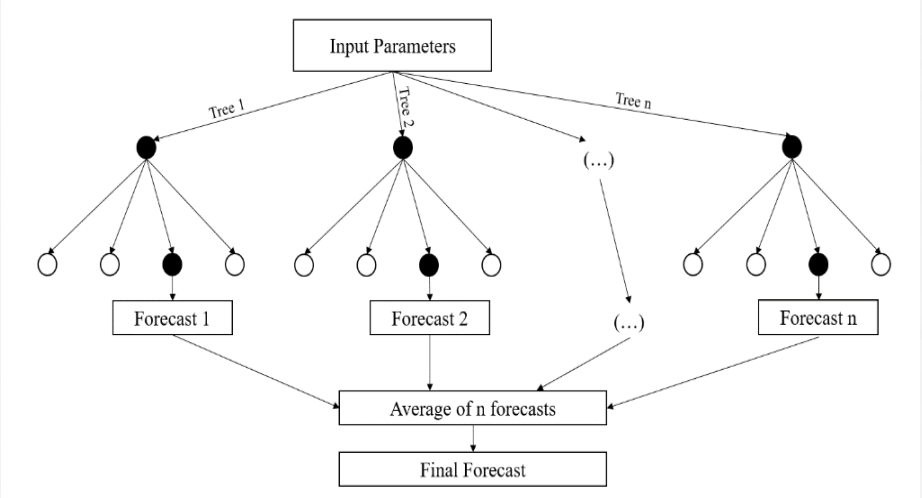
Data Collection

Data	Unit	Dataset/ Station Name	Temporal Resolution	Source	Period
Predictor Variables					
Evaporation (e)	mm	ERA5 hourly data on single levels from 1979 to present	Daily	ECMWF Reanalysis v5 (ERA5)	1979-2023
Large Scale Precipitation (lsp)	m				
Solar Surface Radiation (ssr)	J/m2				
Surface Pressure (sp)	Pa				
Temperature at 2 m (t2m)	K				
Total Cloud Cover (tcc)	-				
Total Column Rain Water (tcrw)	kg/m2				
Total Column Water Vapor (tcwv)	kg/m2				
Target Variable					
Total Precipitation (tp)	mm	ERA5 hourly data on single levels from 1979 to present	Daily	ECMWF Reanalysis v5 (ERA5)	1979-2023
Traditional Forecast Model Data					
Precipitation data	mm	NCEP GDAS/FNL 0.25 Degree Global	Daily	Global Forecast System	2018-2023

Model Description

Random Forest Regression (RF)

- Combines hundreds of small decision trees
- Each tree uses random data & random features to avoid overfitting.
- Handles messy, real-world data (like Dhaka's rainfall) better than single models.



