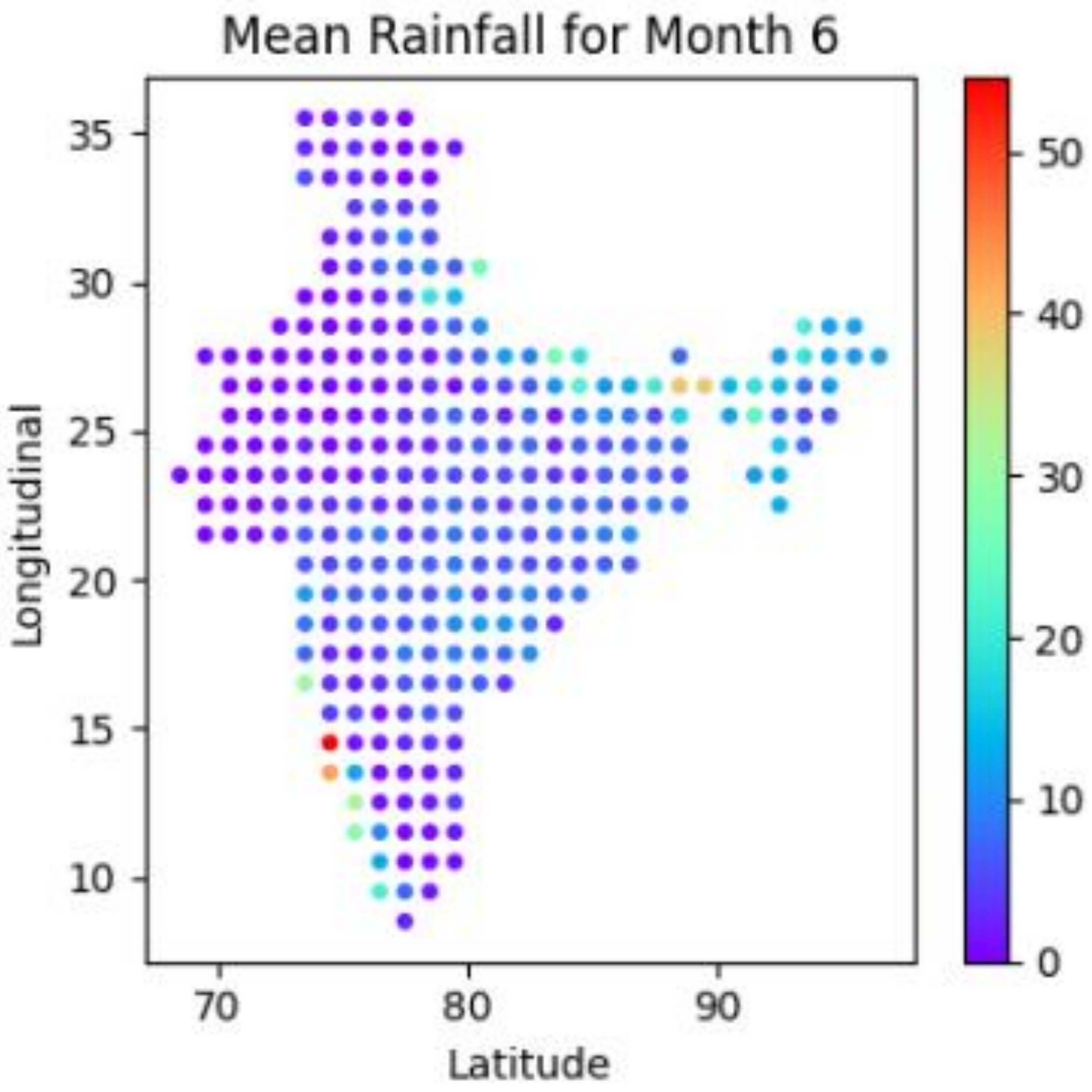


Data Collected across 300 location in India

RAINFALL PREDICTION USING IMD GRIDDED DATASET

Capstone Project
Group 6

Team Members

- Wasudeo Gurjalwar
- Bishwajeet Kumar
- Praveen
- Anantraj Jadhav
- Visali
- Suchitra
- G Sai Balaguru

Project Description

Project Focus:

- Predicting rainfall using historical data from the Indian Meteorological Department (IMD) gridded dataset (2000-2023).

Data Source:

- Daily rainfall observations from 300 locations across India.

Goal:

Develop a predictive model for accurate rainfall forecasting.

Approach:

Use multiple machine learning models

- Supervised Linear regression model
- SARIMA (Seasonal Autoregressive Integrated Moving Average)
- Hybrid deep learning model (CNN + LSTM)

Methodology:

- Train and evaluate each model on the dataset.
- Perform comparative analysis using evaluation metrics:
- (Mean Absolute Error (MAE) , Root Mean Square Error (RMSE), R-squared)

Probable Use cases:

- Identify the most effective model for Water resource management, Agricultural planning, Disaster mitigation

Expected Outcome:

- Insights into the effectiveness of models for improving meteorological predictions.

