**Department of Civil Engineering, IIT Indore**

**Mini Project Report – CE 208**

**Project Topic -** Predicting the possibilities of floods and its management in Assam for next 30 years

**By:** 1) Shubham Kumar - 220004046

2) Srijan Sidharth    - 220004047

3) Valsa Harshith     - 220004049

4) Vishal                  - 220004050

5) Vivek Meena      - 220004051

**Abstract:** Assam, a region prone to frequent and devastating floods, requires robust flood prediction and management strategies to mitigate the impacts on lives, agriculture, and infrastructure. This project aims to address this pressing issue by employing a data-driven approach, specifically utilizing machine learning techniques to forecast rainfall, a key factor in flood prediction. This particular report discusses the current progress made in addressing the issue with the help of a very short-term rainfall prediction model (for few hours) developed using meteorological data and time series forecasting which achieved an accuracy of 86.30% using a model built upon Random Forest Regression algorithm.

**Keywords:** Assam, data**-**driven approach, rainfall prediction, very short-term rainfall prediction, time-series forecasting.

**1. Introduction**

One of the most flood vulnerable regions in the world is Southern Asia, where the frequency of extreme floods is on the rise causing enormous damages to lives, crops and infrastructure (Mirza, 2011). Northeastern India, along with the Brahmaputra catchment area, contains the most flood-vulnerable zones of the country (López et al., 2020). The state of Assam spans across both the Barak and Brahmaputra valleys. Being situated in a tropical climate zone, it undergoes significant fluctuations in temperature and humidity throughout the year, marked by clearly defined wet and dry seasons. It receives ample annual rainfall, following a distinct seasonal pattern. However, due to the abundance of annual precipitation, it experiences natural disaster of flood very frequently that causes huge damage to the biodiversity. During post-independence period, Assam faced major floods in 1954, 1962, 1972, 1977, 1984, 1988, 1998, 2002, 2004 and 2012. Almost every year three to four waves of flood ravage the flood prone areas of Assam (Government of Assam Water Resources).

Efficient management and readiness for flood risks are essential in minimizing the devastating impact of floods, securing the safety of at-risk communities, and preserving the stability of ecosystems and economies in flood-prone areas. An essential part of a disaster risk management is its prediction, to provide accurate and timely information to the decision-makers, allowing them to implement appropriate managerial measures to mitigate the impact of flooding and safeguard lives and property. Flood forecasting involves using scientific methods to predict when and where floods might occur in a specific area. This is often done by analysing a range of factors and predicting the possible patterns between each of them, and their interdependence on each other. Flood forecasting may aim to predict the occurrence of flood based on the timing, duration and magnitude or even the spatial extent by integrating hydrological information. Modern approaches to flood prediction involves methodologies such as, Physical-based, Data-Driven and Hybrid Modelling (SOLOMATINE & PRICE, 2004). Forecasting rainfall is a challenging task that demands precision. Various hardware devices exist to predict rainfall using weather factors like temperature, humidity, and pressure. However, traditional approaches are inefficient.

This study implements a data-driven methodology to investigate flood prediction by forecasting rainfall, by employing machine learning techniques.

Machine Learning (ML) technique is a type of artificial intelligence where accurate predictions can be made with the help of historical records and data. Various machine learning models can enhance flood predicting capabilities, such as, random forests (Breiman, 2001), extreme gradient boosting, decision-tree, artificial neural network, support vector machine, logistic regression, linear regression, etc. This study mainly used random forest to forecast rainfall. (Curth et al., 2024), explains how random forests achieve their excellent performance in practice by interpreting trees and ensembles as adaptive smoothers.

The primary objective of this project is to provide precise flood forecasts for the region of Assam. This study discusses about the current understandings of flood dynamics, outlines the advancements made thus far and describes the prospective actions required to further enhance flood forecasting accuracy and effectiveness.

**2. Literature Reviewed**

This section includes a brief summary about the major reviewed recent studies on rainfall forecasting, ­­flood modelling, and prediction.

(1) (Singha et al., 2022) has discussed about the flood susceptibility mapping in a region of Assam in quite detail using 22 flood causative factors, including morphometric factors, hydrological factors, soil permeability factors, terrain distribution factors, and anthropogenic factors. It included implementation of various machine learning models and then stacking all to develop a hybrid model with a very good accuracy and precision for flood hazard mapping. The approach used for modelling was by classification.

(2) (Ramayanti et al., 2022) discusses about the flood susceptibility mapping in a Beira area, Mozambique. The study utilized SAR data, GIS data, and two deep learning models: the group method of data handling (GMDH) and convolutional neural network (CNN). These models were integrated to develop an accurate Flood Susceptibility Map. The approach for modelling was by regression analysis.

(3) (Salma et al., 2022) proposed a Hybrid machine learning model for prediction of rainfall. The study used CNN algorithm since it has less chance of overfitting as it considers a smaller number of parameters in the network. The raw dataset on per-day basis for the previous 10 years was used for accurate prediction.

(4) (Fayaz et al., 2022) utilized Model Trees for rainfall prediction. The study discusses in detail how are model trees implemented in such problems and examines the ‘M5’ Model Tree algorithm's capabilities for analysing rainfall data in Kashmir. The M5 model tree is a machine learning algorithm that traditional decision trees by allowing linear regression models to be applied at the leaves of the tree, enabling more accurate predictions. This hybrid approach makes it particularly effective for datasets with both numerical and categorical features and it is one of the very effective yet not too complex popular model used widely.

(5) (Myrchiang et al., 2023) implemented 5 machine learning models using Assam’s historical rainfall and geospatial data. The study used 5 days rainfall data integrating with the geospatial data for accurate prediction for different locations in Assam.

(6) (Khan et al., 2020) proposed a hybrid approach using a combination of Convolutional Neural Network and Multi-Layer Perceptron, for 1-day to 5-day in advance daily rainfall prediction. Nine meteorological variables, closely associated with daily rainfall variation, were used as inputs to the hybrid mode, namely, maximum air temperature, geopotential height, longwave radiation, maximum relative humidity, minimum relative humidity, u-wind speed, v-wind speed, sea level pressure and rainfall.

(7) (Latif et al., 2023) summarized some 70 recent research papers on prediction of rainfall using different techniques and approaches. Several articles reviewed were taken directly from this study. The study also analyses several techniques utilised in the research papers summarized and presented several terms of analysis for development of an optimum or a good rainfall forecasting model.

All the various rainfall forecasting studies reviewed either used meteorological datasets (Fredyan et al., 2022) or hydrological data (Frame et al., 2022; Gauch et al., 2021) or combined both for modelling (Rao et al., 2022) for rainfall prediction. This study utilises only meteorological data for rainfall prediction and further aims in future to develop flood susceptibility mapping integrating with geospatial data.

**3. Materials and Methods**

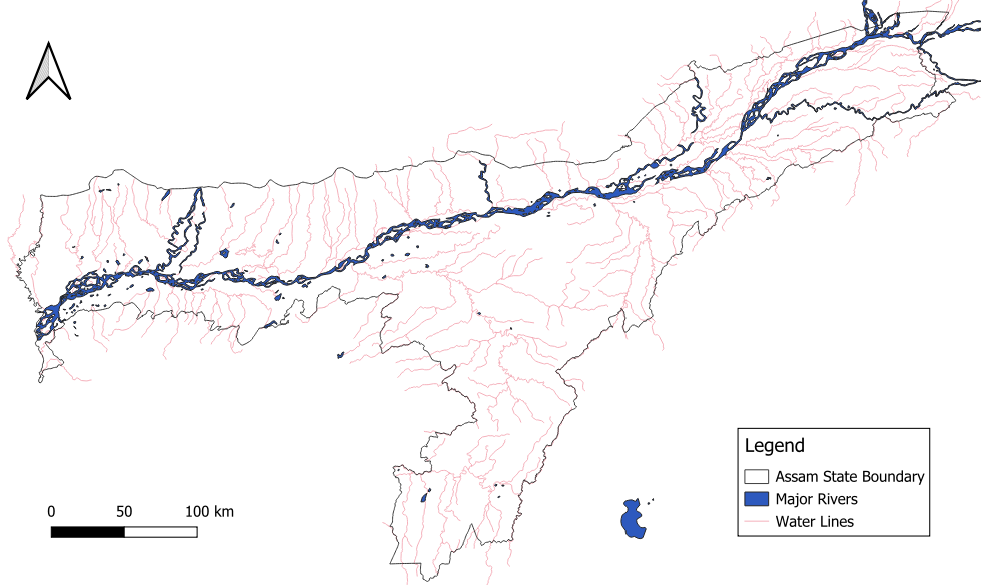
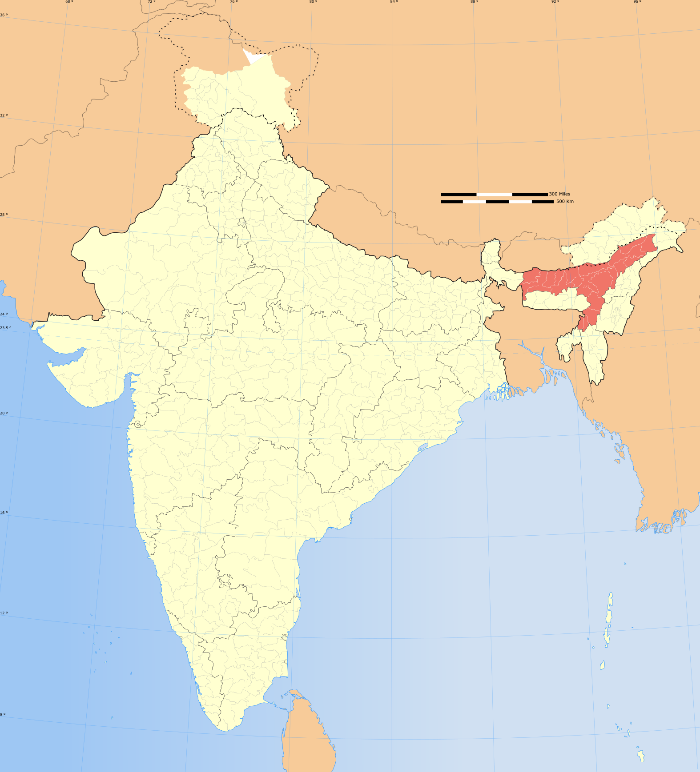
This investigation began with exploration, review and analysis of literature regarding articles about prediction of flood and forecasting rainfall specifically in the region of Assam. The exploration was carried out using ‘Connected Papers’. It builds a graph for several similar research papers to what is entered, enabling easier search for articles. Articles were selected based on certain sets of keywords such as:

{“flood”, “rainfall”}; {“machine learning”, “forecasting”, “prediction”, “modelling”}; {“Assam”, “India”}.