Architecture of a typical Random Forest Regression model

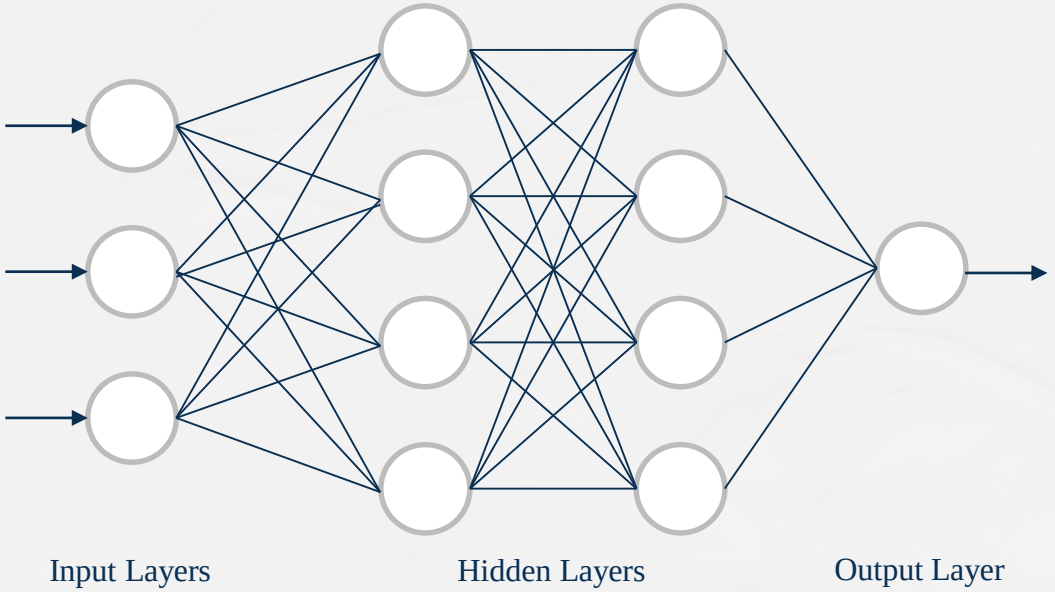
Model Description

Multi-Layer Perceptron (MLP)

Classic neural network with layers of "neurons."

Learns patterns by adjusting connections (weights) between neurons.

Good at finding relationships in clean data, struggles a bit with noisy, chaotic rainfall like Dhaka's!