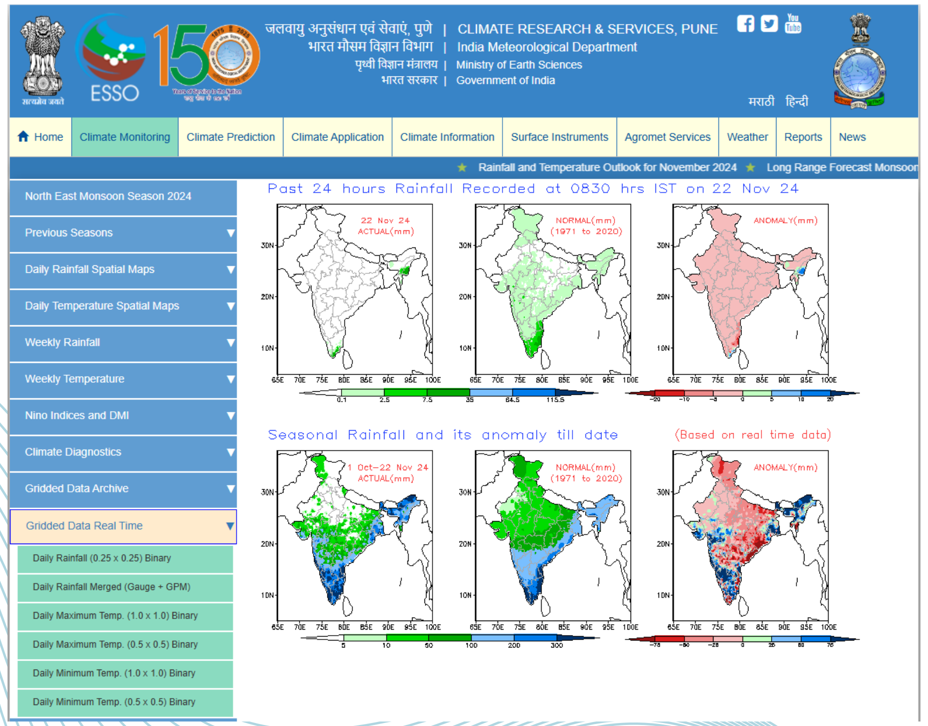
ML Methodologies

	Linear Regression	SARIMA	CNN + LSTM
Model Type	Supervised learning (regression)	Time series forecasting (AR, I, MA components)	Deep learning (Hybrid: CNN for spatial, LSTM for temporal)
Data Handling	Assumes linear relationship between features and output	Requires stationary data; suitable for univariate time series	Handles complex, multi-dimensional datasets, captures both spatial and temporal patterns
Complexity	Simple, computationally efficient	Moderate complexity, requires parameter tuning	High complexity, requires significant computational resources
Performance and Accuracy	Good for linear relationships, limited for complex data	Effective for time series with strong temporal patterns	High accuracy for complex, non-linear relationships, long-term dependencies
Interpretability	Highly interpretable (clear relationship between variables)	Moderate interpretability (autocorrelation)	Low interpretability (often considered a "black box")
Training Time	Very fast, suitable for small datasets	Moderate training time, slower for large datasets	Long training time, requires significant computational power
Handling Seasonality & Trends	Needs manual feature engineering for trends and seasonality	Can handle seasonality and trends with appropriate parameters	Automatically learns seasonality and trends from data
Scalability	Easily scalable to large datasets with fewer features	Scalable, but performance degrades with very large datasets	Highly scalable for large, complex datasets, but requires more resources
Use Case Suitability	Best for simpler tasks with linear relationships	Best for univariate time series forecasting with trends and seasonality	Best for complex, high-dimensional, sequential data with spatial and temporal dependencies
Strengths	Fast and easy to implement, interpretable	Effective for time series with stationarity and autocorrelation	Accurate for complex patterns, handles both spatial and temporal dependencies
Weaknesses	Limited for non-linear or complex relationships	Struggles with non-stationary data and multiple variables	Requires large data, computationally intensive, less interpretable

Data set

IMD
Pune
[Climate Research & Services, Pune]

Get RAW Data from IMD Pune

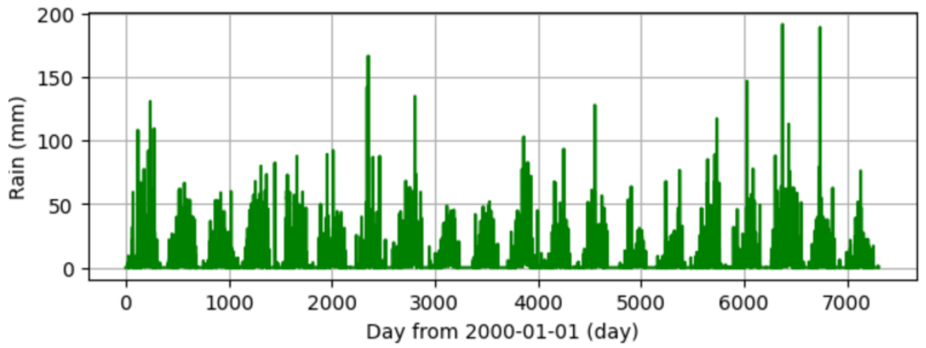


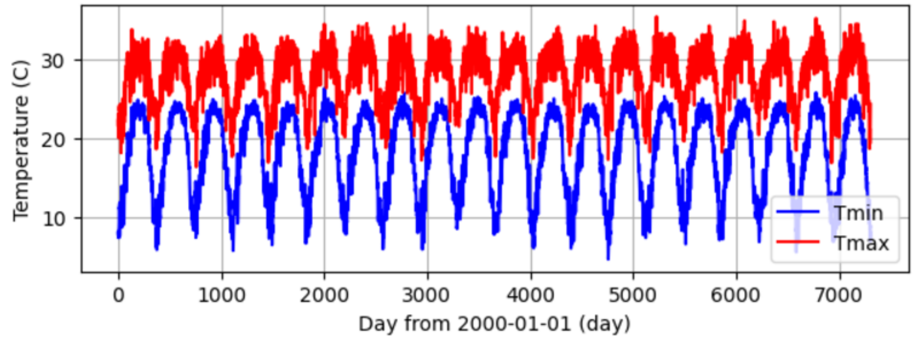
Convert RAW Data to Interim Data

- Handling Missing Data
- Combining various parameter Data (rain, Temp)
- Date conversion to day, week & month

Processed Data (Daily)