The search and selection of articles was done manually considering combinations of words from the sets. There was not much of region specificity, so a few similar articles from other countries were also selected; although articles on studies in India were of focus. The majority of the articles were selected from within the past 10 years to ensure that the study remains relevant to the latest findings regarding this topic.

3.1 Study Area

The state of Assam is located 89° 42′ E to 96° E longitude and 24° 08′ N to 28° 02′ N latitude and covers an area of 78,438 km2. Assam experiences tropical monsoon climate with high humidity and heavy rainfall. It covers an altitude range of 45 m – 1960 m. The average humidity is 90%. The months of November, December, January and February constitute the winter season. The winter temperature varies between 10 °C to 12 °C. In April and May, depressions over the Bay of Bengal enter the region and strong winds, locally known as “Bordoichilla” visit the area with occasional rainfall. The temperature goes to 20 °C – 28 °C. Most of the rainfall in the State received by the influence of south –west monsoon between June to September. The average annual rainfall is 166 cm in Brahmaputra valley, and 183 cm in Barak valley. The summer temperature varies between 30 °C to 36 °C. Nearly 75% of precipitation occurs during the monsoon period (July to September) each year.



**Figure 1.** Study area map

3.2 Data Acquisition

For the implementation phase in the project, three different types of data were collected for different purposes, they are namely, Remote sensing data/Satellite image data for Flood inventory mapping, Meteorological data of various locations in Assam for model development, training and testing; and Geospatial data for study area analysis.

i) Remote sensing data was collected in the form of Annual flood layer of Assam developed by Bhuvan National Remote Sensing Centre (NRSC) – Indian Geo Platform of NRSC launched by ISRO, from the flood annual layer of Web Map Service (WMS) of Thematic Services for 12 years from 1999 to 2010.

ii) Raw time series meteorological datasets for the point locations obtained after inventory mapping were obtained from NASA Lagney Research Centre (LaRC) – POWER Project at hourly and daily temporal levels for 9 years, 2 months (01-01-2015 to 01-03-2024) and 30 years (01-03-1994 to 01-03-2024) respectively. All the available parameters were taken from the API are tabulated below:

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Parameters Category | Number | Parameter Names |
| 1. | Solar Fluxes and Related | 9 | All Sky Surface Shortwave Downward Irradiance, Clear Sky Surface Shortwave Downward Irradiance, All Sky Insolation Clearness Index, All Sky Surface Longwave Downward Irradiance (thermal infrared), All Sky Surface Photosynthetically Active Radiation (PAR) Total, Clear Sky Surface Photosynthetically Active Radiation (PAR) Total, All Sky Surface UVA Irradiance, All Sky Surface UVB Irradiance, All Sky Surface UV Index |
| 2. | Parameters for solar cooking | 3 | All Sky Surface Shortwave Downward Irradiance, Clear Sky Surface Shortwave Downward Irradiance, Wind Speed at 2 Meters |
| 3. | Temperatures | 7 | Temperature at 2 Meters, Dew/Frost Point at 2 Meters, Wet Bulb Temperature at 2 Meters, Earth Skin Temperature, Temperature at 2 Meters Range, Temperature at 2 Meters Maximum, Temperature at 2 Meters Minimum |
| 4. | Humidity/Precipitation | 3 | Specific Humidity at 2 Meters, Relative Humidity at 2 Meters, Precipitation |
| 5. | Wind/Pressure | 11 | Surface Pressure, Wind Speed at 10 Meters, Wind Speed at 10 Meters Maximum, Wind Speed at 10 Meters Minimum, Wind Speed at 10 Meters Range, Wind Direction at 10 Meters, Wind Speed at 50 Meters, Wind Speed at 50 Meters Maximum, Wind Speed at 50 Meters Minimum, Wind Speed at 50 Meters Range, Wind Direction at 50 Meters |

iii) Geospatial for entire assam state was collected for the spatial map preparation and analysis. This data is crucial for spatial modelling such as flood susceptibility mapping. The geospatial data is not yet integrated with the present model. However, this data is also important for understanding the geographical aspects such as elevation, slope, aspect, profile curvature, topographic ruggedness index (TRI), etc.

3.3 Platforms Used

For preparing inventory mapping and spatial maps, QGIS (v. 3.36.1) was used. For data preprocessing and model selection, Google Colab platform (mainly due to the easy manageability of Exploratory Data Analysis), with a T4 GPU hardware accelerator was used. An AMD Ryzen 5 5625U with Radeon Graphics was used to perform hyperparameter tuning in the Visual Studio Code platform. Python libraries namely, os, requests (for automated fetching of dataset from API for several locations), pandas for data handling, math, numpy, matplotlib, scipy, sklearn, and seaborn were used for this study.

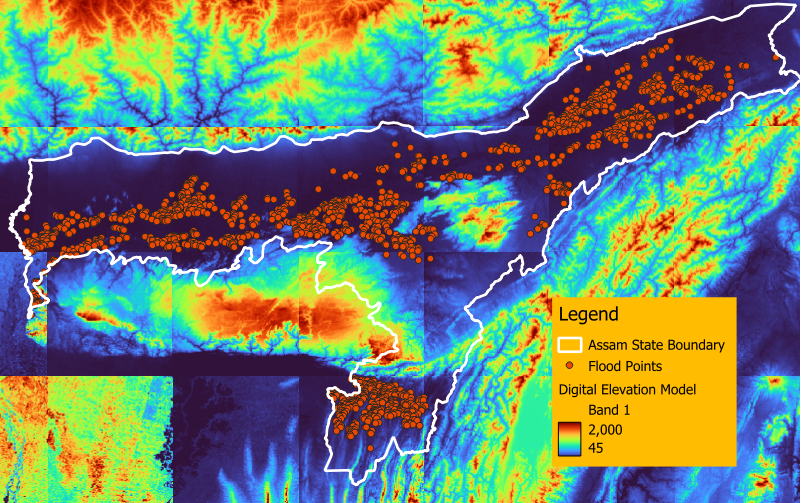
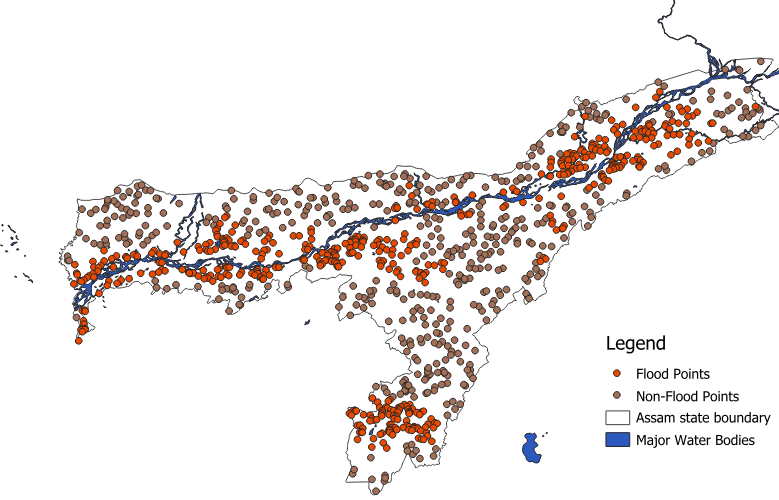
3.4 Methods

In the context of flood prediction, employing rainfall forecasting as regression utilizing past rainfall data to predict future precipitation levels. By treating rainfall as a continuous variable, regression models can estimate the expected rainfall over a given period. This approach offers a quantitative understanding of how precipitation patterns relate to flood occurrences, supporting informed decision-making in flood-prone areas.

3.4.1 Flood Inventory Mapping

An accurate spatial flood prediction is dependent upon on the different flood causing factors suitable for the study area, such, topography, land use, hydrological characteristics, climate, etc. An inventory map, which shows the areas most susceptible to flooding, is essential for predicting future flood occurrences in the study area. Flood inventory mapping involves systematically charting existing floods within a region, employing methodologies such as field surveys, interpretation of aerial photographs or satellite image analysis, comprehensive review of historical flood records, technical reports, governmental publications, and expert interviews (Wubalem et al., 2021). The flood-prone area of the entire country constitutes approximately 10.2% of the total land area, whereas in Assam, it accounts for 39.58% of the state's area. This indicates that the flood-prone area in Assam is nearly four times larger than the national average for flood-prone areas in the country.

In order to accurately access the risk and impact of flooding, this study analysed the historical flood data. Annual flood layer data of Assam developed by Bhuvan National Remote Sensing Centre (NRSC) – Indian Geo Platform of NRSC launched by ISRO, was collected from the flood annual layer of Web Map Service (WMS) of Thematic Services from 1999 to 2010 to develop flood inventory mapping in Assam. The satellite images of inundated areas in Assam thus obtained, were processed in the QGIS platform to provide the points of flood occurrence during each year. Using the Cell Statistics tool for Raster analysis in QGIS, points of maximum occurrences in each of the maps were obtained and identified as flood point locations. Thus, there were 1595 flood points obtained. The coordinates for each of these points were extracted for future use. For the future flood modelling, we chose 500 points of these 1595 points randomly; and chose more 500 points randomly other than these flood points in the study area identified as non-flood points. For the future flood model analysis, the database can be assigned a binary composition as: the flood points being represented as ‘1’ and non-flood points represented as ‘0’. A flood inventory map is essential for performing a spatial analysis, and is generally used for mapping the flood susceptibility of the area (Addis, 2023; Ramayanti et al., 2022; Singha et al., 2022).



**Figure 2.** Flood inventory map

3.4.2 Flood Affecting Factors

Predicting rainfall is a complex task influenced by a range of meteorological factors that exhibit dynamic variability across different locations. Flood characteristics vary depending on the geo-environmental conditions of each area. The selection of features for rainfall modelling also depends on the availability of relevant data. Observational meteorological data pertain to atmospheric conditions, for instance temperature, dew point, wind direction, wind speed, cloud cover, cloud layer(s), ceiling height, visibility, current weather, and precipitation amount (Latif et al., 2023). It is crucial to extensively utilise Meteorological features to forecast variations in precipitation. A few papers used temperature and humidity as their input data, whereas a few others use radar echo and CHIRPS spatiotemporal data respectively as their meteorological independent variables.

To perform flood susceptibility mapping effectively, it's crucial to identify the factors that condition floods. Model accuracy also depend upon correctness of the data that is used to train the model. After analysis of various literatures and collecting various weather data sources, one of the API was chosen for data collection and further analysis and all the conclusions made are mentioned at every possible step.

3.4.3 Machine Learning Algorithms Used in The Current Model

Random Forest

Random Forest (RF) is a machine learning model introduced by (Breiman, 2001), characterized by its ensemble of decision trees. The primary objective of ensemble learning is to combine the training of multiple decision trees into a unified prediction. Random Forest allows for the estimation of decision tree classifier accuracy across various subsample points through bootstrapping (). By use of simple averaging, Random Forest enhances prediction accuracy and regulates overfitting.