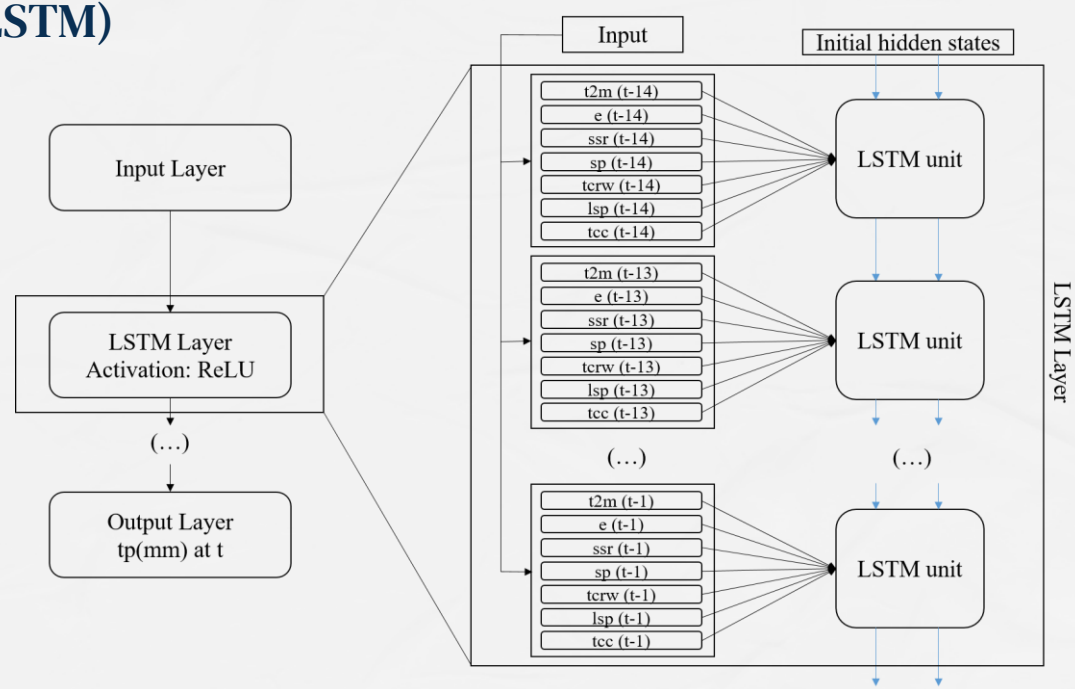


Architecture of a typical Multi-Layer Perceptron model

Model Description

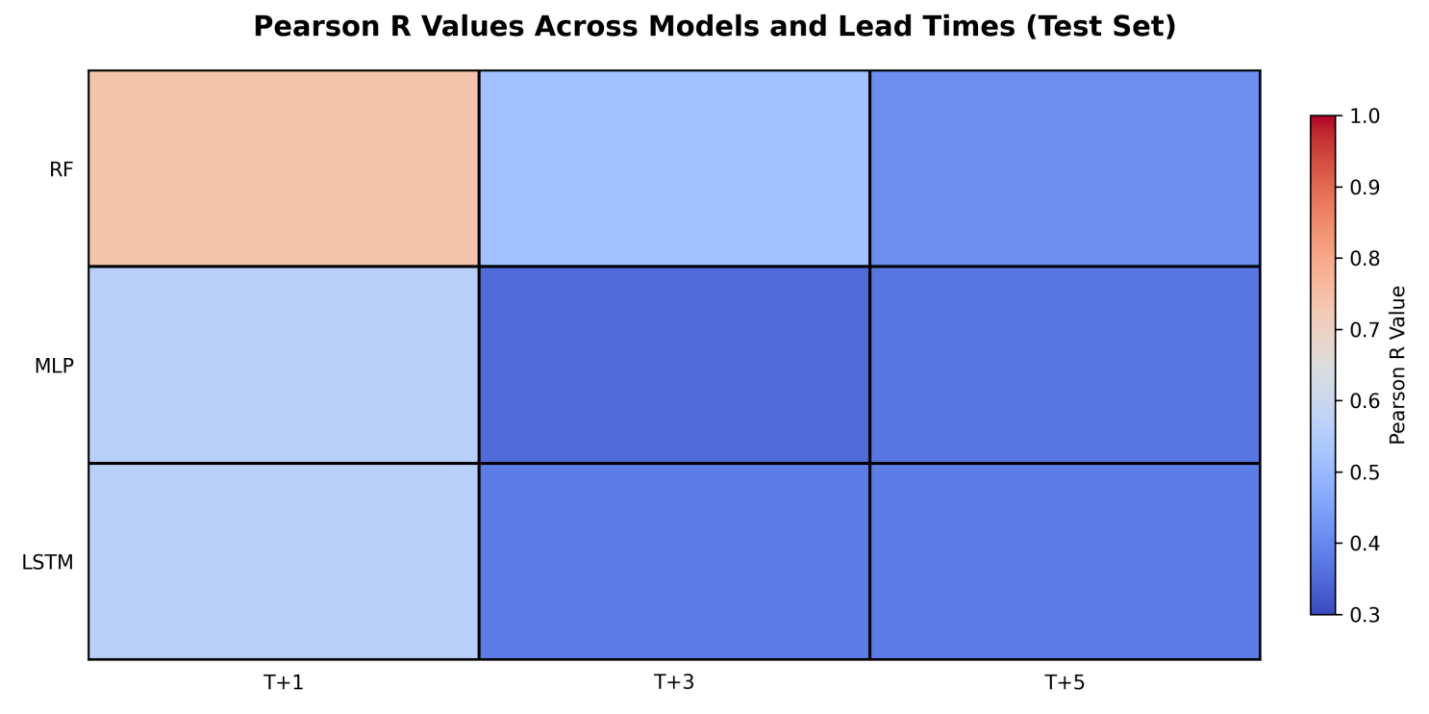
Long Short-Term Memory (LSTM)

- Designed to remember past weather patterns.
- Looks back at **14 days** of data to predict future rainfall.
- Works best for long, smooth trends, but Dhaka's rainfall is too unpredictable!