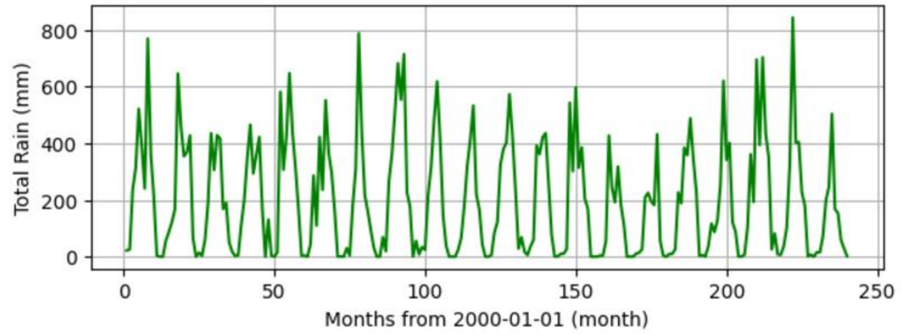
- Day
- Rain
- TempMin
- TempMax

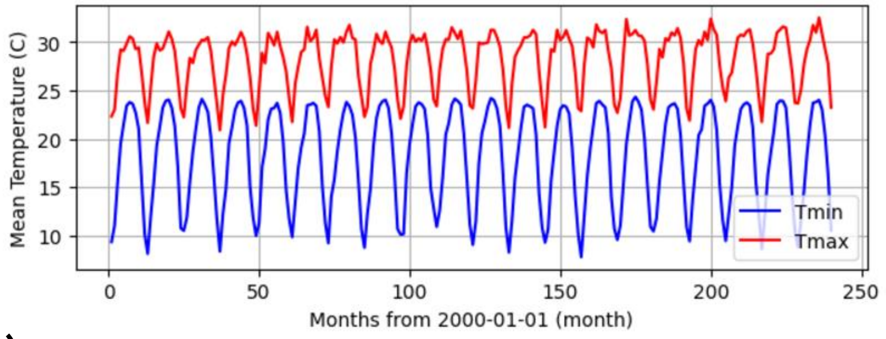




Processed Data (Monthly)

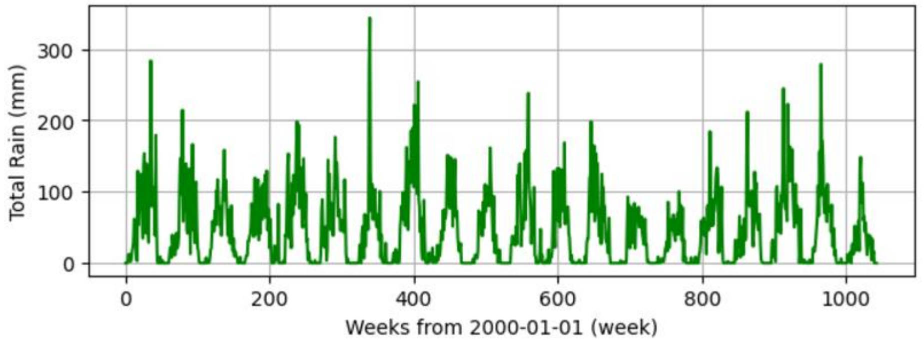
- Month
- Mean Rain
- Cumulative Rainfall
- Average Temperature (Max + Min)

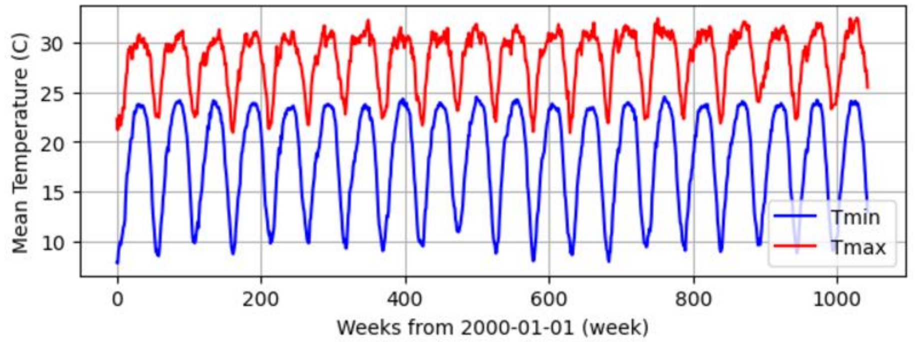




Processed Data (Weekly)

- Week
- Mean Rain
- Cumulative Rainfall
- Average Temperature (Max + Min)