Gradient Boosting

The gradient boosting algorithm is a very efficient ensemble technique utilized for both regression and classification tasks involving tabular data. Gradient Boosting operates as an additive framework, allowing the sequential execution of differentiable loss functions to minimize error gradients. At each stage, a regression tree is fitted to the negative gradient of the calculated loss function. At times, as per the problem often Gradient Boosting is preferred over Random Forest.

Decision Tree

A decision tree is a simple versatile and intuitive machine learning algorithm used for both classification and regression tasks. They operate by recursively splitting the dataset based on the values of input features, creating a hierarchical structure resembling a tree. It takes a tree structure with nodes, branches and leaves. However, decision trees are prone to overfitting, especially when the trees are deep and complex.

Linear Regression

Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input variables and the output, represented by a straight line in two dimensions or a hyperplane in higher dimensions. The goal of linear regression is to find the best-fitting line that minimizes the difference between the observed values and the values predicted by the model. This is typically achieved by minimizing the sum of squared differences between the observed and predicted values, a method known as Method of Least squares.

Multi-Layer-Perceptron

A multilayer perceptron (MLP) is a type of artificial neural network characterized by its layered architecture, consisting of an input layer, one or more hidden layers, and an output layer. Each layer contains nodes, also known as neurons, which are interconnected with weighted connections. The MLP is a powerful tool for supervised learning tasks such as classification and regression. It can learn a non-linear function approximator for either classification or regression problem. MLPs are versatile and can learn complex patterns in data, making them suitable for a wide range of applications in areas such as image recognition, natural language processing, and financial forecasting. However, they require careful tuning of parameters and may be prone to overfitting with large, high-dimensional datasets.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple yet powerful supervised learning algorithm used for both classification and regression tasks. The principle behind KNN is based on the assumption that data points with similar features tend to belong to the same class or have similar target values. In KNN, the prediction for a new data point is determined by the class or average of the values of its nearest neighbors in the feature space.

Support Vector Machine

Support Vector Machines (SVMs) are powerful supervised learning models. The core idea behind SVMs is to find the optimal hyperplane that best separates different classes in the input space while maximizing the margin between the classes. SVMs are effective in high-dimensional spaces and are versatile due to their ability to handle linear and non-linear data using different kernel functions. However, SVMs may not predict well if the data is not linearly separable or if the appropriate kernel function and parameters are not chosen correctly and struggles with large, high-dimensional and noisy datasets, which can lead to overfitting or poor generalization performance.

3.4.4 Design of the Machine Learning Model

The working and architecture of the current model is described in detail. The meteorological data was collected for the past 10 years was fed to the model developed. The dataset was split in into training and testing set and then pre-processed first, performing the Exploratory Data Analysis and Feature Engineering, and then feature selection was performed for the dataset. After the feature selection, the model selection was performed for finding the most accurate working models for the dataset. Finally, the selected model was tuned for its optimum hyper-parameters and then validated.

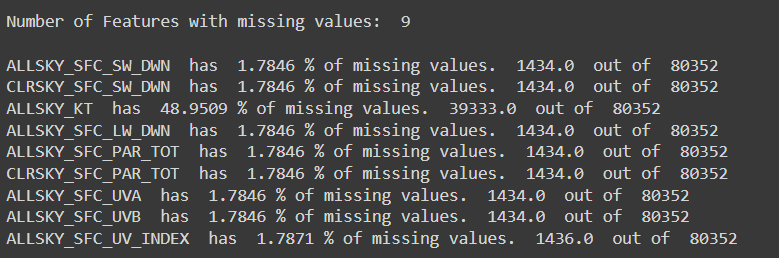
Various authors have employed different strategies to predict rainfall at various locations. Rainfall prediction and the factors considered for prediction may vary depending on the location (Barrera-Animas et al., 2022). The study discussed here was performed on the dataset for a particular location in Assam. All the steps taken along with the conclusions, if any, along with the improvements possible are covered in quite detail hereafter.

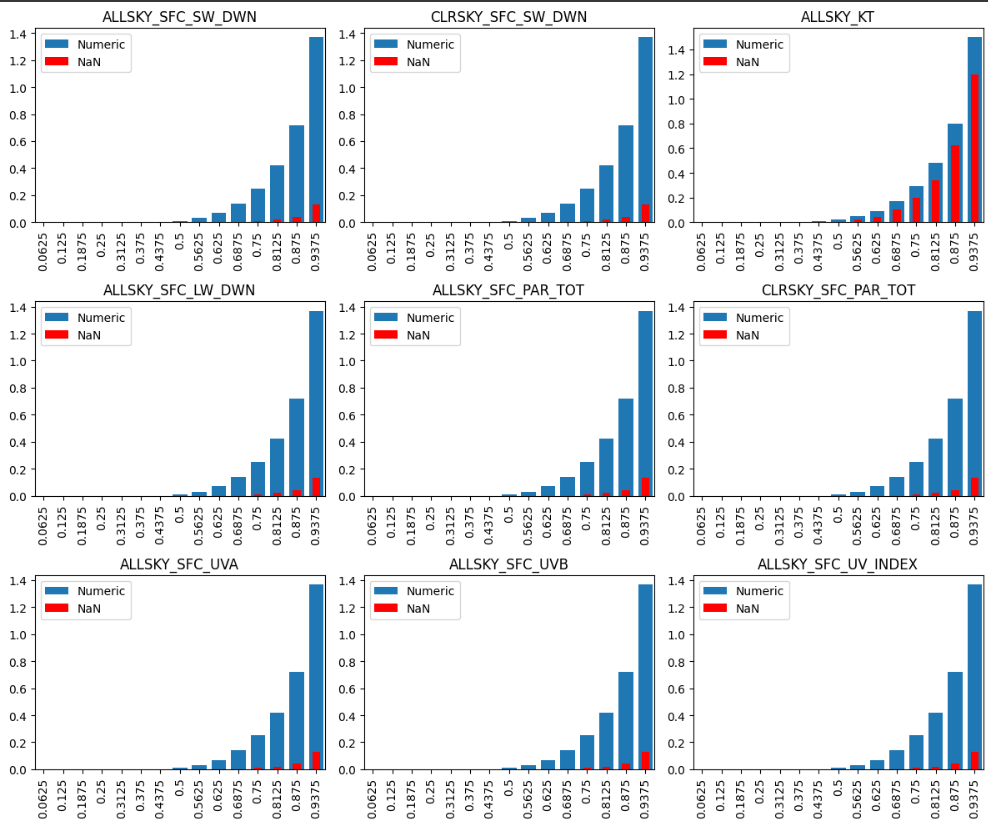
**3.4.4 Data Preprocessing**

The dataset contains records for 9 years and 2 months and 80,352 records. The dataset was first split into 75% and 25% and, then pre-processing was done for both the splits independently. This is crucial step to avoid any Data Leakage (Apicella et al., 2024). The parameter ‘PRECTETCORR’ (Corrected Precipitation) was assigned as the target variable to be forecasted and rest all 25 other parameters were declared as features/predictor variables. The dataset used does not contain any missing values in the target variable.

**Exploratory Data Analysis**

*Missing Values*

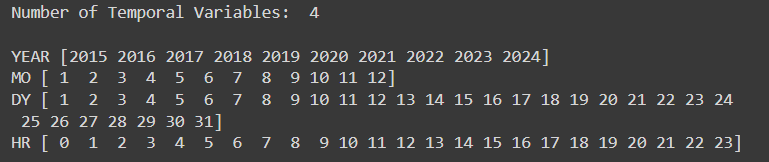
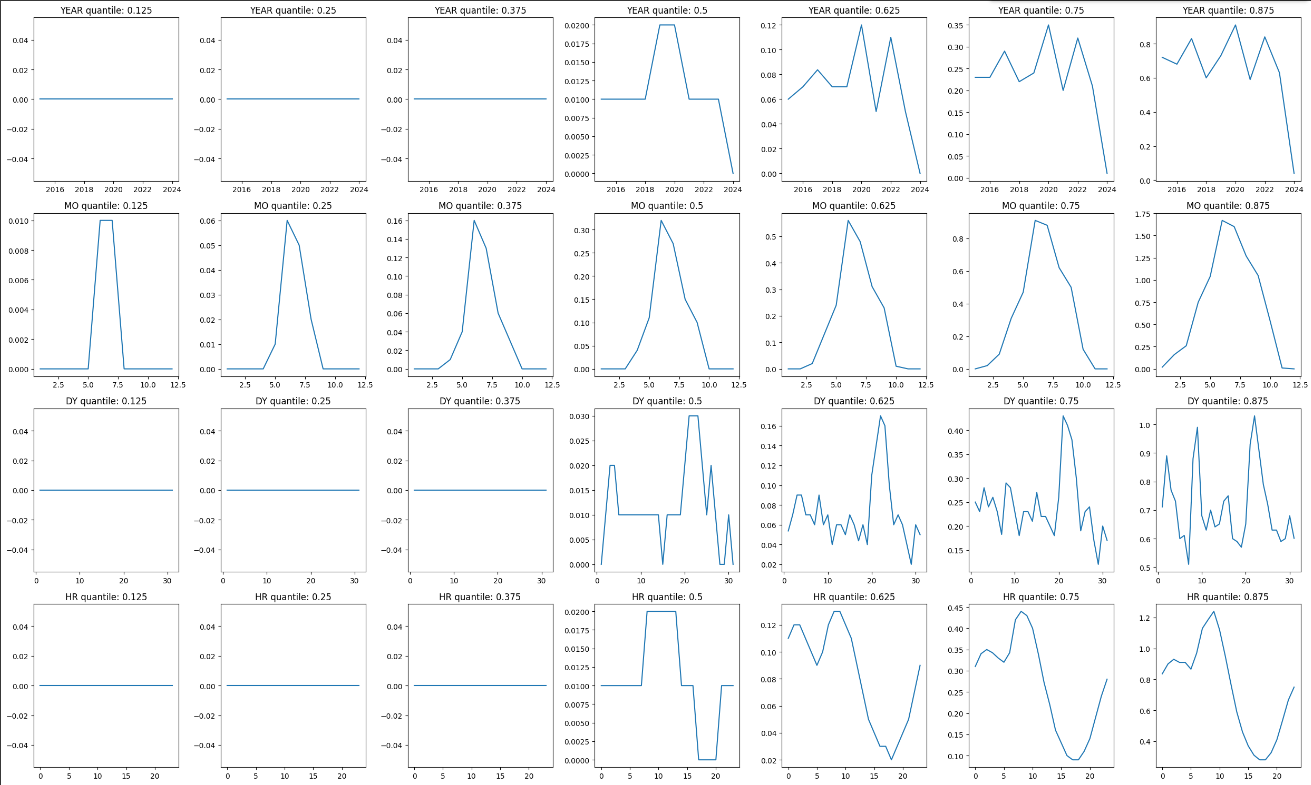
The missing values preset in some of features in the dataset were analysed first, comparing the missing values at different percentiles with respect to the target variable (Corrected Precipitation) for both of missing and non-missing values.



