Prediction of Rainfall Pattern Using CHIRPS High-Resolution Data over India in AIML & DATA SCIENCE Project

1. Introduction

Rainfall plays a critical role in shaping the socioeconomic fabric of India, which is predominantly an agrarian country. Predicting rainfall patterns is vital for agriculture, water resource management, disaster planning, and climate studies. The complexity of India's monsoons, influenced by diverse geographical features and global climatic phenomena, makes rainfall prediction a challenging task.

The advent of Artificial Intelligence and Machine Learning (AIML) and Data science (DS) has revolutionized climate science, offering advanced tools to analyze large datasets and predict patterns. This study leverages Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) high-resolution rainfall data to explore and predict rainfall patterns across India. Using AIML, we attempt to understand trends, identify anomalies, and evaluate the variability of rainfall across different regions of India.

2. Study Area

India is geographically diverse, comprising five major regions: North, South, East, West, and Central. Each region experiences distinct climatic conditions influenced by factors such as topography, proximity to water bodies, and prevailing wind systems. This study analyzes rainfall data from these regions to understand spatial and temporal variability.

Data and Methodology

Data Source:

- The CHIRPS dataset provides high-resolution, long-term precipitation data.
- The dataset for this study includes daily rainfall values from 2012 to 2023.

- Additional information is obtained from the India Meteorological Department (IMD) website: www.imd.gov.in.

Data Overview:

- Variables: Rainfall (in mm), Latitude, Longitude, Time (daily).
- Format: NetCDF (.nc).
- Coverage: Spatial coverage over India with high-resolution grids.

	LONGITUDE	LATITUDE	TIME	RAINFALL
0	66.5	6.5	2023-01-01	NaN
1	66.5	6.5	2023-01-02	NaN
2	66.5	6.5	2023-01-03	NaN
3	66.5	6.5	2023-01-04	NaN
4	66.5	6.5	2023-01-05	NaN
...	...	...	...	...
6356470	100.0	38.5	2023-12-27	NaN
6356471	100.0	38.5	2023-12-28	NaN
6356472	100.0	38.5	2023-12-29	NaN
6356473	100.0	38.5	2023-12-30	NaN
6356474	100.0	38.5	2023-12-31	NaN

[6356475 rows x 4 columns]

Techniques and Tools:

- **Data Extraction**: Python libraries such as xarray and pandas were used to extract and preprocess the data.

  Code:

```
import xarray as xr

import pandas as pd

file_path = 'path_to_nc_file.nc'

dataset = xr.open_dataset(file_path)

df = dataset.to_dataframe().reset_index()

df['TIME'] = pd.to_datetime(df['TIME'])

df['MONTH'] = df['TIME'].dt.month

df['YEAR'] = df['TIME'].dt.year
```

```
data = df[['YEAR', 'MONTH', 'LATITUDE', 'LONGITUDE', 'RAINFALL']]

data = data.dropna(subset=['RAINFALL'])
```

- **Visualization**: Matplotlib, Seaborn, and Cartopy were employed to plot rainfall maps and trends.

  Code:

```
import matplotlib.pyplot as plt

import cartopy.crs as ccrs

plt.figure(figsize=(10, 8))

ax = plt.axes(projection=ccrs.PlateCarree())

ax.coastlines()

sc = plt.scatter(

   data['LONGITUDE'], data['LATITUDE'], c=data['RAINFALL'], cmap='Blues', s=10, transform=ccrs.PlateCarree()

)

plt.colorbar(sc, label='Rainfall (mm)')

plt.title('Rainfall Distribution')

plt.show()
```

- **Machine Learning**: Random Forest Regression was utilized for predictive modeling.

  Code:

```
import xarray as xr

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error

from sklearn.model_selection import train_test_split

# Function to process a single .nc file

def process_nc_file(file_path):
```

```
    dataset = xr.open_dataset(file_path)
    df = dataset.to_dataframe().reset_index()
    df['TIME'] = pd.to_datetime(df['TIME'])
    df['MONTH'] = df['TIME'].dt.month
    df['YEAR'] = df['TIME'].dt.year
    df = df[['YEAR', 'MONTH', 'LATITUDE', 'LONGITUDE', 'RAINFALL']]
    return df
# Paths to the .nc files for years 2014 to 2023
file_paths = [
    "F:\\CSIR project\\RF25_ind2014_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2015_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2016_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2017_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2018_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2019_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2020_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2021_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2022_rfp25.nc",
    "F:\\CSIR project\\RF25_ind2023_rfp25.nc",
]
# Process all files and combine into a single DataFrame
rainfall_data = pd.concat([process_nc_file(fp) for fp in file_paths], ignore_index=True)
# Filter for June, July, August, September (6, 7, 8, 9)
rainfall_data_filtered = rainfall_data[rainfall_data['MONTH'].isin([6, 7, 8, 9])]
# Group by year and month to get average rainfall
data = rainfall_data_filtered.groupby(['YEAR', 'MONTH']).agg({'RAINFALL':
'mean'}).reset_index()
# Handle missing values in the dataset
data = data.dropna(subset=['RAINFALL'])
```

```
# Prepare features (X) and target (y)
X = data[['YEAR', 'MONTH']]
y = data['RAINFALL']
# Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train the Random Forest Regressor
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
# Evaluate the model
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
```