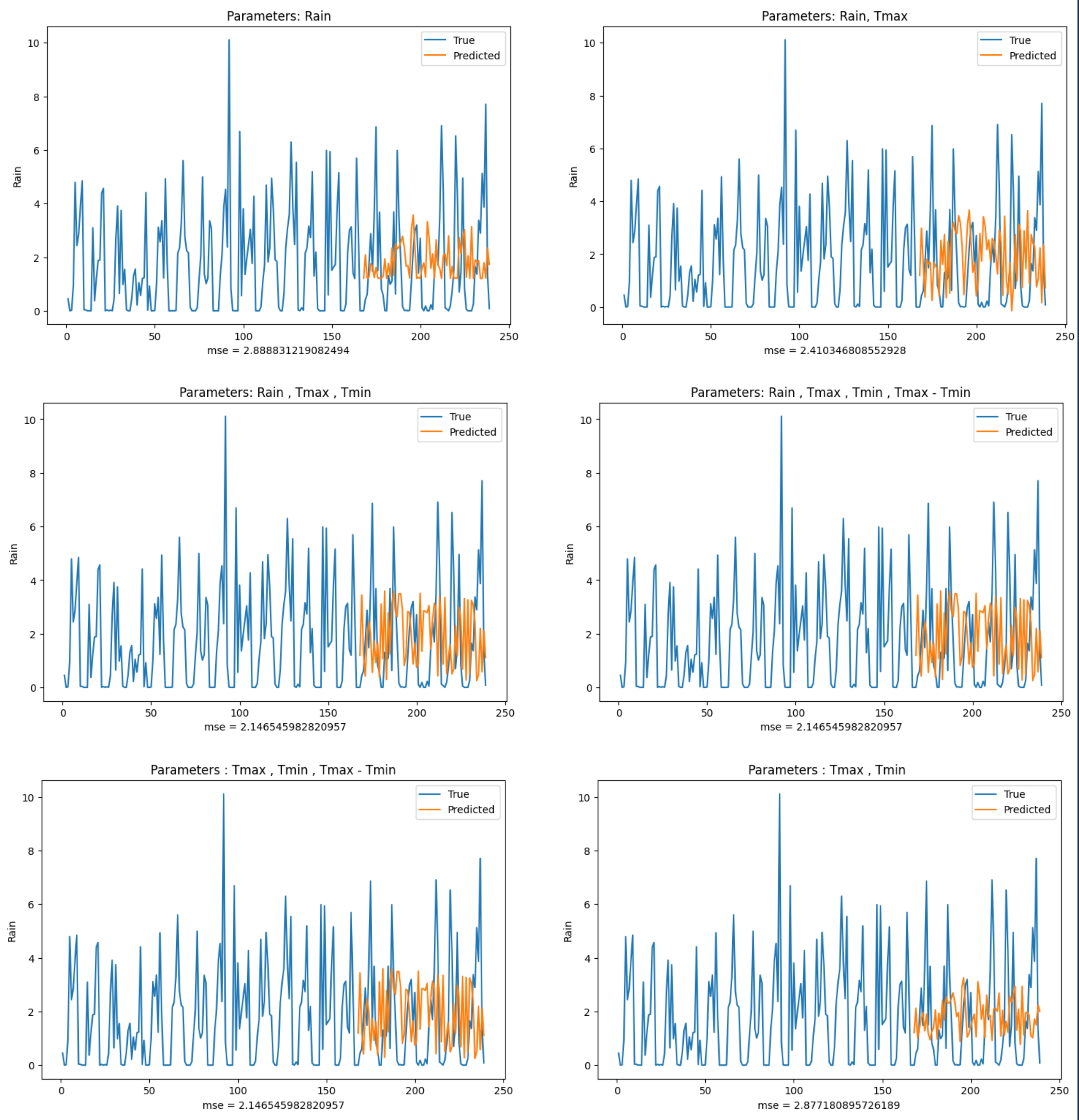




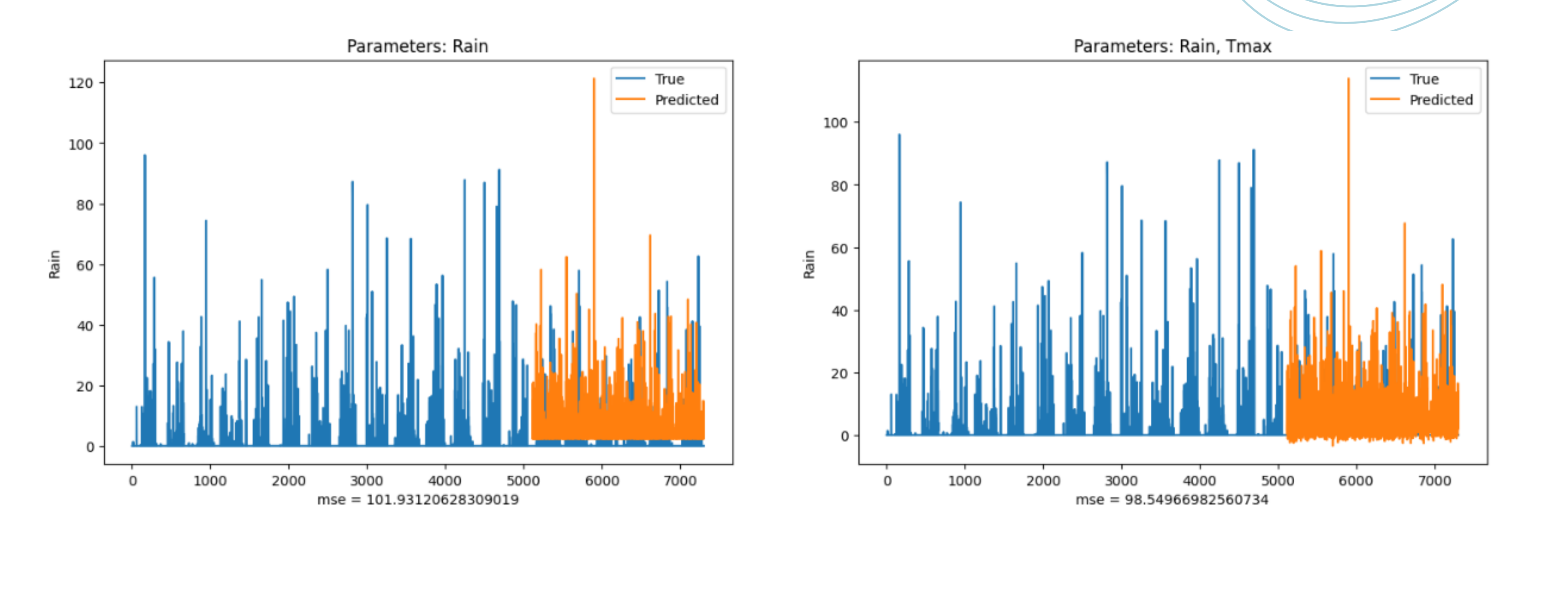
Result (Linear Regression)

Combination of various Parameters (rain, Tmax, Tmin, Tmax-Tmin

Month Prediction for Single Location



Daily Prediction for Single Location



Optimal Parameters for each Location

```
  position,          LRcoef_rain, LRcoef_tmax LRcoef_tmin LRintercept
"['(10.5, 76.5)']",[ 0.4672153  -1.36820134  1.17267643],19.440931413856823
"['(10.5, 77.5)']",[ 0.4672153  -1.36820134  1.17267643],19.440931413856823
"['(10.5, 78.5)']",[ 0.4672153  -1.36820134  1.17267643],19.440931413856823
"['(10.5, 79.5)']",[ 0.4672153  -1.36820134  1.17267643],19.440931413856823
```

Observation Based on monthly and Daily Data Set

Monthly data set

- the error is reducing when we are using more parameters for prediction like Tmin, Tmax