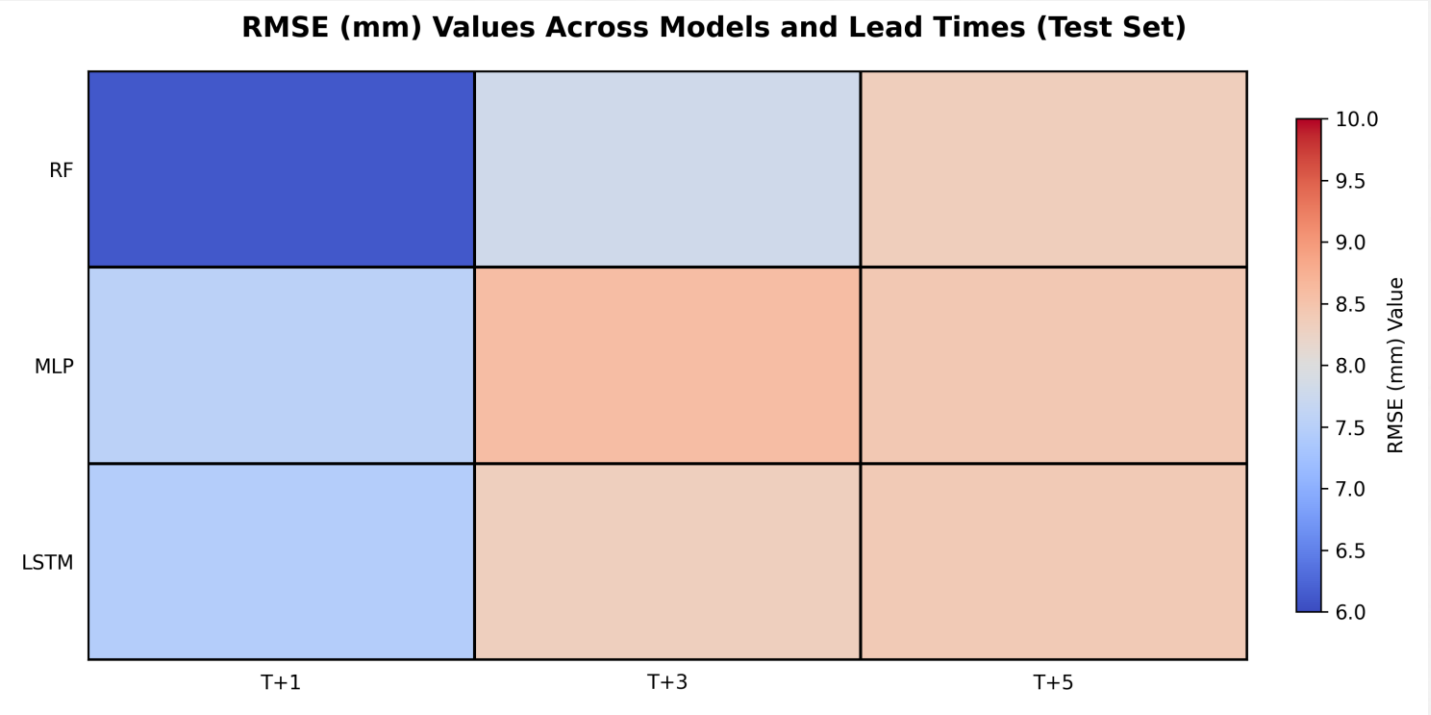
Architecture of the LSTM model used to forecast rainfall