Regional Analysis

North India

North India, encompassing the Himalayan region and the Indo-Gangetic plains, experiences significant rainfall variability. This region is predominantly influenced by the southwest monsoon and western disturbances.

South India

South India receives rainfall from both the southwest monsoon and the northeast monsoon, with the Western Ghats playing a pivotal role in shaping rainfall distribution.

East India

East India, characterized by high humidity and dense forests, receives substantial rainfall due to the Bay of Bengal branch of the southwest monsoon.

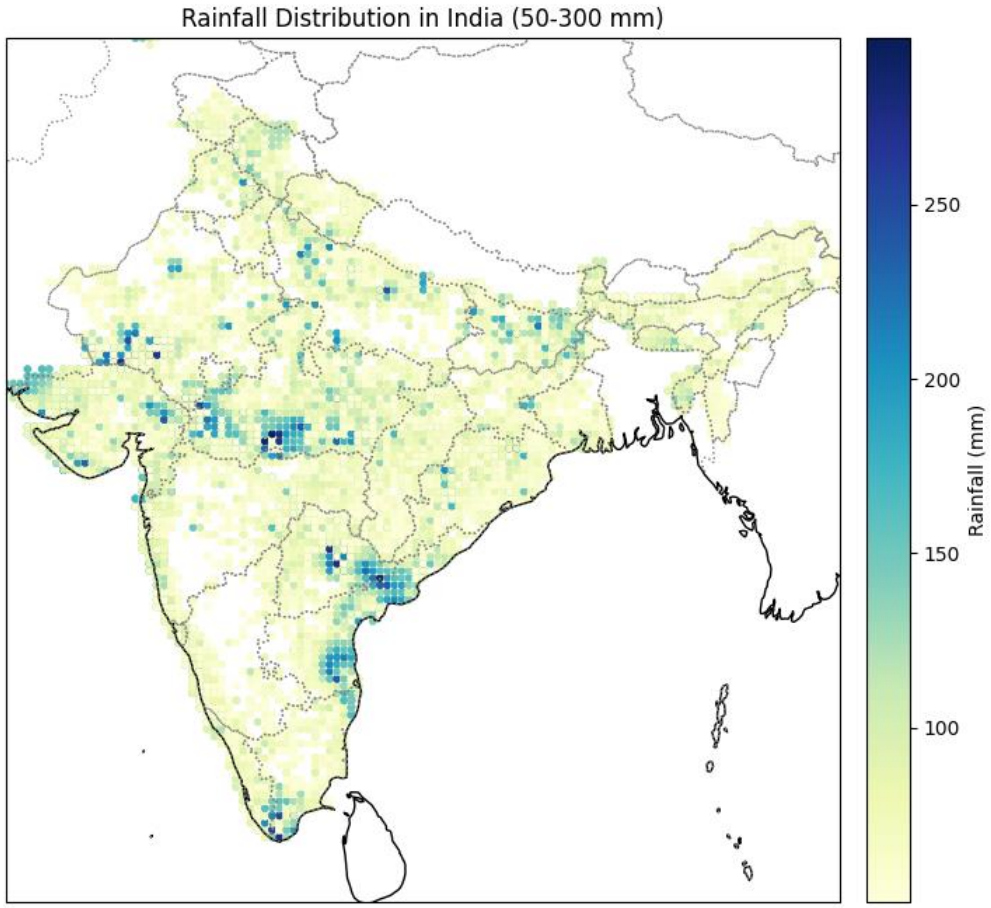
West India

West India, dominated by arid and semi-arid regions, sees stark contrasts in rainfall between coastal areas and interior zones.

Central India

Central India forms the heart of the Indian subcontinent, experiencing moderate rainfall, crucial for its agrarian economy.