Daily Data set

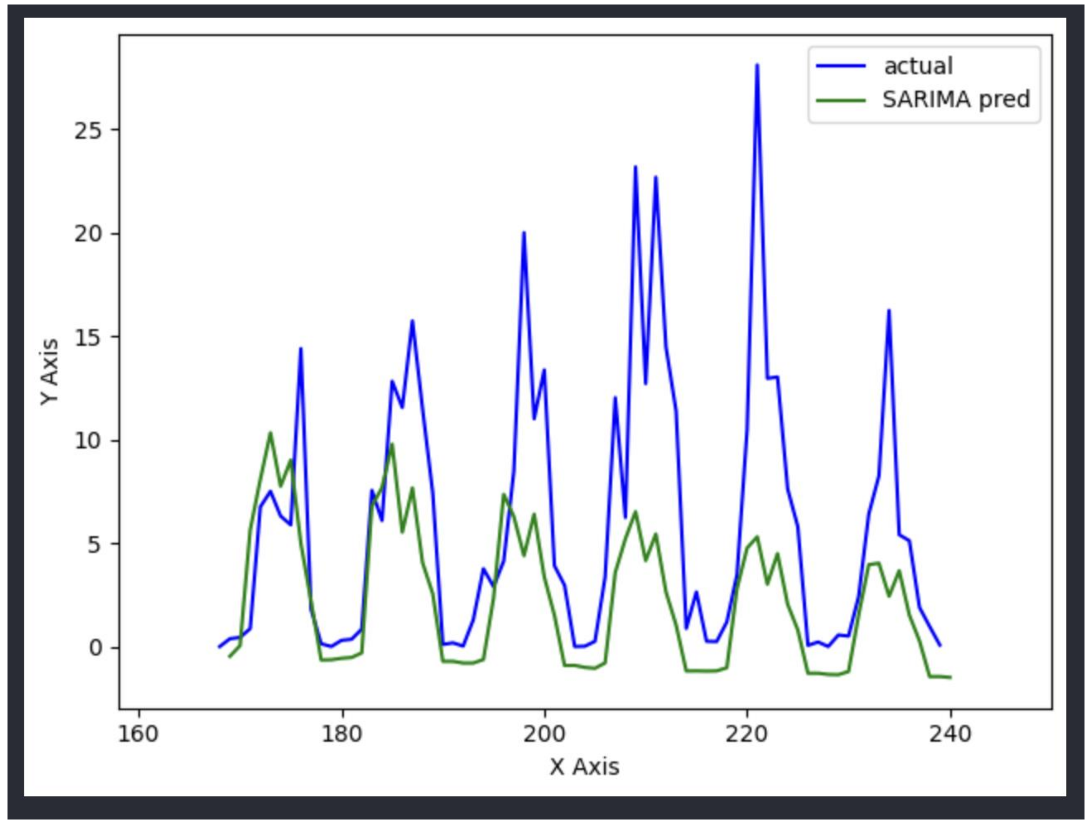
- The error is more as the numbers of zeros (no rain fall days are more) are more and we should not remove this data and manipulate as well

Result (SARIMA)

Validation dataset Rainfall Prediction

Hyperparameter Tunning of SARIMA

Month Prediction for Single Location

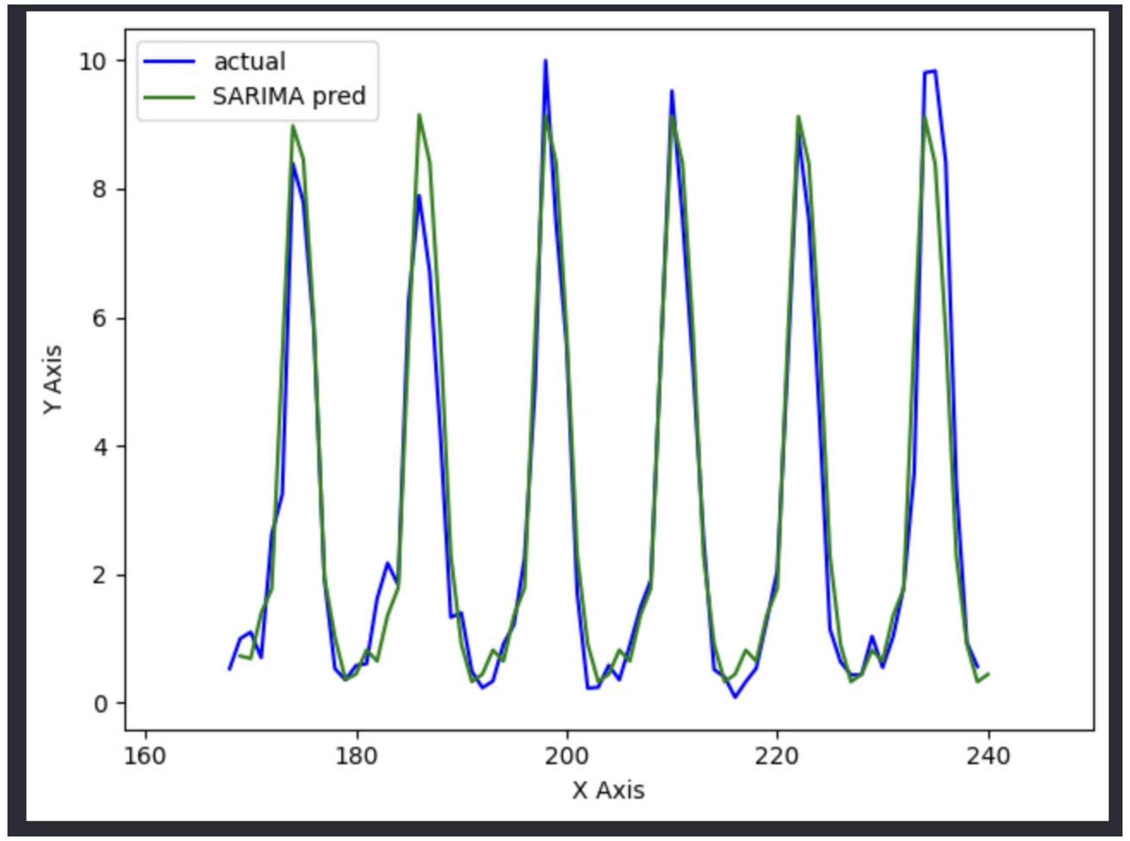



```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                               y   No. Observations:                    168
Model:             SARIMAX(1, 1, 0)x(3, 0, 0, 12)   Log Likelihood                -496.125
Date:                         Fri, 22 Nov 2024   AIC                          1002.249
Time:                                 17:39:03   BIC                          1017.839
Sample:                                      0   HQIC                         1008.577
                                         - 168
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4096      0.063     -6.495      0.000      -0.533      -0.286
ar.S.L12       0.1013      0.064      1.589      0.112      -0.024       0.226
ar.S.L24       0.2986      0.083      3.606      0.000       0.136       0.461
ar.S.L36       0.3494      0.084      4.151      0.000       0.184       0.514
sigma2        20.9263      1.698     12.323      0.000      17.598      24.254
===================================================================================
Ljung-Box (L1) (Q):                   1.14   Jarque-Bera (JB):                62.94
```