Conclusions:

Thus, if we see some trend between Target Values, we may not simply ignore the NaN values. E.g., for "ALLSKY\_KT", NaN values and non-NaN follow similar increasing trend with target variable after around 50% percentile. We cannot drop the column rightaway. We need to take care of this while doing feature engineering. Additionally, we may or may not drop other columns' NaN values, because of less dependence of the Target variable on their NaN values.

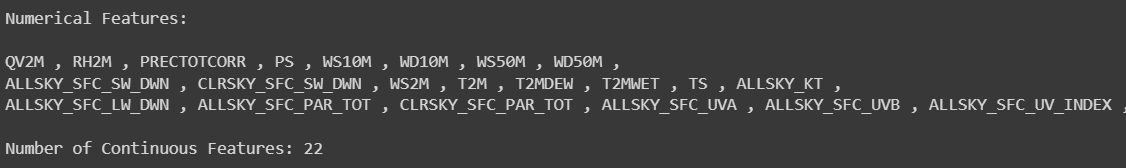
*Temporal Variables*

There are 4 temporal variables in the dataset, namely, Year (YEAR), Month (MO), Date (DY), and Hour (HR). The dependence of the Target variable was analysed with the temporal variables.

Conclusions:

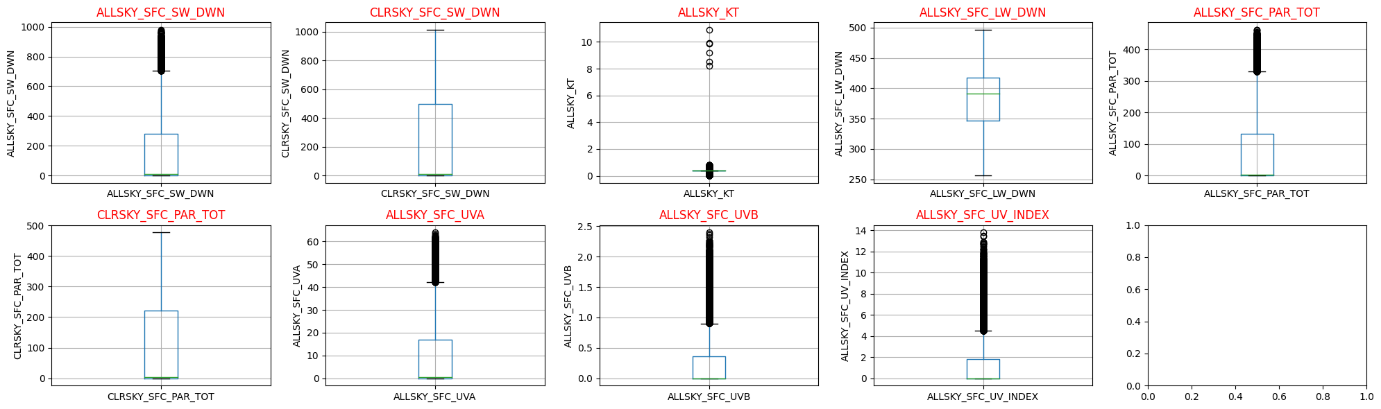
* There is a good trend between month and rainfall as expected, i.e., rainfall in Assam is observed to be maximum during monsoon, around June
* Similar observations for hour with rainfall for this dataset. Maximum rainfall occurs around morning to afternoon
* There is no such general trend between year and rainfall. Still, we may get some information for improvement. We'll drop it for now.
* It seems that month follows some unusual and unexpected trend. For heavy rainfall, it has more randomness; for moderate rainfall, there is a considerable gap between consecutive comparable precipitation.

*Numerical Variables*

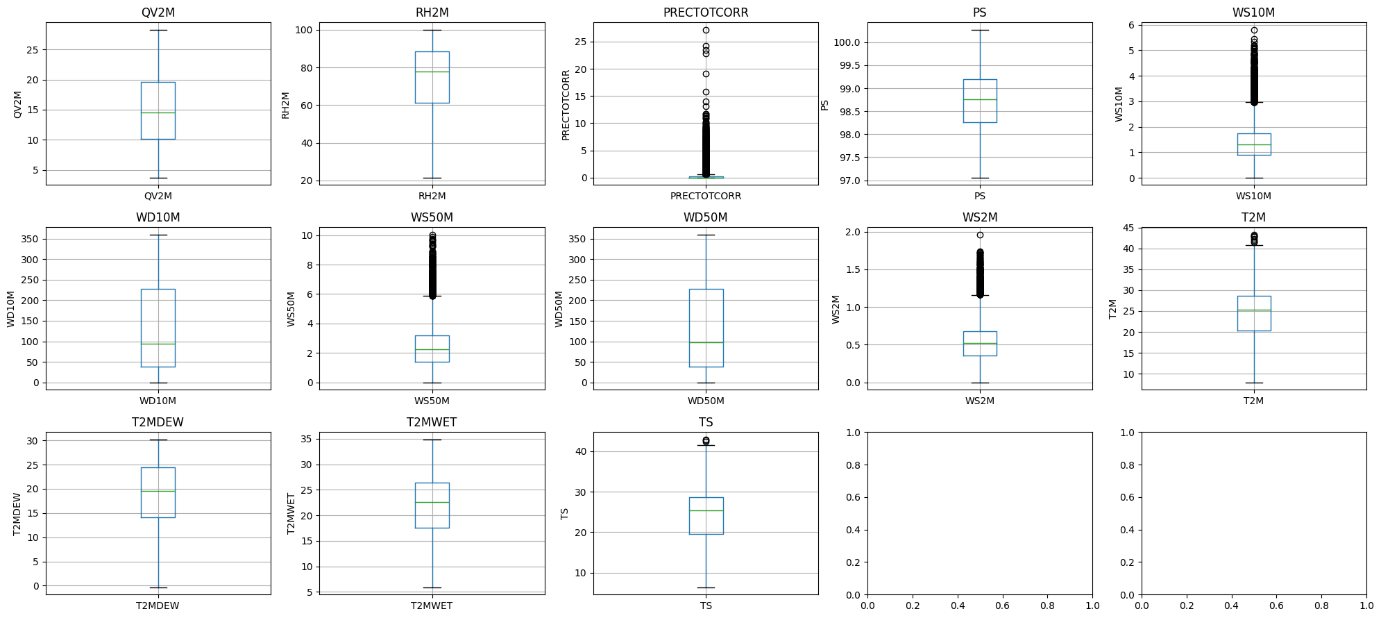
There are 22 numerical features in the dataset. These features were analysed for outliers to help in data cleaning by imputation and to understand the data to avoid underfitting.

During Exploratory Data Analysis, visualizing the distribution of numerical features is essential for gaining information about the underlying patterns and characteristics of the data to understand necessary measures to be taken. Histograms, box plots, and density plots are commonly used techniques to visualize distributions.

*Visualising Outliers:*

Outliers are data points that significantly differ from the rest of the dataset. They can skew statistical analyses and machine learning models by introducing noise or bias. Detecting and handling outliers is crucial for maintaining the integrity and accuracy of analyses and predictions. However, for such a problem statement where prediction of such a variable has to be performed, which can vary a lot at any point of time, and thus removal of outliers is not a good option, we need to check how the outliers’ removal affects the prediction and may need to handle them with more sophisticated methods for a good prediction performance.

(Visualising outliers for features with missing values)



(Visualising outliers for other features)

Conclusions:

* There are outliers in the features, thus, we'll need to handle them carefully. We cannot remove outliers directly since, as per the problem statement we may need to predict values around outliers.
* Also, the prediction was performed after further preprocessing, to analyse the prediction in both instances of presence of outliers and after their removal.
* For the missing values present in the features, we may replace them with the median or mode.