Result & Discussion



The RF model exhibits higher Pearson R values than the other models

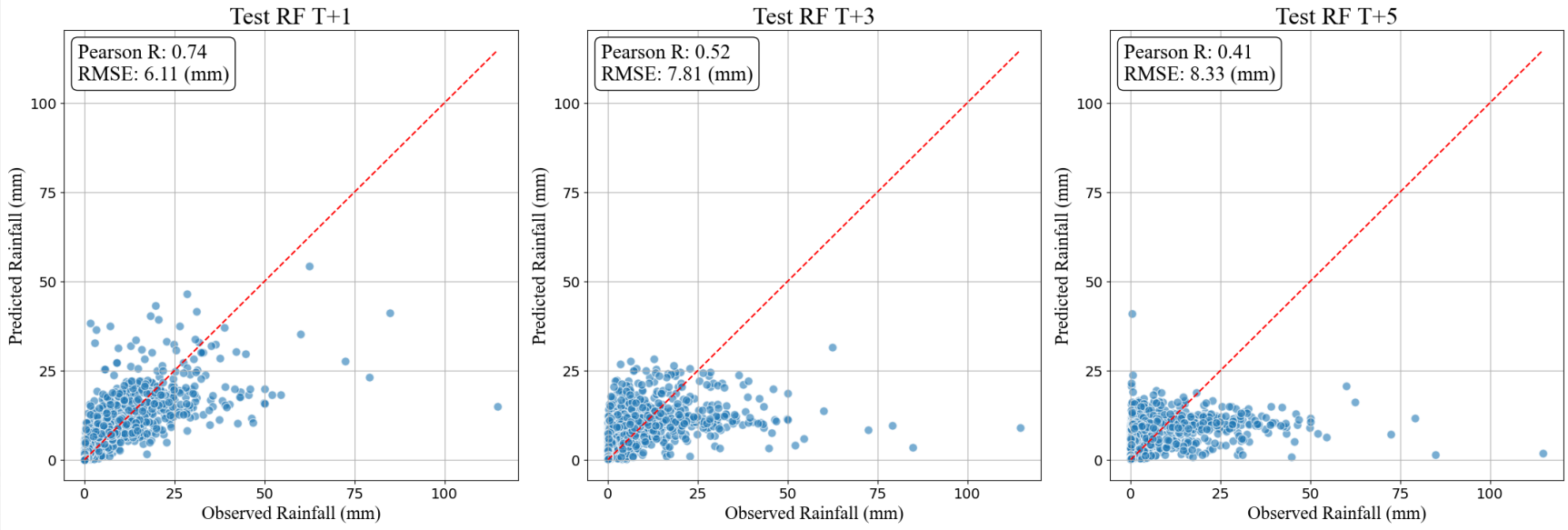
Result & Discussion



The RF model exhibits lower RMSE than the other models, indicating better forecast accuracy

Result & Discussion

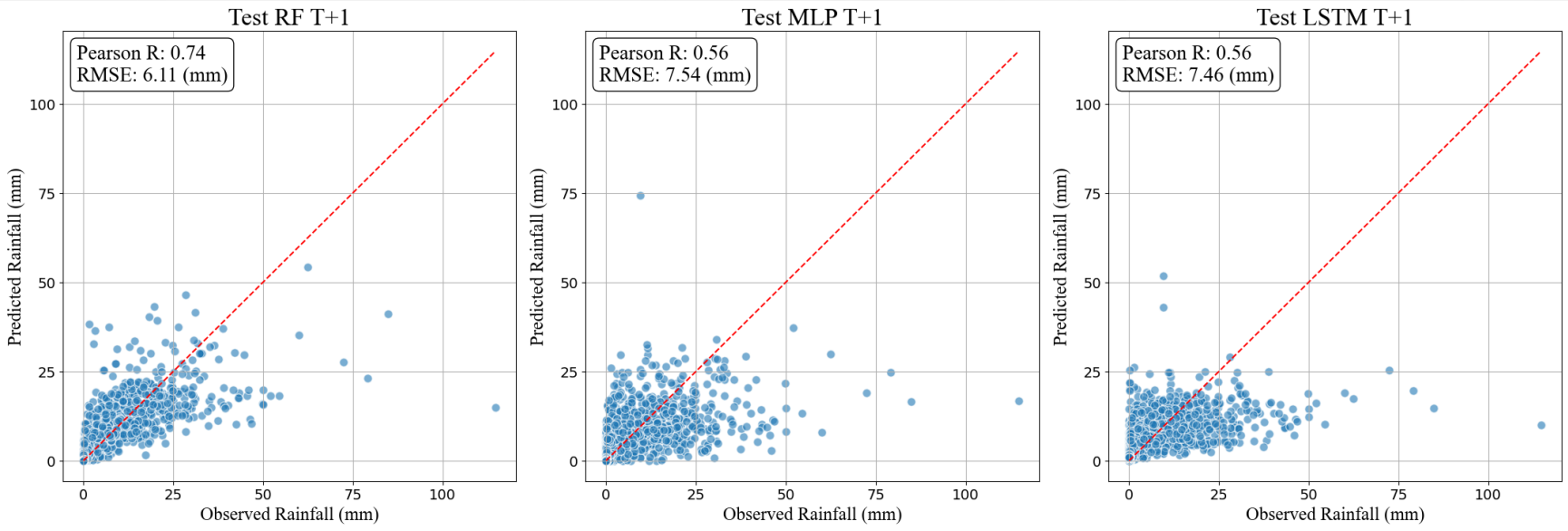
Regression Plots for Random Forest Model