Optimal Parameters for each Location

1	pos	param	trainMSE	valMSE
2	(10.5, 76.5)	((2, 1, 1), (2, 0, 2, 12))	12.935	33.141
3	(10.5, 77.5)	((2, 1, 1), (0, 0, 1, 12))	8.049	8.332
4	(10.5, 78.5)	((1, 1, 2), (2, 0, 1, 12))	8.305	5.648
5	(10.5, 79.5)	((2, 1, 1), (1, 0, 1, 12))	16.336	18.629
6	(11.5, 75.5)	((1, 0, 0), (2, 0, 1, 12))	38.281	100.936
7	(11.5, 76.5)	((0, 1, 4), (3, 0, 0, 12))	10.092	14.189

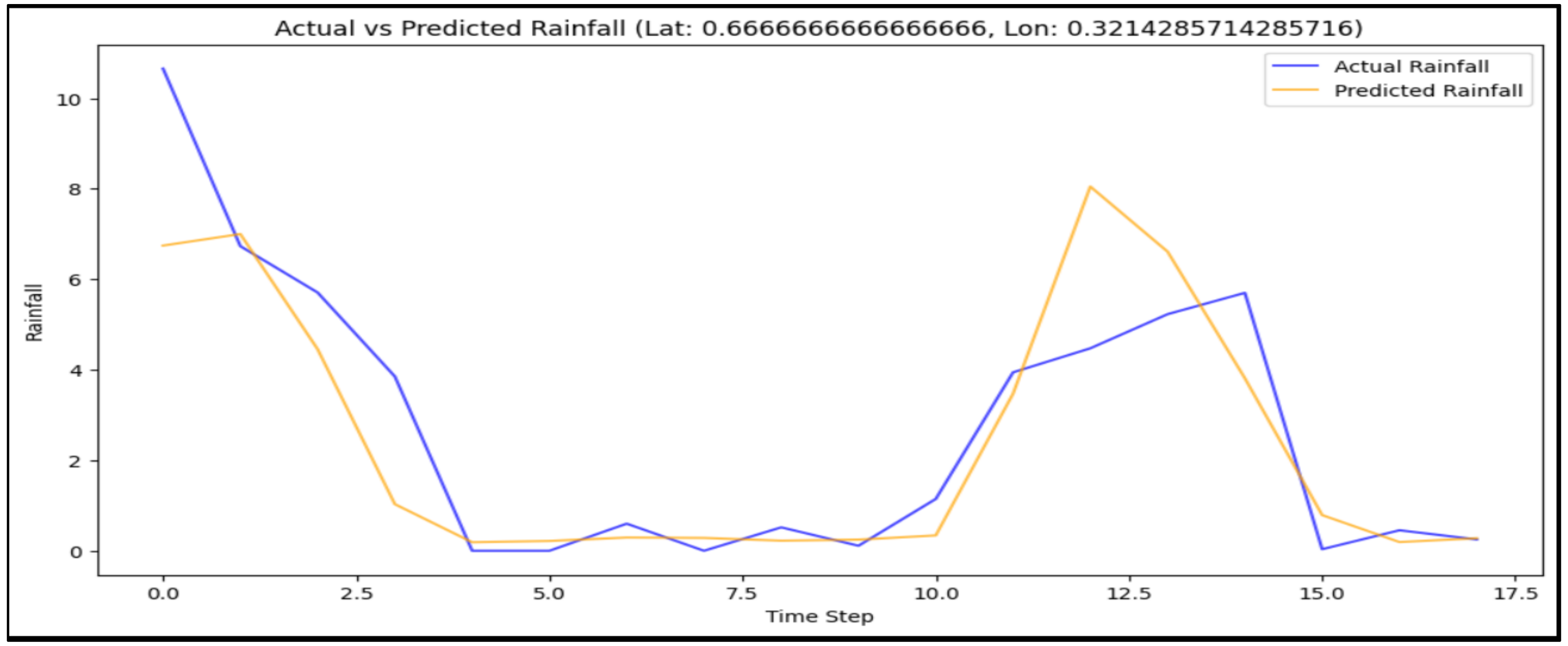
Month Prediction for all location (Average)