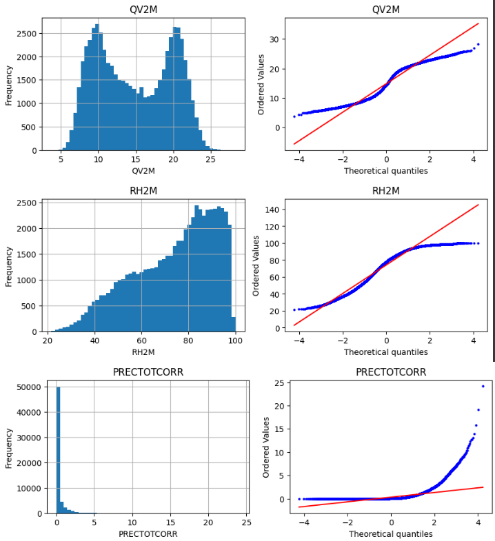
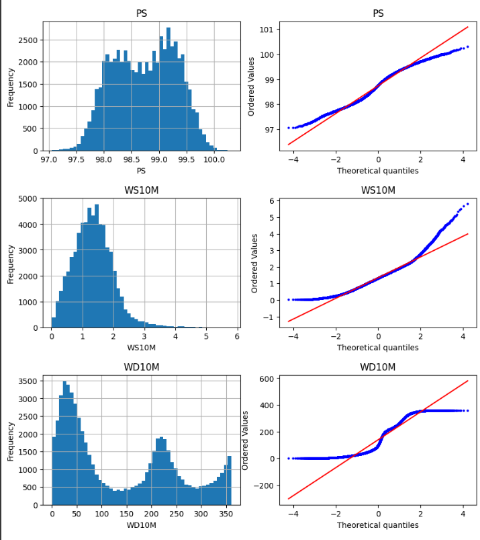
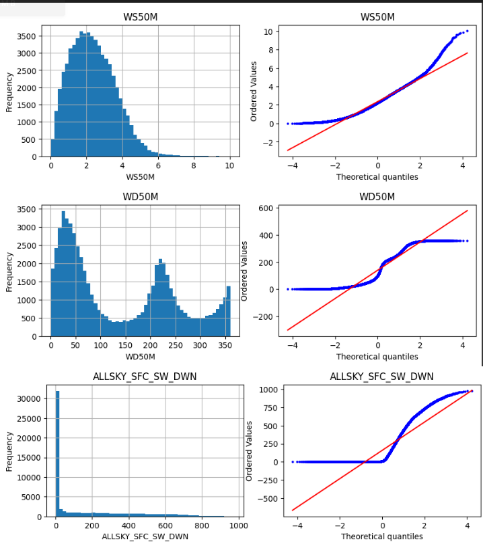
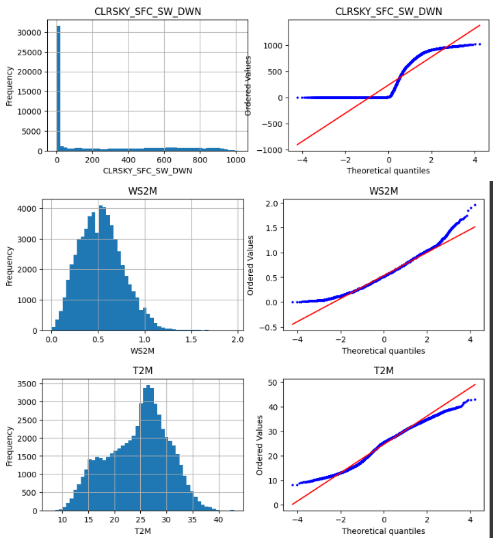
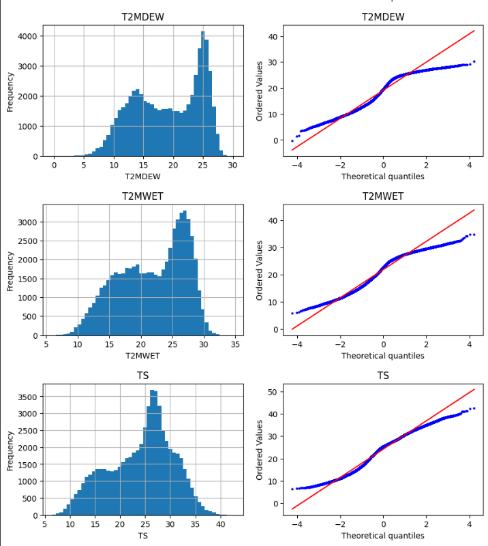
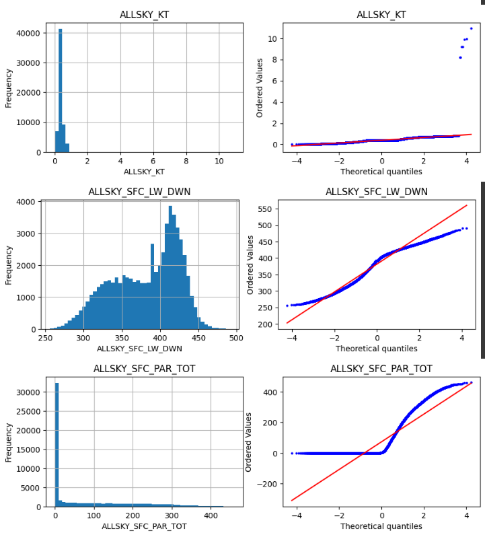
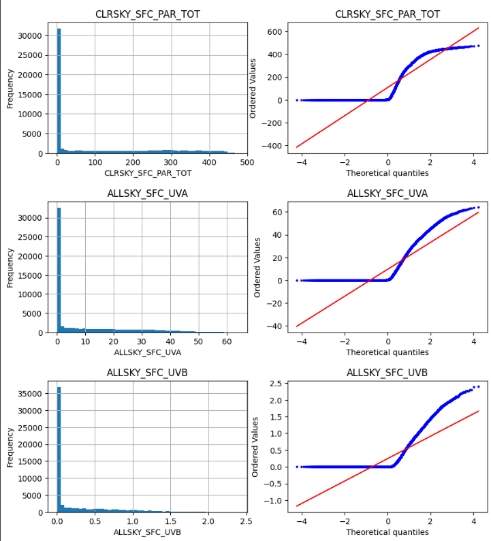
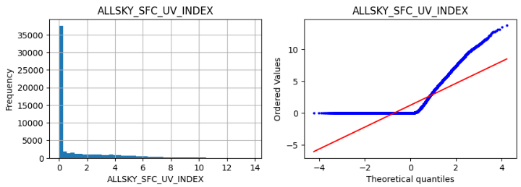
After the completion of one cycle of EDA, we will be first imputing the missing values based on these conclusions, followed by a necessary feature engineering and then further continuing the Exploratory Data Analysis.

**Imputation**

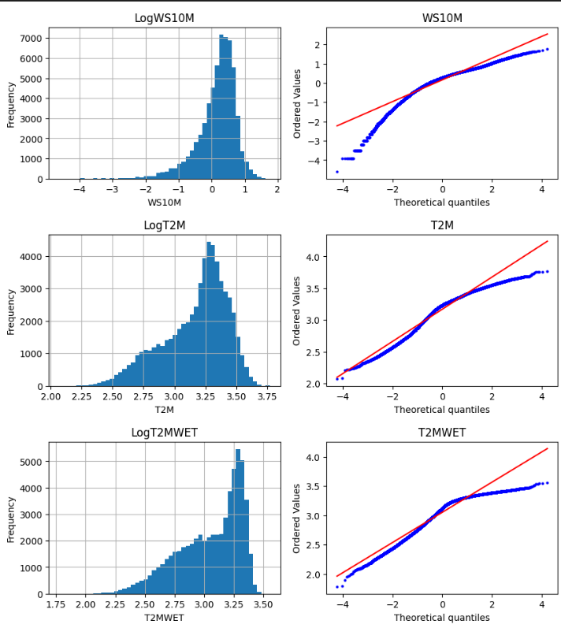
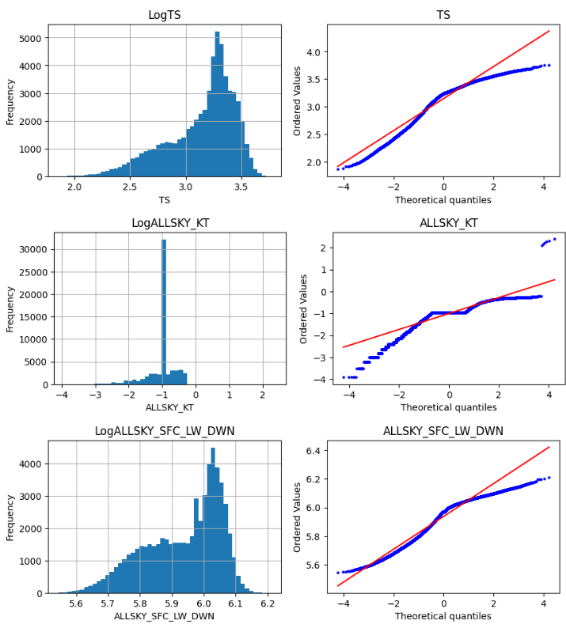
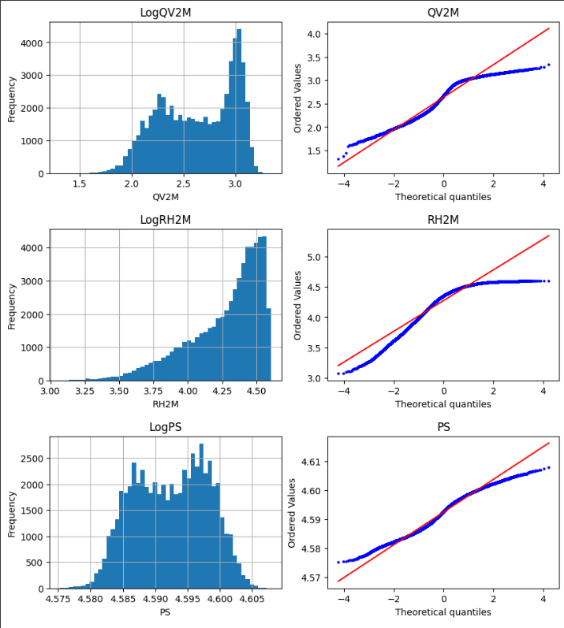
Due to the presence of outliers in the features containing missing values, replace the missing with median or mode of the feature column is preferred. Here, we chose median to replace the missing values. Additionally, for these features, feature engineering was done prior to replacement, with new columns added for each feature assigned binary values as ‘1’ for row with missing values and ‘0’ for row with non-missing values. It is important to be kept in mind while this imputation to only use the information obtained from the training dataset to avoid an overfitting.