The model performance decreases as the lead time increases

Result & Discussion

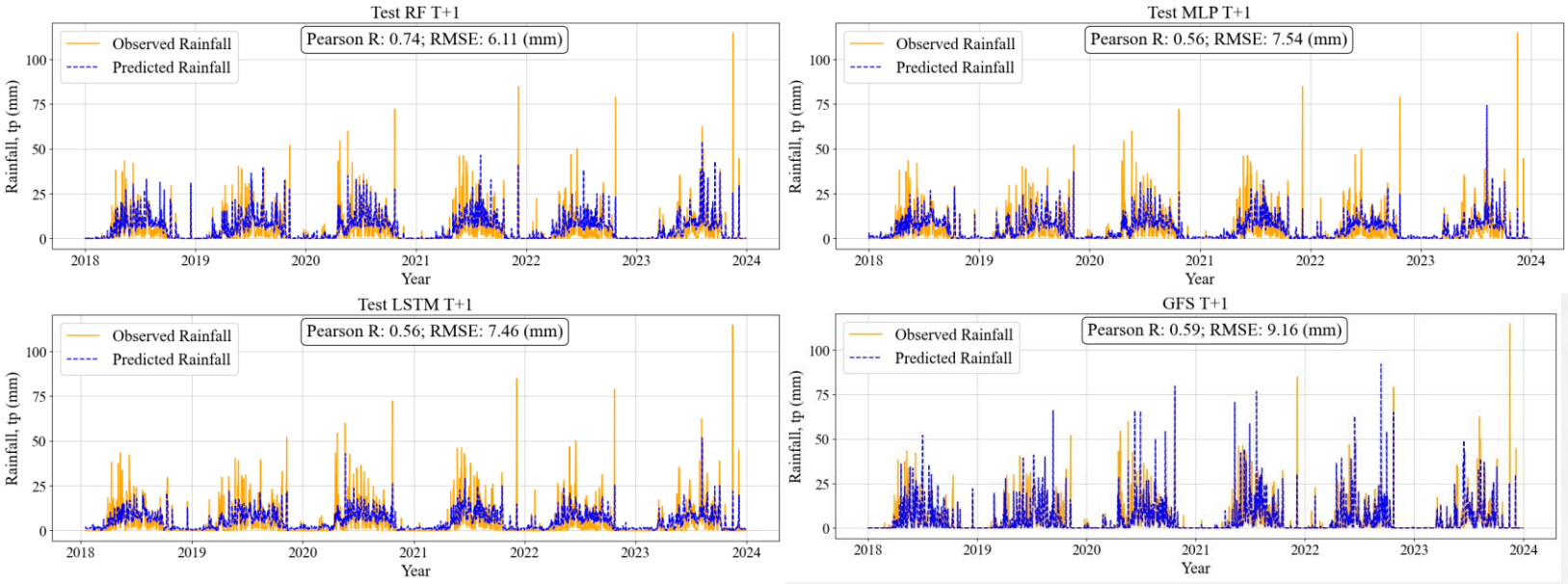
Regression Plots for T+1 Day Lead Time



The Random Forest model performs significantly better than the other models

Result & Discussion

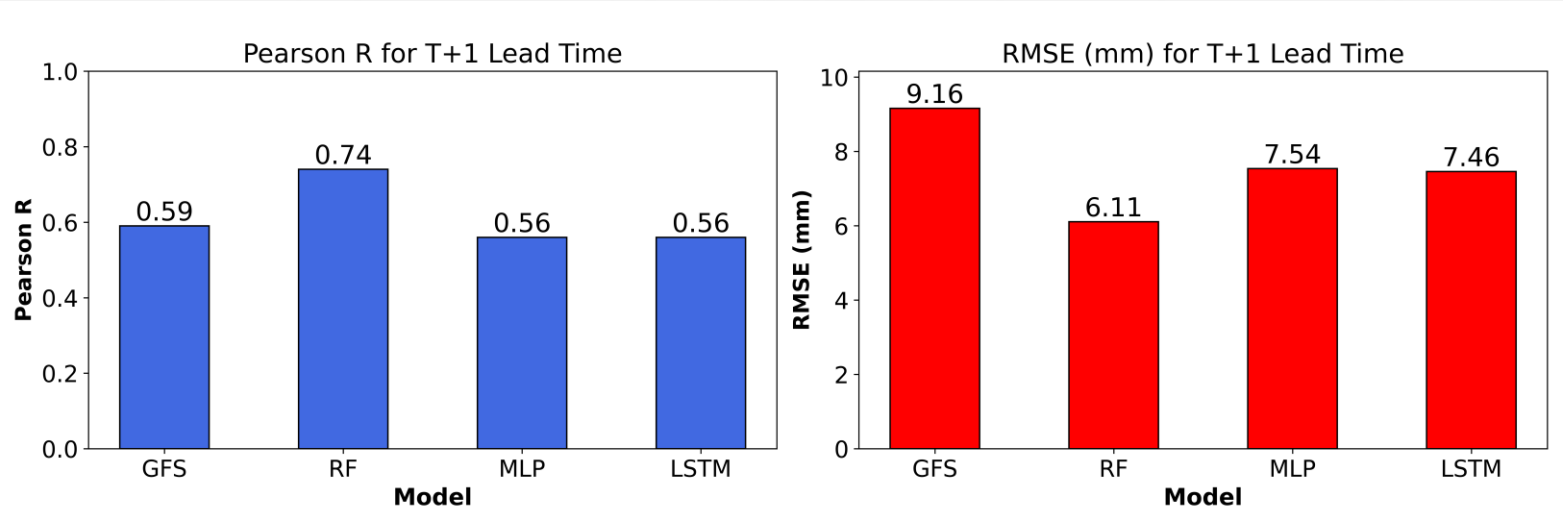
Time Series Plots for T+1 Day Lead Time



The Random Forest model tends to capture complex patterns and outliers more effectively than other methods

Result & Discussion

Comparison of Metrics for T+1 Day Lead Time



The Random Forest model performs significantly better even than the traditional model

Conclusions

- **T-day ahead forecasts performed best**, with shorter lead times showing higher accuracy
- As lead time increased, accuracy declined
- **Random Forest (RF)** outperformed all models with the highest accuracy.
- Long-term forecasting is challenging; all models struggled with increasing uncertainty at longer lead times

Conclusions

Random Forest's strength lies in its ensemble decision tree approach, capturing complex patterns without relying on temporal dependencies.

Feature Selection & Correlation

Some predictor variables may be highly correlated, affecting model performance. RF handles this better by automatically selecting important features, while MLP and LSTM may struggle with multicollinearity.

Temporal & Spatial Dependencies

While LSTM is designed to capture long-term dependencies. RF performs well with less sequential data and can capture short-term fluctuations better.

Limitations & Recommendations

- **Incorporating real-time data sources**
- **Exploring multi-model ensemble methods**
- **Incorporating additional input variables**

Thank you for your patience!

References

- MacroTrends. (n.d.). https://www.macrotrends.net/cities/20119/dhaka/population
- Weather and Climate. (n.d.). *Average monthly precipitation - Rainfall, Dhaka, Bangladesh.* https://weather-and-climate.com/average-monthly-precipitation-Rainfall,Dhaka,Bangladesh
- Picture: https://bdnews24.com/bangladesh/b1b7a43dc9be, https://images.app.goo.gl/3zaiFTCcwLhX94Zc9
- https://images.app.goo.gl/czCBzt6gaRHNRn238
- https://medium.com/@bhatadithya54764118/day-16-random-forests-enhancing-decision-trees-for-improved-predictions-6a8a32134b7e