Fig: 2023 Rainfall distribution



Methodology

Data Analysis

1. **Data Cleaning**: Addressing missing values and removing anomalies.
2. **Exploratory Data Analysis (EDA)**: Identifying trends, seasonal patterns, and regional variability.
3. **Visualization**: Heatmaps, line plots, and bar charts to depict spatial and temporal distribution.

Machine Learning Approach

1. Feature Selection: Year, month, and geographical coordinates.
2. Model: Random Forest Regression for predicting rainfall.
3. Evaluation: Mean Squared Error (MSE) to assess model performance.

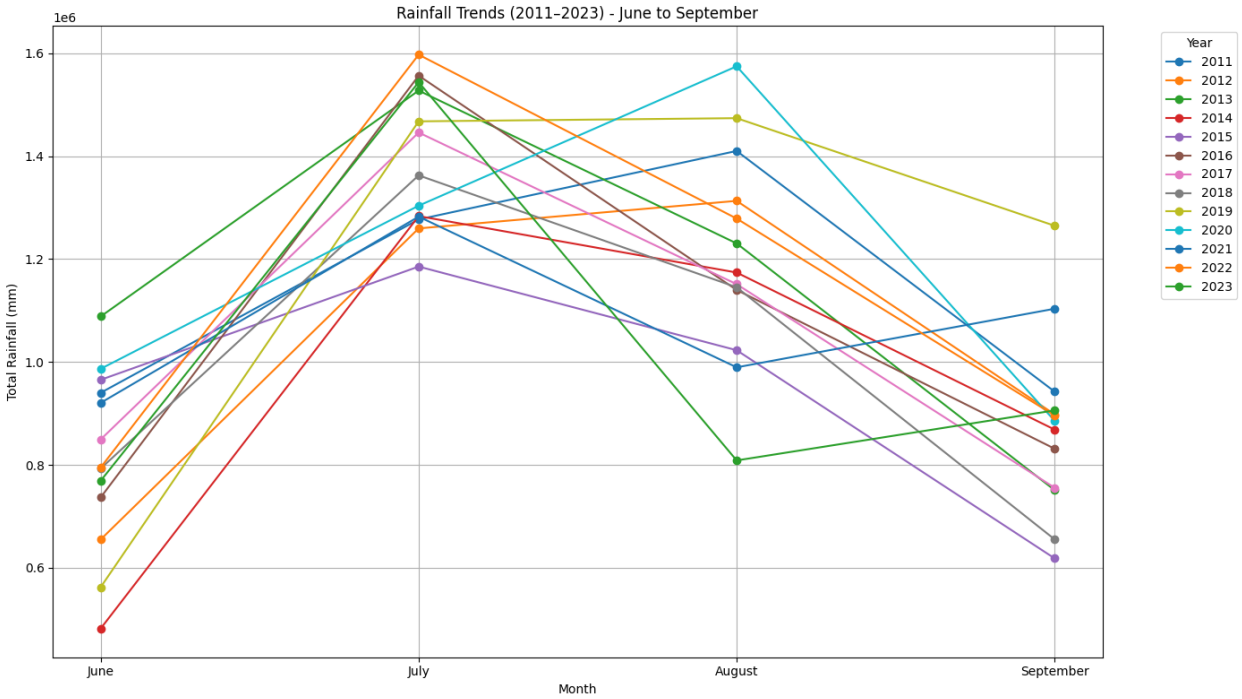
Tools Used

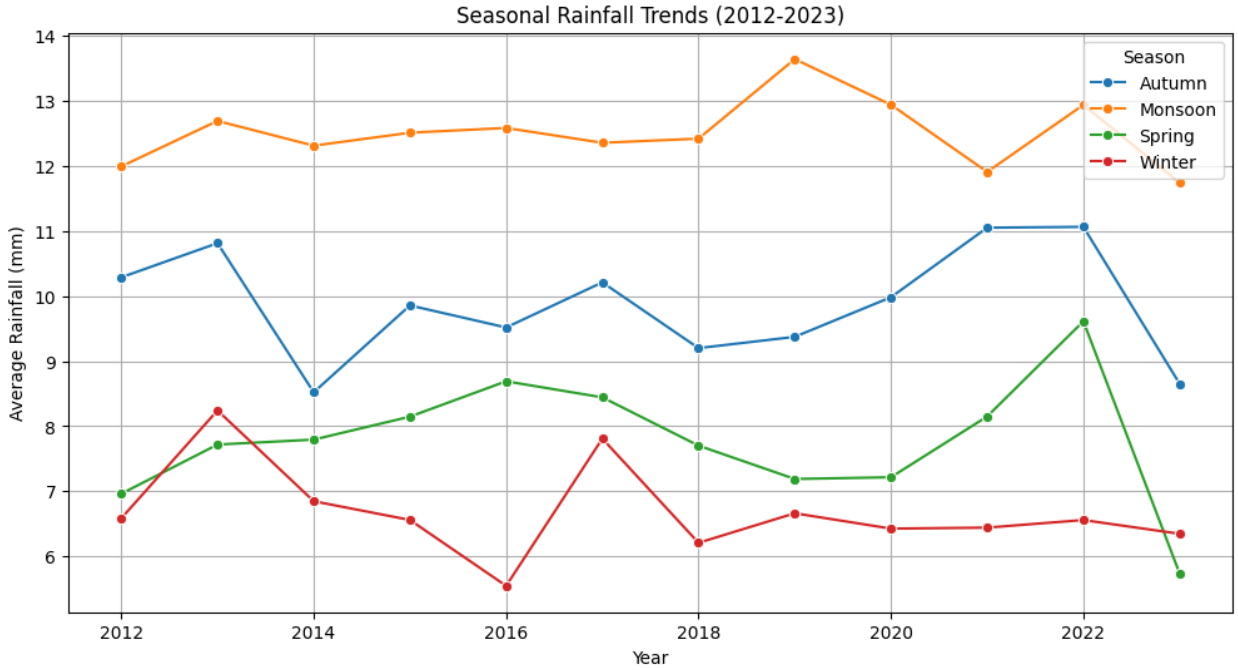
1. **Python Libraries**: xarray, pandas, numpy, matplotlib, seaborn, scikit-learn.
2. **Software**: Jupyter Notebook.
3. **Mapping Tools**: Cartopy for spatial visualization.