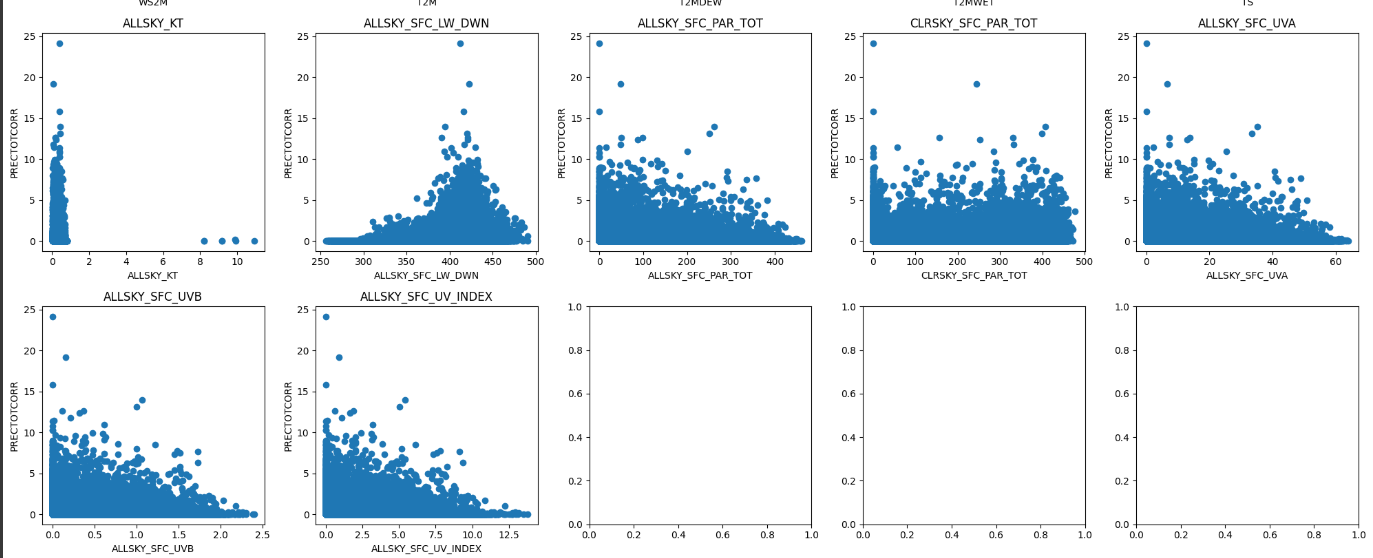
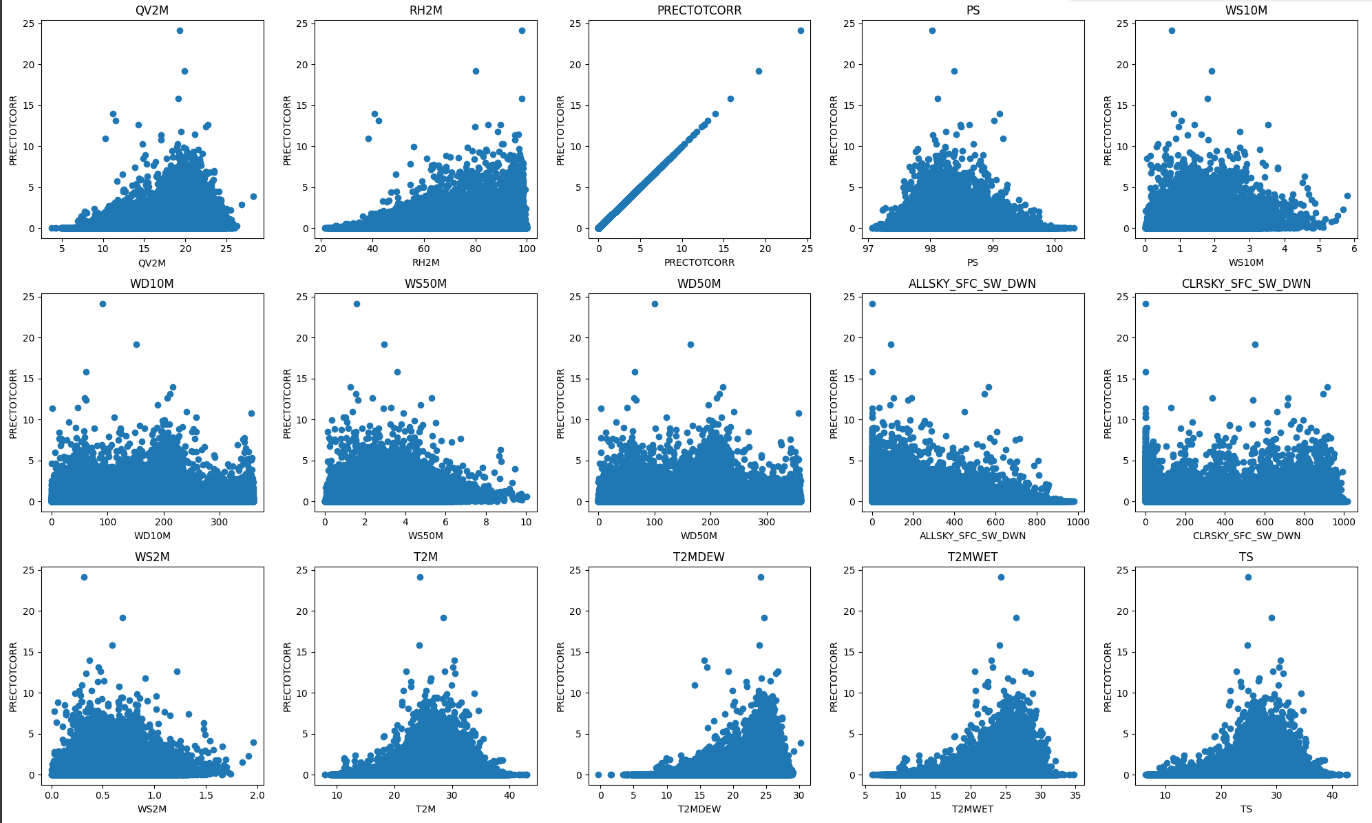
*Visualising Density plots*:

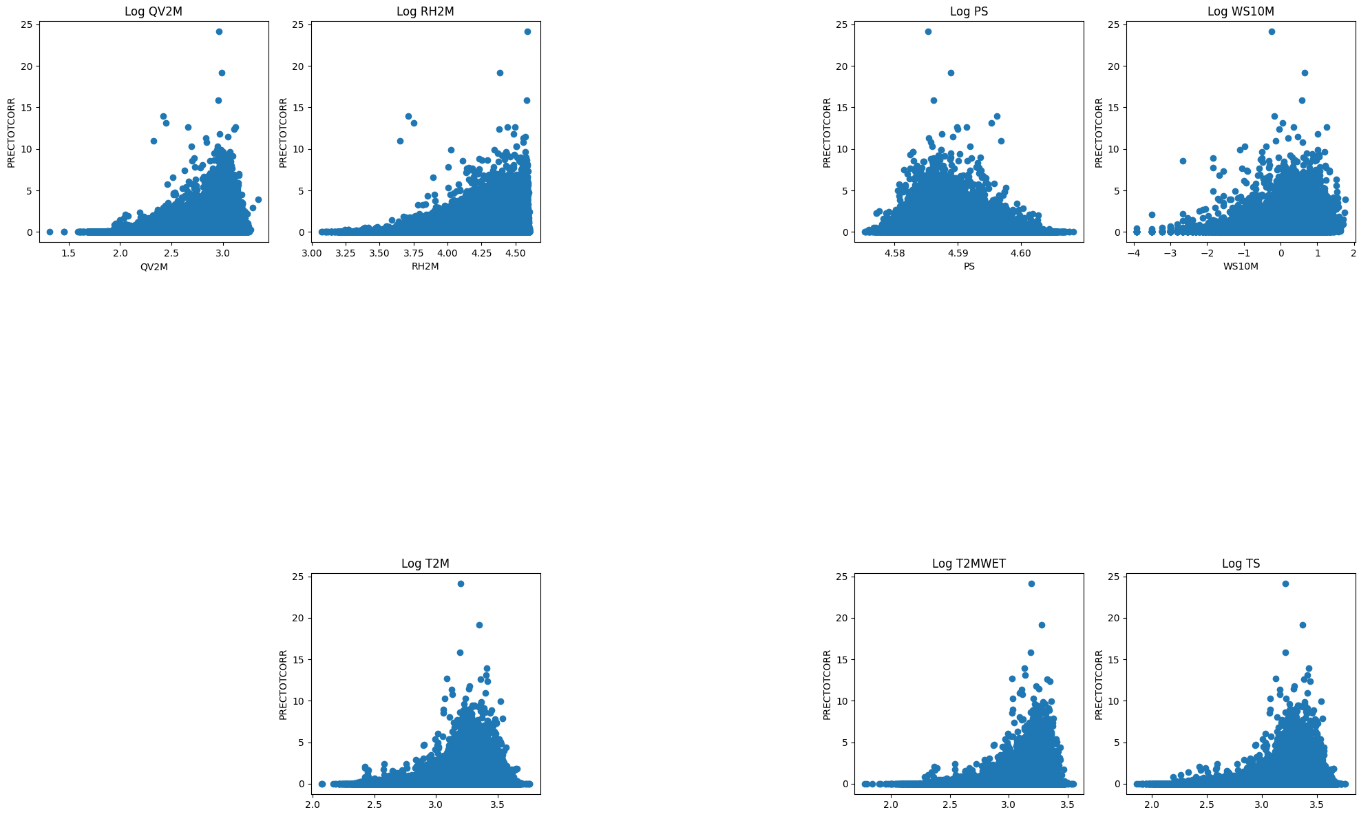
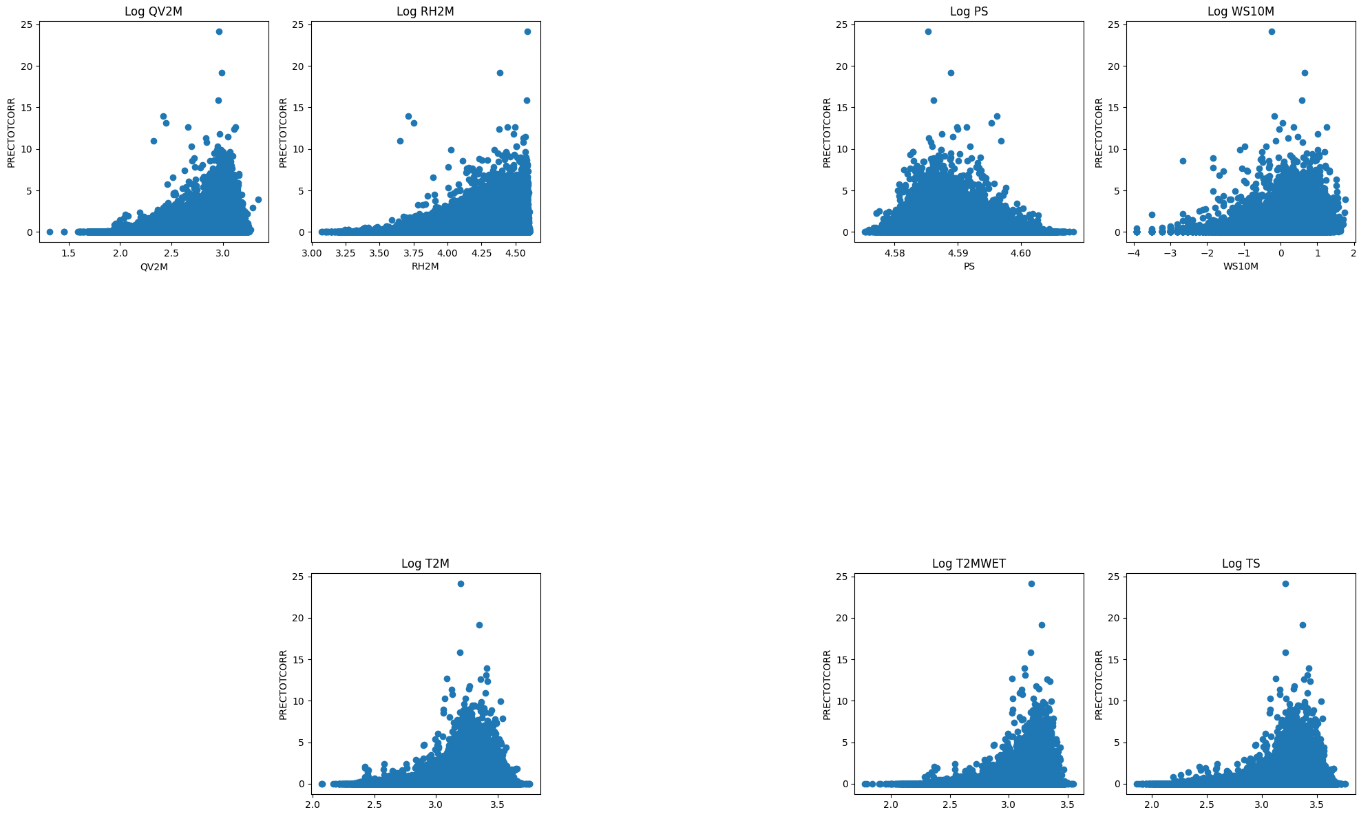
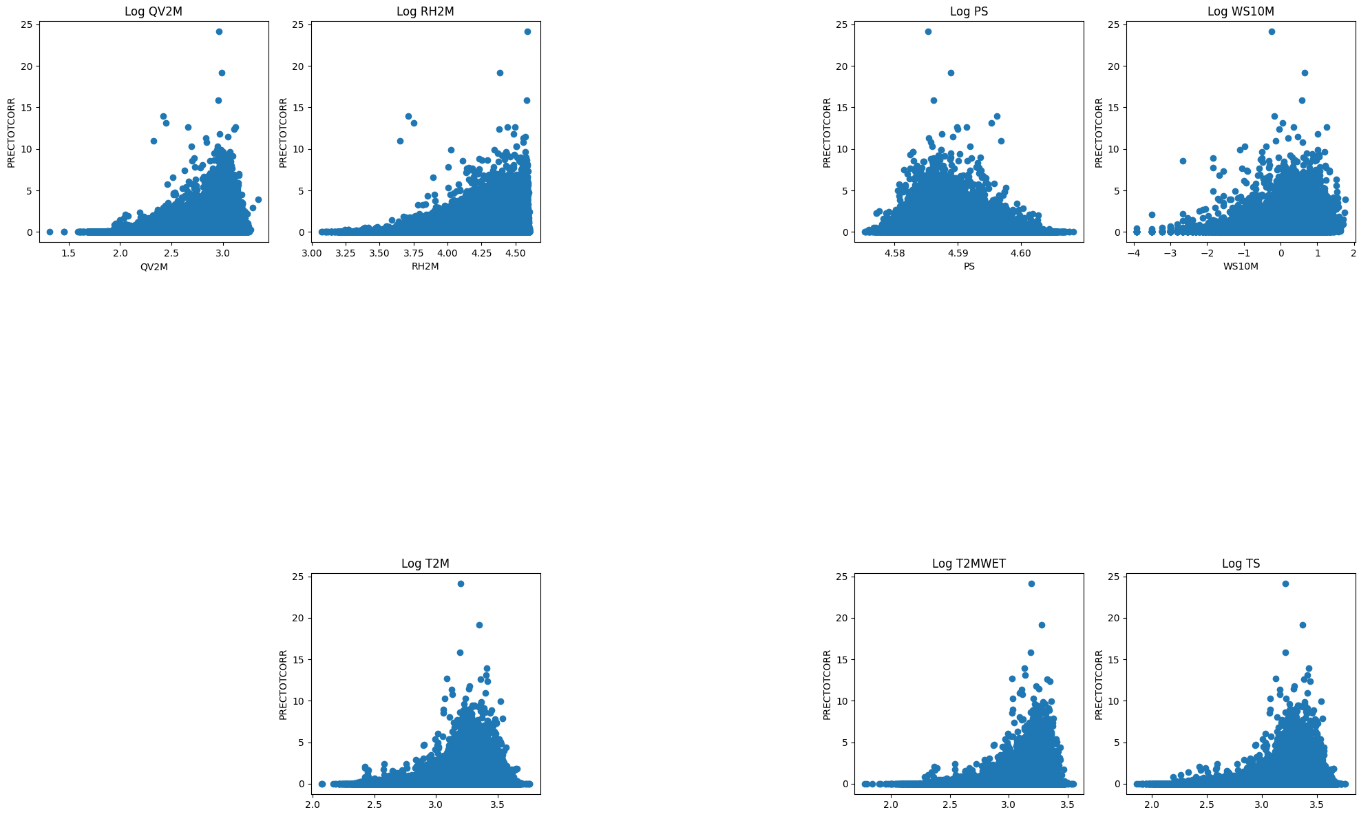
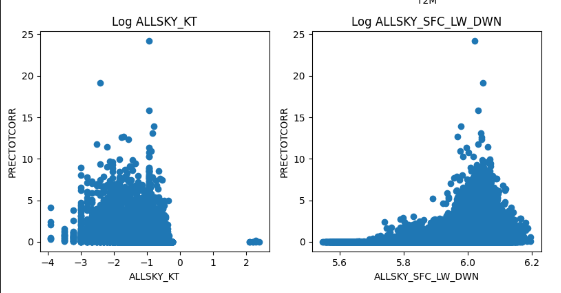
The plots shown below are discretized with 50 bins. Understanding the distribution helps in identifying outliers, assessing skewness, and detecting any potential data anomalies. Moreover, visualizing the distribution can guide feature engineering decisions, such as transformations or binning, to improve model performance. For example, if a feature is heavily skewed, applying transformations like log or square root can help normalize it. This will help in improving model performance. The below plots show the density plots for each of features along with a comparison with a regression line for the normal distribution showing the best fit for a normal distribution.