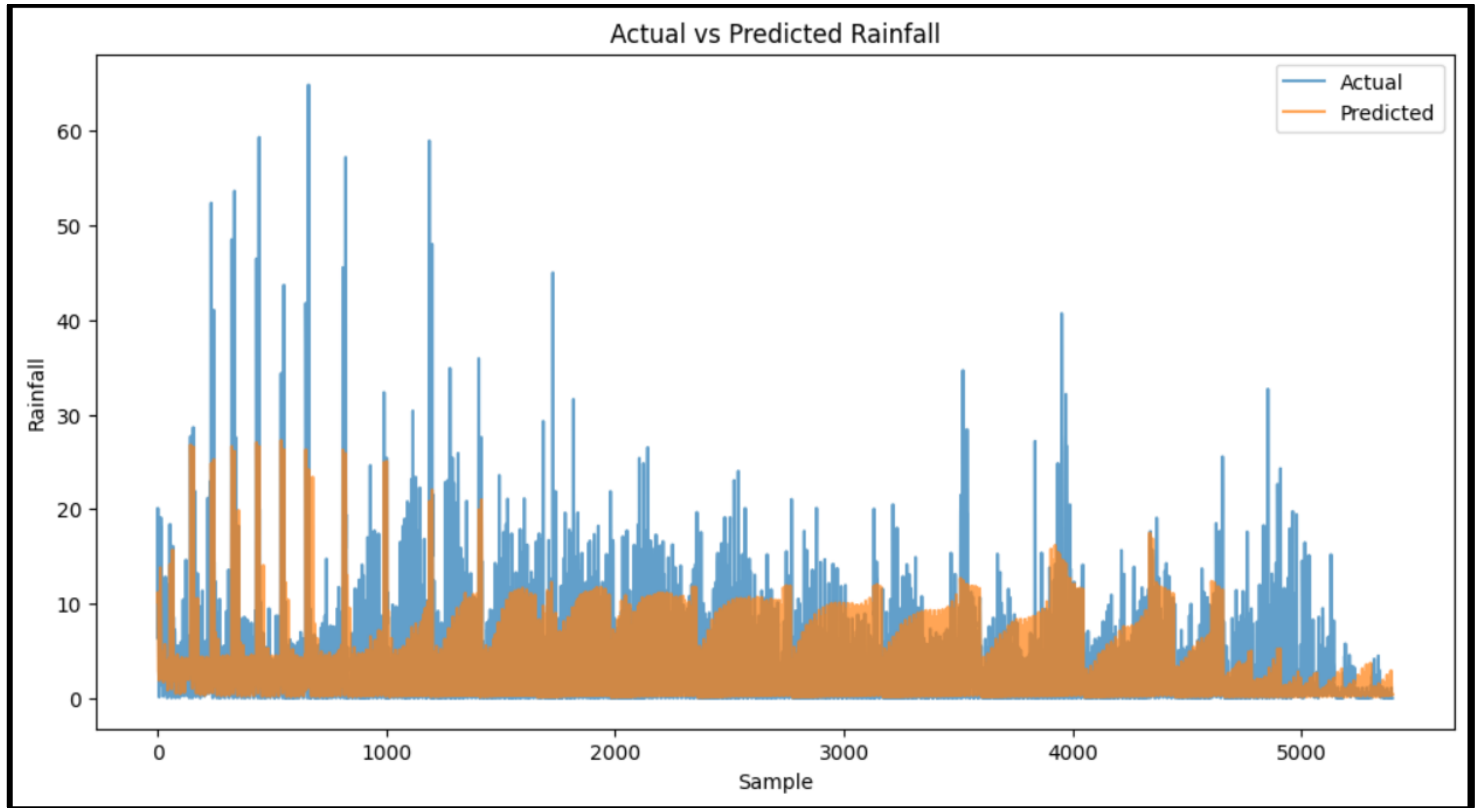

```
                                  SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  168
Model:             SARIMAX(1, 0, [1, 2], 12)   Log Likelihood                -228.160
Date:                         Thu, 21 Nov 2024   AIC                            466.321
Time:                                 21:49:58   BIC                            481.941
Sample:                                      0   HQIC                           472.660
                                         - 168
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0054      0.008      0.722      0.470      -0.009       0.020
ar.S.L12       0.9982      0.002    504.406      0.000       0.994       1.002
ma.S.L12      -0.9364      0.069    -13.582      0.000      -1.072      -0.801
ma.S.L24       0.1532      0.067      2.290      0.022       0.022       0.284
sigma2         0.6800      0.061     11.197      0.000       0.561       0.799
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                25.51
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               1.09   Skew:                             0.02
Prob(H) (two-sided):                  0.74   Kurtosis:                         4.91
===================================================================================
```

Result (CNN +LASTM)

Validation Prediction for one Location



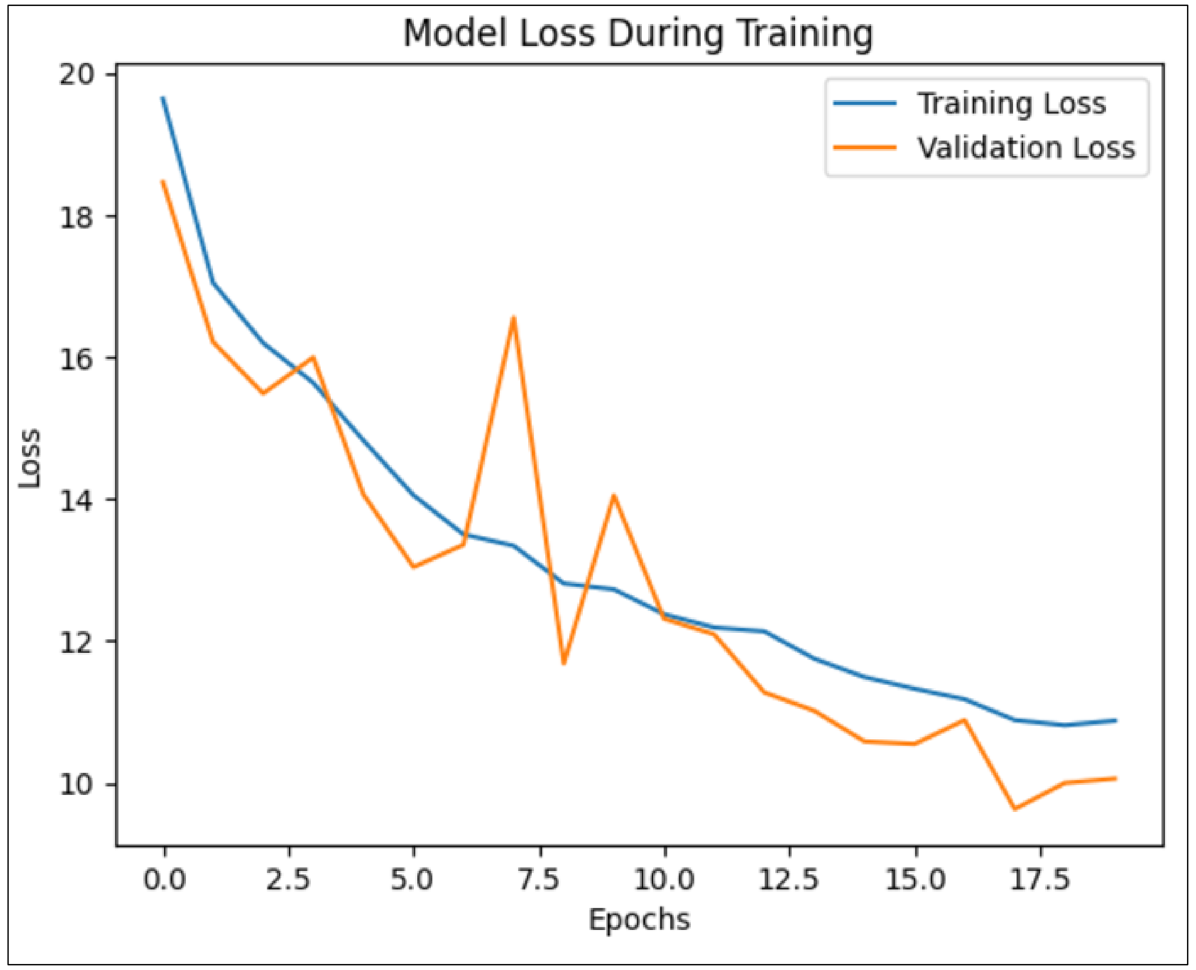
Validation Prediction for All Location



Model Summary

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 28, 64)	1,024
max_pooling1d (MaxPooling1D)	(None, 14, 64)	0
dropout (Dropout)	(None, 14, 64)	0
lstm (LSTM)	(None, 14, 50)	23,000
dropout_1 (Dropout)	(None, 14, 50)	0
lstm_1 (LSTM)	(None, 50)	20,200
dense (Dense)	(None, 1)	51