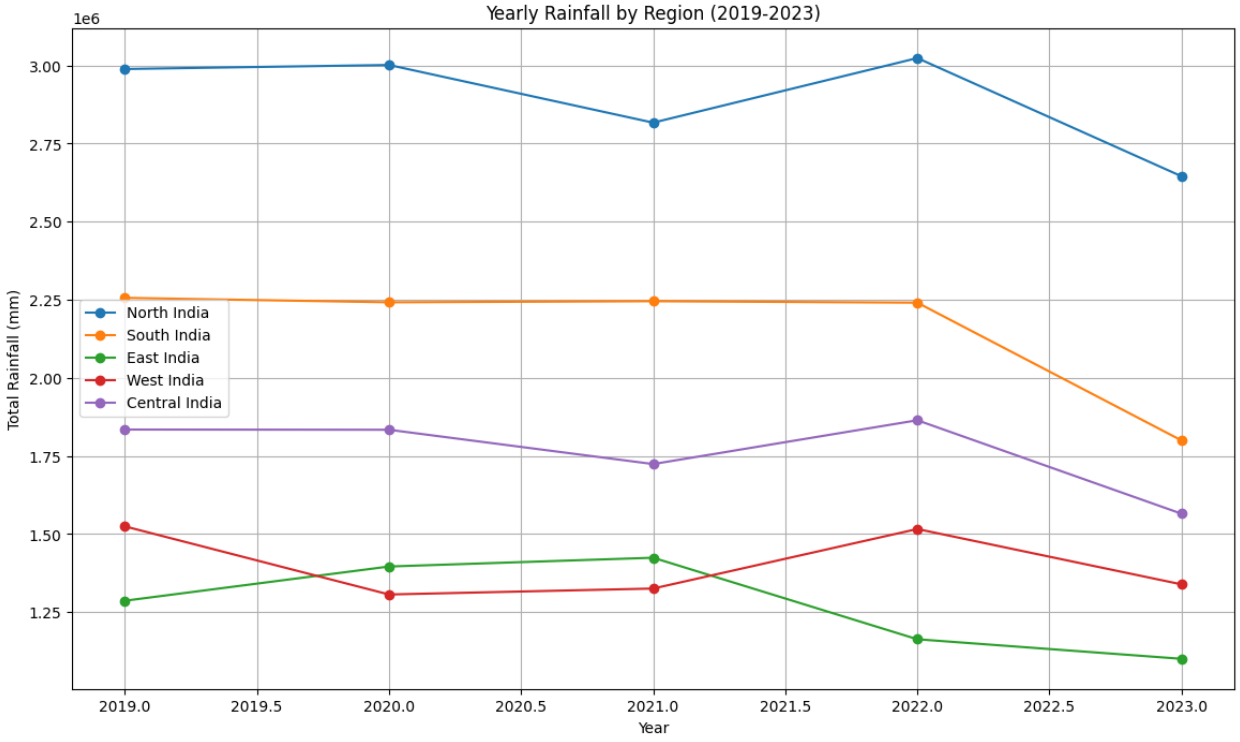
3. Results and Conclusion

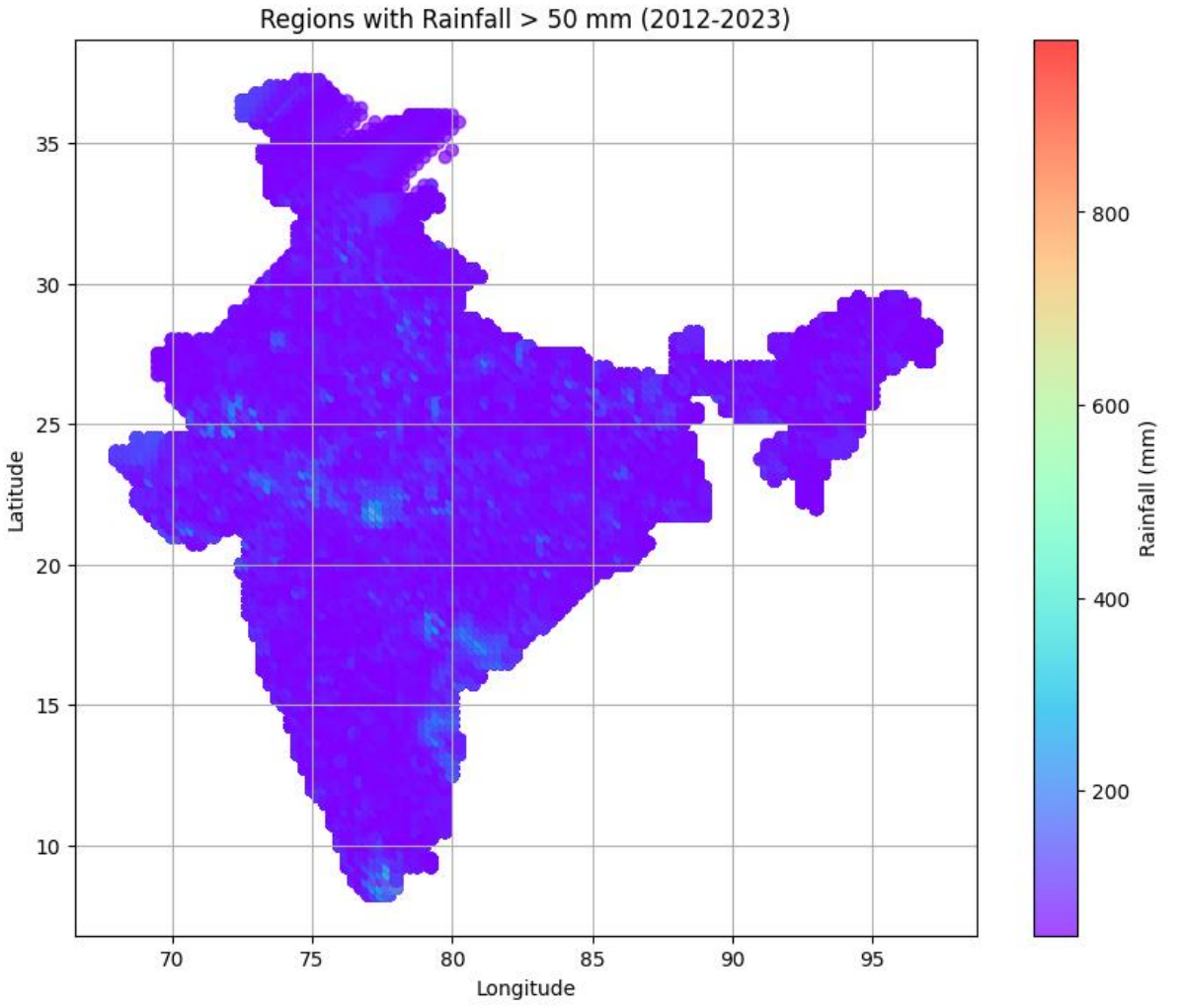
Results

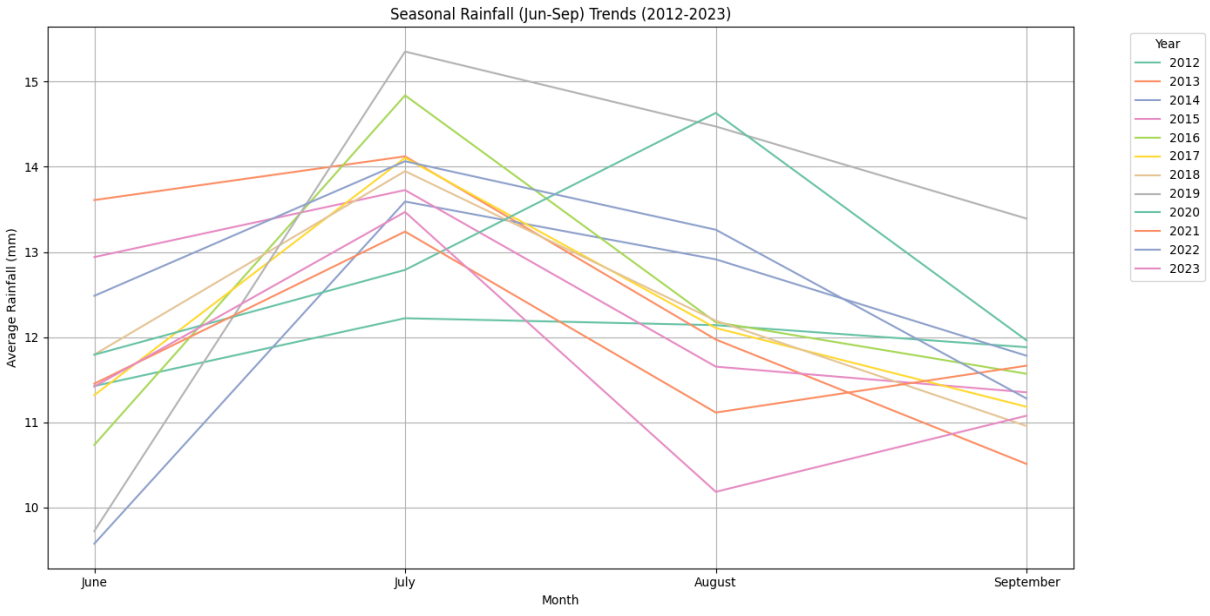
1. **Rainfall Trends**:
    - North and East India exhibit the highest rainfall during the monsoon months (June-September).
    - Central and South India show moderate rainfall with noticeable inter-annual variability.