**

We can visualise from here that many of the features do not fit the normal distribution characteristics. The features with positive values were transformed into their corresponding log normal distributions.

*Visualising Log Transformed Density plots:*

*Visualising Correlation mapping:*

**

(Correlation maps for log normally transformed features)

These maps tell if there is any simple relationship between input features and target variable to be forecasted. The relationship is clearer and stronger if the correlation map is a straight line or curve and it signifies no simple relation if these maps are not such.

Conclusions:

There is no simple relationship between any feature and target variable. We can add lag variables for now and then improve on that with the help of these features. Trying to identify the underlying relationships is not covered for now.

**Feature Engineering**

Handling Temporal Variables

From Exploratory Data Analysis of Temporal Variables, we can conclude:

* We require the month feature; we cannot drop it.
* It is not clear how to handle date feature as of now, we may check the accuracy in both the cases of it being considered and dropped.
* We must consider hourly feature, as it will be crucial for short term predictions and also for the accuracy of long-term prediction.
* It may be important to consider the year feature if the dataset is for a longer period of time; as it can tell historical records of rainfall in the region and help to guess further. This might me more helpful for long-term forecasting.

Additionally, it may be possible to extract several new features helpful for more improvement, for now, we don't cover it.

Handling Numerical Variables:

We need to identify the accurate distribution for each feature, possibly also by incorporating feature extraction. For now, the positively valued features were transformed to their corresponding log normal distributions (This will be improving accuracy to some extent, as discussed in the results).

Also, the outliers were left to be handled for now, as per the complexity of the problem statement. Removal of outliers can lead to serious problems in accurate prediction.

Adding Lag Variables:

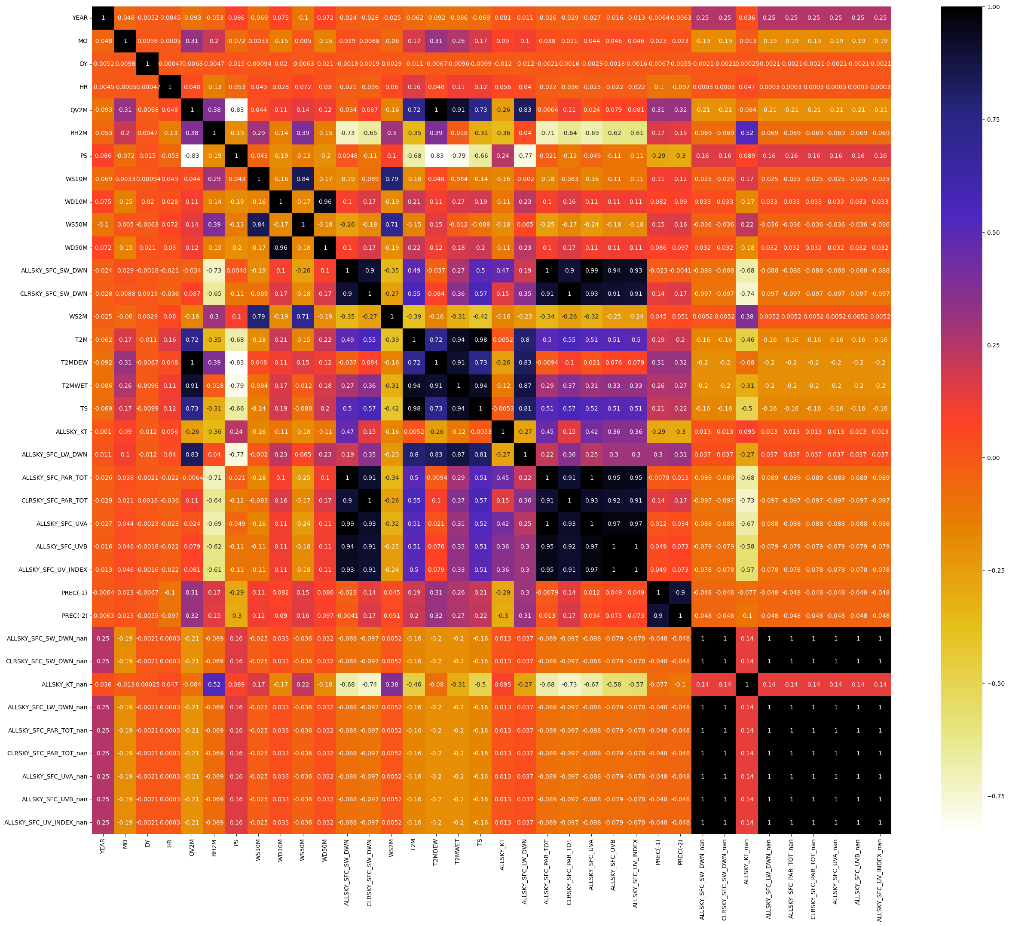
For time series forecasting, we need to add previous target variables’ columns. At first, 48 previous time series features were added, that is, for previous 48 hours. This is a crucial step for accurate prediction as the weatherly parameters are highly unrelated in terms of simple linear relations.

**Feature Selection**

*Identifying the most important features:*

Using “SelectFromModel” function from “feature\_selection” sub-module of the “sklearn” module, the features with importance above the mean importance were derived using “RandomForestRegressor” as the estimator.