Total params: 44,275 (172.95 KB)
Trainable params: 44,275 (172.95 KB)
Non-trainable params: 0 (0.00 B)



Result Testing Evaluation Metric

#	Location	MSE	RMSE	R2
Linear Regression	Single	2.15	1.466	0.358
SARIMA Type 1	Single	4.4	2.09	0.218
SARIMA Type 2	Single	1.8	1.341	0.545
CNN + LSTM	All	1.8	1.34	0.629

Contribution

Name →	Wasudeo	Bishwajeet	Praveen	Anantraj	Vishali	Suchitra	G Sai Balaguru	Total
Data cleanup and acquisition	10	10	70	10	-	-	-	100
ML Model Selection/ Training	33.33	33.33	33.33	-	-	-	-	100
Hyper parameter tuning	33.33	33.33	33.33	-	-	-	-	100
Metrics	33.33	33.33	33.33	-	-	-	-	100
Presentation	90	-	10	-	-	-	-	100
Documentation	20	40	40	-	-	-	-	100

Conclusion

Supervised Learning Regression

- Linear Regression can be interpreted easily
- Linear regression give rain fall prediction as a function of previous month rain fall and temps
- Adding more parameter helps in reducing the error

- Supervised SARIMA
  - Linear Regression can be interpreted easily
  - SARIMA helps in identifying patterns, trends and seasonality from the past data but lag in incorporating other parameters
  - Further tunning is possible to reduce the error

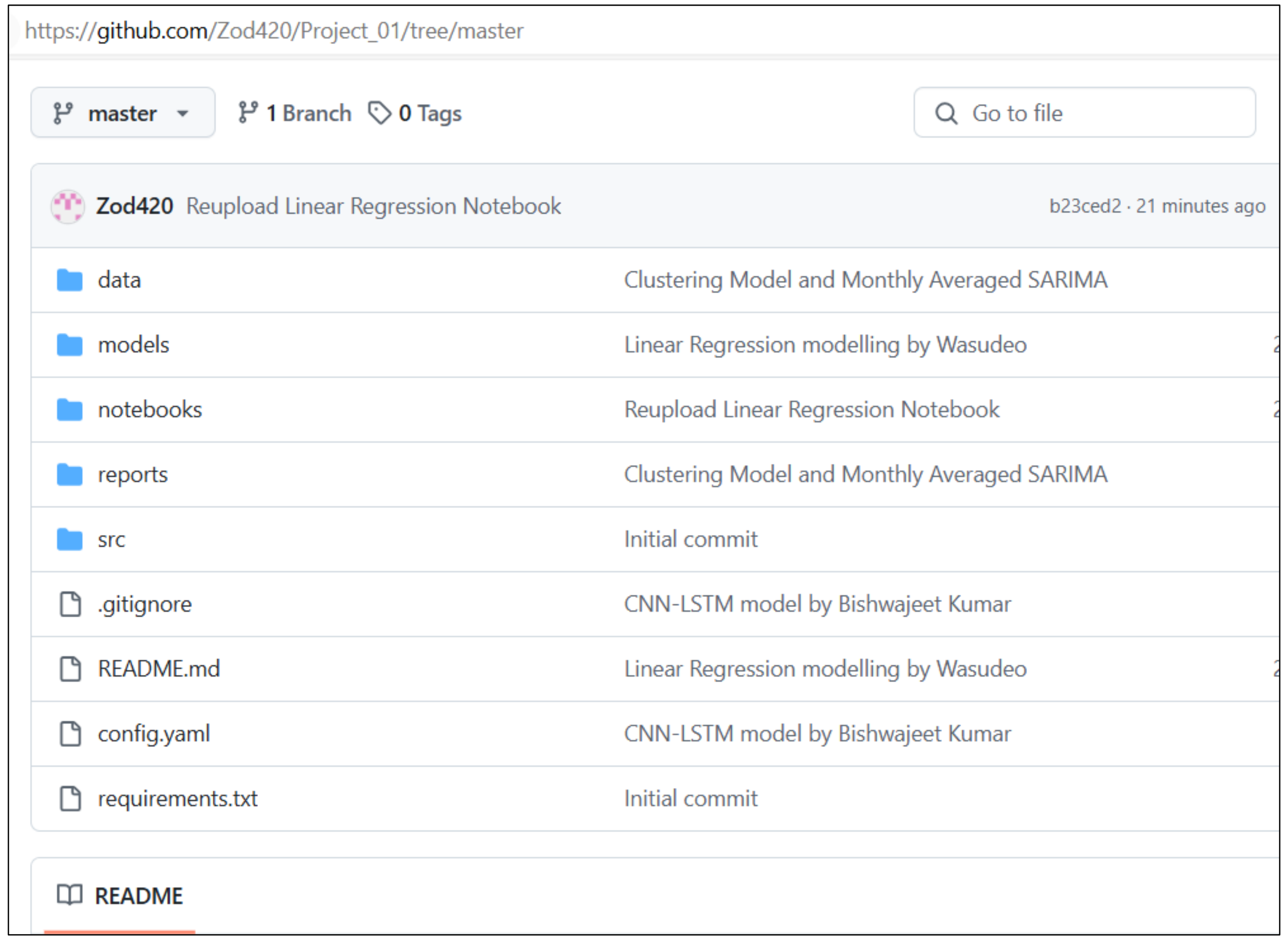
Deep learning (CNN + LSTM)

- Can incorporate location also to have a single model for any location
- Easy to add on more feature to incorporate the prediction
- Further tunning is possible to reduce the error

Future Action :

- Exploration of weekly data to check the accuracy of the models
- More parameters can be incorporated to have more accurate rain fall prediction such as humidity, wind speed altitude pressure etc.

Git Hib Repo ([GitHub - Zod420/Project_01: IISC ML project](https://github.com/Zod420/Project_01/tree/master))



THANK YOU!

For Your Support and Cooperation