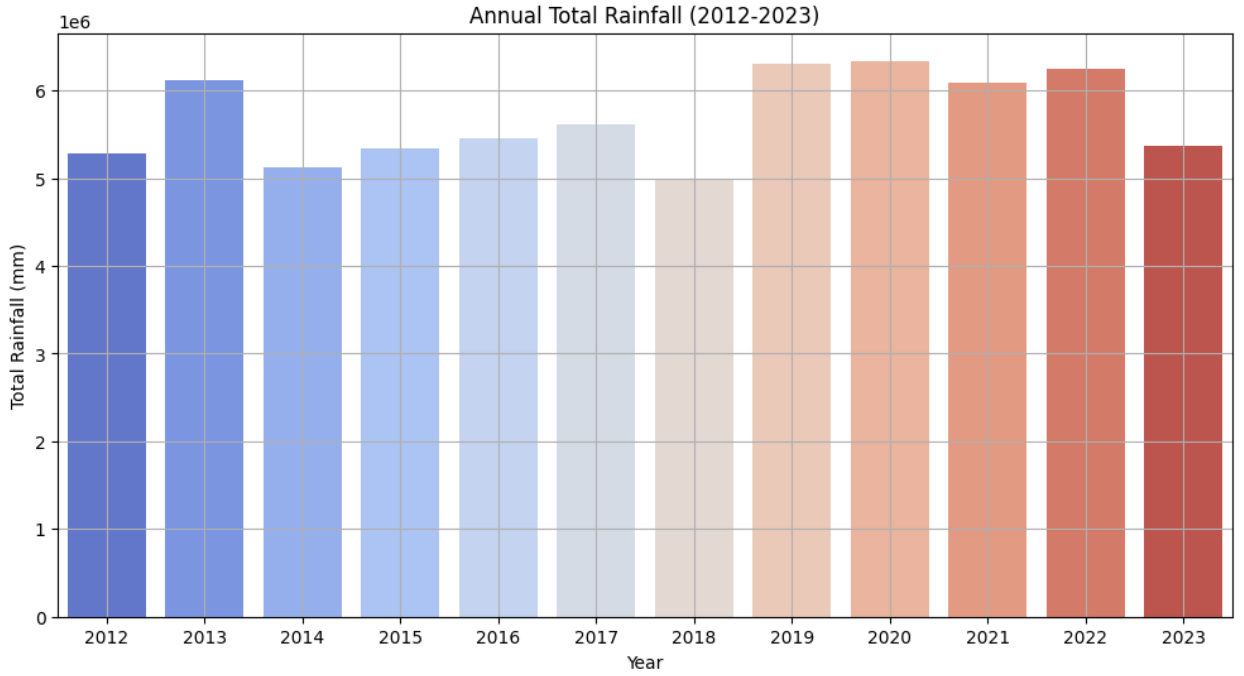


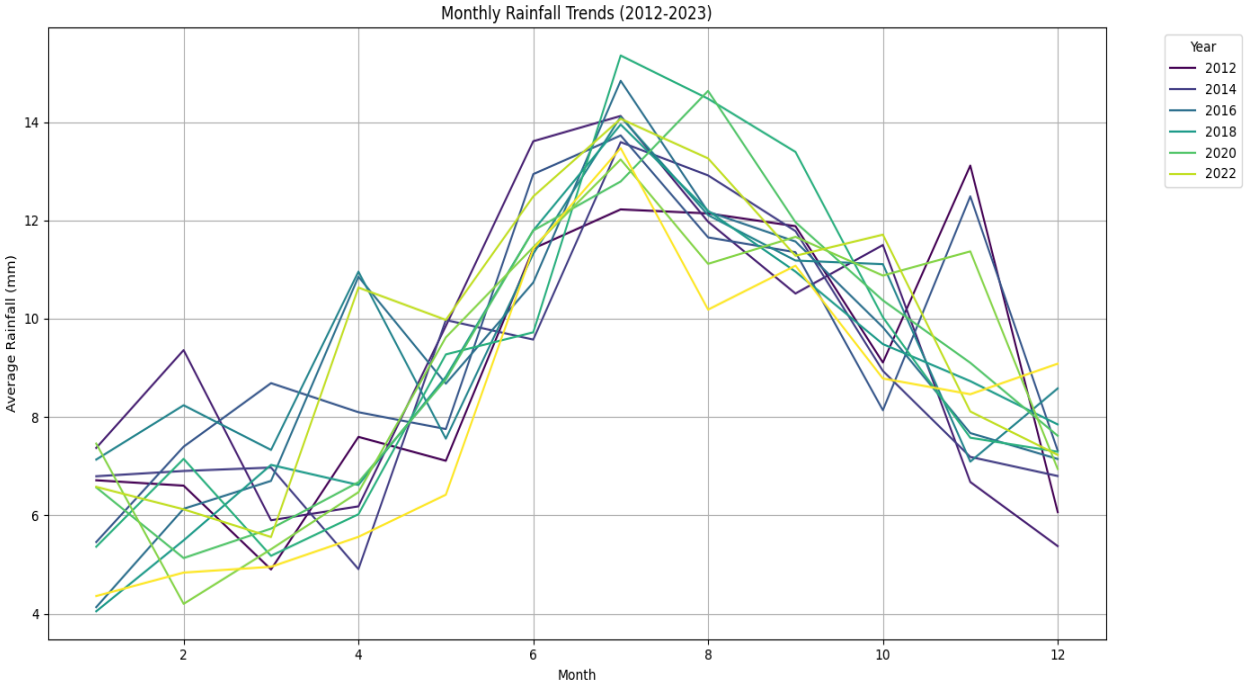

Seasonal Rainfall Trends (2012-2023)
Season
Autumn
Monsoon
Spring
Winter
Average Rainfall (mm)
Year
14
13
12
11
10
9
8
7
6
2012
2014
2016
2018
2020
2022


Yearly Rainfall by Region (2019-2023)
1e6
North India
South India
East India
West India
Central India
Total Rainfall (mm)
Year
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
2019.0
2019.5
2020.0
2020.5
2021.0
2021.5
2022.0
2022.5
2023.0