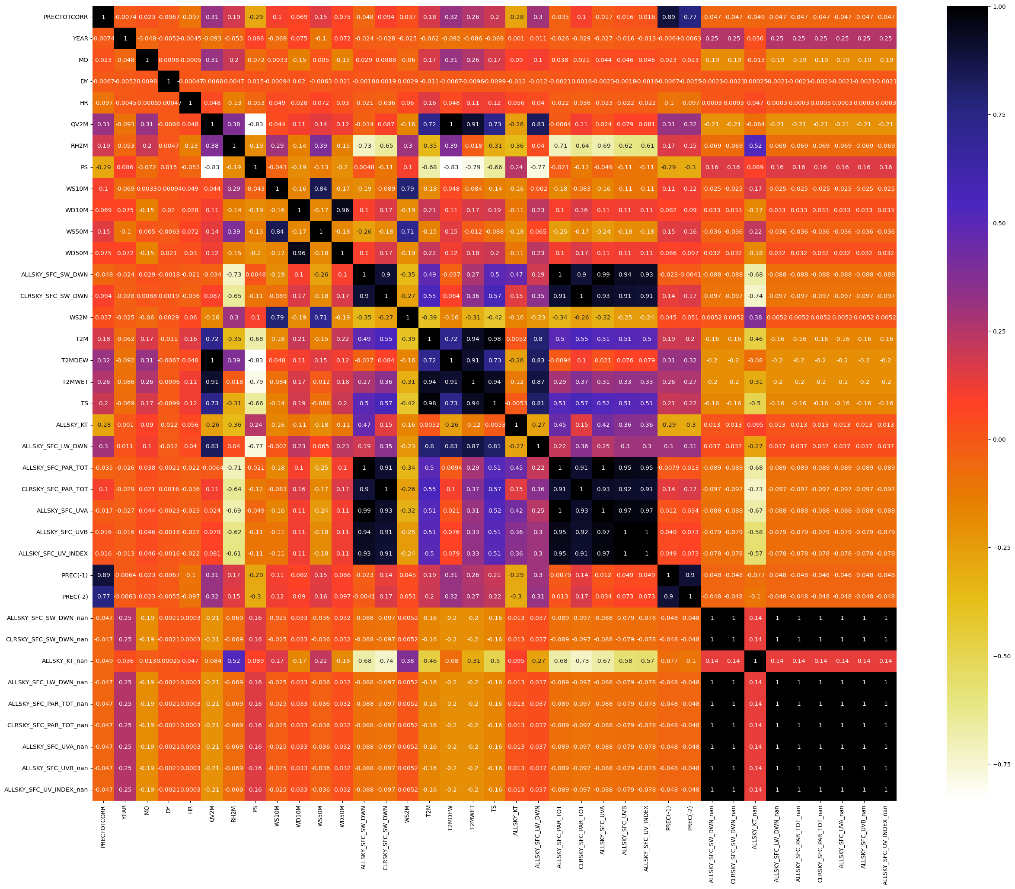
Result: Only the previous 2-hour precipitation features were identified as the most important features for the prediction. Rest of the features were not considered for simplicity of the model.



*Feature selection using Pearson’s correlation heat mapping:*

These heatmaps show the percentage by which any two features are correlated to each other and represents the results in the form of a matrix of values along with the colour for each value for an easier visualisation.

For the feature selection process using the Pearson’s correlation map, a threshold of 80% was chosen. Therefore, for any two features correlated 80% or more to each other, only one of the features was chosen for the final dataset.



Another such a heat map was generated including the target variable for identifying the input features correlated with the target. A threshold of 0.01 was chosen and all the features which are correlated lesser than 1% were removed.

The final dataset contains only 14 features out of total 82 columns obtained, including all of features extracted.

**Feature Scaling**

Using “scale” function from the sklearn’s preprocessing module, the features of the final pre-processed dataset was scaled.

**Model Selection**

For model selection, the chosen appropriate algorithms were: Random Forest, Gradient Boosting, Decision Tree, Linear Regression, Multi-Layer-Perceptron, K-Nearest Neighbors, Support Vector Machine.

The models’ performances were evaluated using “mean\_absolute\_error” (MAE), “mean\_squared\_error” (MSE), and “r2\_score” (R2) from sklearn’s metrics module

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S. No. | Model | Training Metrics | | | Testing Metrics | | |
|  |  | R2 | MAE | RMSE | R2 | MAE | RMSE |
| 1. | Random Forest | 0.9769 | 0.0296 | 0.1182 | 0.8597 | 0.0834 | 0.2849 |
| 2. | Gradient Boosting | 0.8797 | 0.0788 | 0.2700 | 0.8552 | 0.0847 | 0.2894 |
| 3. | Decision Tree | 1.0000 | 0.0000 | 0.0000 | 0.6713 | 0.1139 | 0.4361 |
| 4. | Linear Regression | 0.8041 | 0.1079 | 0.3446 | 0.8273 | 0.1065 | 0.3161 |
| 5. | Multi-Layer-Perceptron | 0.8653 | 0.1012 | 0.2857 | 0.8399 | 0.1037 | 0.3044 |
| 6. | K-Nearest Neighbors | 0.8620 | 0.0880 | 0.2892 | 0.8156 | 0.1072 | 0.3266 |
| 7. | Support Vector Machine | 0.8063 | 0.0977 | 0.3426 | 0.8066 | 0.1013 | 0.3344 |

Most of the models worked quite accurate, the least accurate one being the Decision-Tree Regression. For the final step of Hyper-parameter tuning, we choose only one of the most accurate models; we chose to tune the model based on Random Forest algorithm which showed a test accuracy of 85.97% at an instance.

**Hyper-parameter Tuning**

The final dataset was exported to perform Hyper-parameter tuning on a different platform for faster results. There were 7 hyper-parameters chosen to be tuned with the probable optimum value range sets, represented using key value pair in the dictionary.

rf\_parameters ={

'n\_estimators' : [100, 200, 500,1000],

'max\_depth' : [None, 5, 8, 15, 20],

'min\_samples\_split' : [2, 8, 15, 20],

'min\_samples\_leaf' : [2, 8, 15, 20],

'max\_features' : ["auto",5,7,8,10],

'bootstrap' : [True,False],

'criterion' : ['squared\_error', 'absolute\_error']

}

These value ranges were chosen after multiple testing for optimum parameters. Overall, it took around 2 hours for complete hyper-parameter tuning process.

Results: The best parameters for the model were obtained as:

best\_params = { 'n\_estimators': 500, 'min\_samples\_split': 2, 'min\_samples\_leaf': 2, 'max\_features': 8, 'max\_depth': 15, 'bootstrap':True, 'criterion':'squared\_error' }

**4. Results and Discussions**

The model was validated for the optimum parameters obtained, the accuracy in terms of correlation coefficient increased about 0.5% on an average.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S. No. | Training metrics | | | Testing metrics | | |
| R2 | MAE | RMSE | R2 | MAE | RMSE |
| Random Forest | 0.9525 | 0.0429 | 0.1696 | 0.8630 | 0.0802 | 0.2815 |

The model obtained is optimum is only for a specific location (27.677821° N, 94.971273° E) for which is trained and tuned for optimum hyper-parameters, although it predicts accurately for other locations too.

**Google Colab notebook Source-Code:** https://colab.research.google.com/drive/1Nu7R\_VYWglYFN9cIA5Z\_oGD0FmI8dqxD?usp=sharing

**5. Conclusions**

This study was able to achieve a ‘very short-term rainfall forecasting’ (for about a few hours). This report covers about such a forecasting one hour-prior to precipitation occurrence to an accuracy of about 86.30%. Such a high accuracy was possible mainly due to the lagged target variables added to the data during feature engineering, other features in the dataset did not provide much help except for helping in improving the accuracy by around 8% (compared to simple time series forecasting). The study further aims to achieve short-term, and long-term rainfall forecasting, and further implement the forecasting integrating with the geo-spatial data, thus, achieving flood forecasting.

**6. References**

1. Mirza, M. M. Q. (2011). Climate change, flooding in South Asia and implications. *Regional environmental change*, *11*(Suppl 1), 95-107.
2. López, P. L., Sultana, T., Kafi, M. A. H., Hossain, M. S., Khan, A. S., & Masud, M. S. (2020). Evaluation of global water resources reanalysis data for estimating flood events in the Brahmaputra River Basin. *Water Resources Management*, *34*, 2201-2220.
3. Solomatine, D. P., & PRICE, R. K. (2004). Innovative approaches to flood forecasting using data driven and hybrid modelling. In *Hydroinformatics: (In 2 Volumes, with CD-ROM)* (pp. 1639-1646).
4. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
5. Curth, A., Jeffares, A., & van der Schaar, M. (2024). Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers. *arXiv preprint arXiv:2402.01502*.
6. Singha, C., Swain, K. C., Meliho, M., Abdo, H. G., Almohamad, H., & Al-Mutiry, M. (2022). Spatial analysis of flood hazard zoning map using novel hybrid machine learning technique in Assam, India. *Remote Sensing*, *14*(24), 6229.
7. Myrchiang, D., Ado, M., & Amitab, K. (2023, December). Machine Learning-Based Flood Prediction of Assam. In *2023 International Conference on Computational Intelligence, Networks and Security (ICCINS)* (pp. 1-6). IEEE.
8. Wubalem, A., Tesfaw, G., Dawit, Z., Getahun, B., Mekuria, T., & Jothimani, M. (2021). Comparison of statistical and analytical hierarchy process methods on flood susceptibility mapping: In a case study of the Lake Tana sub-basin in northwestern Ethiopia. *Open Geosciences*, *13*(1), 1668-1688.
9. Addis, A. (2023). GIS​–​ based flood susceptibility mapping using frequency ratio and information value models in upper Abay river basin,​ Ethiopia. *Natural Hazards Research*, *3*(2), 247-256.
10. Ramayanti, S., Nur, A. S., Syifa, M., Panahi, M., Achmad, A. R., Park, S., & Lee, C. W. (2022). Performance comparison of two deep learning models for flood susceptibility map in Beira area, Mozambique. *The Egyptian Journal of Remote Sensing and Space Science*, *25*(4), 1025-1036.
11. Singha, C., Swain, K. C., Meliho, M., Abdo, H. G., Almohamad, H., & Al-Mutiry, M. (2022). Spatial analysis of flood hazard zoning map using novel hybrid machine learning technique in Assam, India. *Remote Sensing*, *14*(24), 6229.
12. Fayaz, S. A., Zaman, M., Kaul, S., & Butt, M. A. (2022). How M5 Model Trees (M5-MT) on continuous data are used in rainfall prediction: An experimental evaluation. *Revue d'Intelligence Artificielle*, *36*(3), 409.
13. Fredyan, R., & Kusuma, G. P. (2022). Spatiotemporal convolutional LSTM with attention mechanism for monthly rainfall prediction. *Commun. Math. Biol. Neurosci.*, *2022*, Article-ID.
14. Khan, M. I., & Maity, R. (2020). Hybrid deep learning approach for multi-step-ahead daily rainfall prediction using GCM simulations. *IEEE Access*, *8*, 52774-52784.
15. SALMA, S., & Ashwitha, A. (2022). Hybrid CNN-LSTM Model: Rainfall Analysis and Prediction for Karnataka Region. *Journal of Theoretical and Applied Information Technology*, *100*(22).
16. Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., ... & Nearing, G. S. (2022). Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *26*(13), 3377-3392.
17. Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045-2062.
18. Rao, G. V., Reddy, K. V., Sridhar, V., Srinivasan, R., Umamahesh, N. V., & Pratap, D. (2022). Evaluation of NCEP-GFS-based Rainfall forecasts over the Nagavali and Vamsadhara basins in India. *Atmospheric Research*, *278*, 106326.
19. Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Machine Learning with Applications*, *7*, 100204.
20. Apicella, A., Isgrò, F., & Prevete, R. (2024). Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning. *arXiv preprint arXiv:2401.13796*.