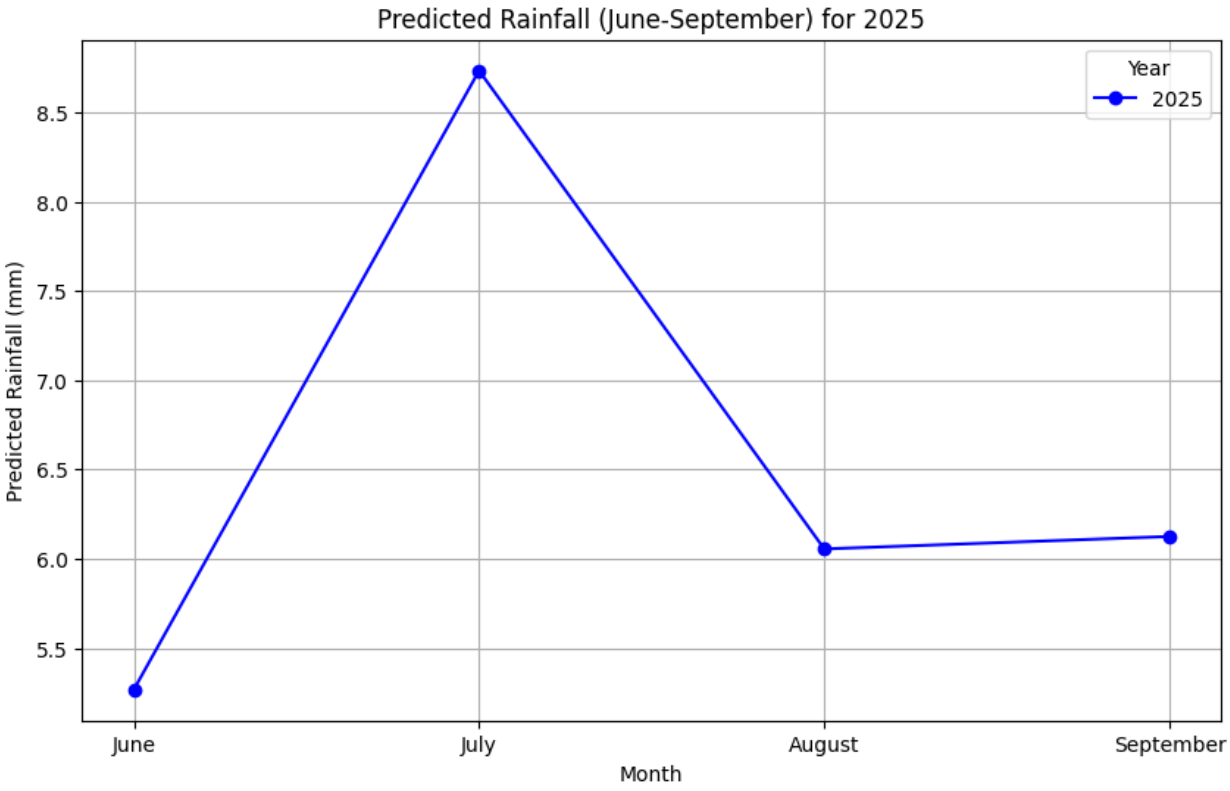

Regions with Rainfall > 50 mm (2012-2023)
Latitude
Longitude
Rainfall (mm)
35
30
25
20
15
10
70
75
80
85
90
95
800
600
400
200


Seasonal Rainfall (Jun-Sep) Trends (2012-2023)
Average Rainfall (mm)
Month
June
July
August
September
15
14
13
12
11
10
Year
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023


Annual Total Rainfall (2012-2023)
1e6
Total Rainfall (mm)
0
1
2
3
4
5
6
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year


Monthly Rainfall Trends (2012-2023)
Average Rainfall (mm)
4
6
8
10
12
14
2
4
6
8
10
12
Month
Year
2012
2014
2016
2018
2020
2022

2. **Predictive Analysis**:
    - Machine learning models struggled due to data quality issues, particularly missing and imbalanced values.
    - Random Forest predictions for 2025 indicated limited reliability due to low correlation between input features and rainfall.



3. **Spatial Insights**:
    - Regions with >50 mm rainfall predominantly align with monsoon activity.
    - Arid regions in West India consistently exhibit minimal rainfall.

Conclusion

The analysis highlights the challenges in modeling rainfall due to data limitations, including missing values and sparse distribution of extreme rainfall events. While AIML techniques provide valuable insights, the dataset's quality and additional climatic variables are crucial for enhancing model accuracy. The study underscores

the need for integrated datasets and advanced methodologies to address these limitations.

4. References:

1. IMD Website: www.imd.gov.in.
2. IMD Pune CHIRPS Data: Rainfall Data in NetCDF Format.
3. CSIR Fourth Paradigm Institute (CSIR-4PI): Studies on climate modeling and data analysis.
4. National Remote Sensing Centre (NRSC): Resources on spatial data and rainfall mapping.
5. World Meteorological Organization (WMO): Guidelines on rainfall prediction and climate models.
6. Climate Hazards Group: CHIRPS Documentation and Tools.
7. Funk, C., et al. "The Climate Hazards InfraRed Precipitation with Stations—A New Environmental Record for Monitoring Extremes." *Scientific Data*, 2015.
8. Guhathakurta, P., & Rajeevan, M. "Trends in the Rainfall Pattern over India." *Current Science*, 2008.
9. Kaggle: Rainfall Prediction Datasets.
10. Python Documentation: Matplotlib, Pandas, Scikit-learn.