# Integration of Modern-Era Retrospective analysis for Research and Applications,Version 2 (MERRA-2) data and Lightning Detection Sensor (LDS) data for lightning event prediction

Prisha Sharma[1*], Vinod Kr Sharma[2], Abhinav Kr Shukla[2], Sameer Saran[2]
[1]Thapar Institute of Engineering & Technology, Patiala
[2]Regional Remote Sensing Centre – North, RRSC, NRSC, New Delhi
*Email : psharma1_be22@thapar.edu

(Received on 27 August 2024; In final form on 25 February 2025)

***DOI***: *https://doi.org/10.58825/jog.2025.19.1.183*

**Abstract:** Lightning, a complex and potentially destructive atmospheric phenomenon, poses significant risks to public safety and infrastructure resilience particularly to the nation with limited resources and inadequate early warning system. In tropical regions, the frequency of lightning strikes, particularly during the monsoon season, underscores the importance of early warning systems. The development of accurate detection and timely warning infrastructures is essential to mitigate the impact of lightning events and enhance disaster preparedness. At present 46 lightning detection sensors (LDS) are installed across India, by the Indian Space Research Organization (ISRO). Each LDS is having 300 Km range detection. The network is designed with a 50% overlap to ensure high geolocation accuracy and maintain redundancy in regions with a strong LDS presence. Though this study is focused on a region with a strong existing LDS network, we recognized that there are under developed nations with scarce resources, and inadequate early warning systems where the LDS network is weak or nonexistent. To address this concern, the primary objective of the study is to establish the correlation between the atmospheric parameters and lightning event to predict lightning before it occurs during monsoon season in tropical region. The study will be helpful to predict the lightning for the regions having the sparse LDS network or areas without LDS network by analysing the available MERRA-2 data and factors (Humidity, pressure and precipitation data) causing the lightning using AI techniques. This research relies on data from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) dataset and the LDS dataset, focusing on the region between 86.68°E to 88.52°E longitude and 24.70°N to 26.42°N latitude during June 2022. To improve grid and temporal resolution, data refinement techniques were applied. Following this, statistical analyses, including the Chi-squared test, ANOVA, and Spearman-Rho correlation, were conducted to identify the parameters most strongly correlated with lightning occurrences. Humidity, pressure, and precipitation emerged as the most predictive factors. Using these parameters, the Extra Trees model with bagging was employed to predict lightning occurrences, and a Random Forest classifier was used to predict lightning intensity based on the number of strikes. These models were validated using additional datasets. The findings from this study have the potential to significantly advance early warning systems, particularly in nations with limited or no LDS coverage, thereby enhancing resilience to natural hazards such as lightning across a wider area.

**Keywords:** MERRA-2, Lightning Detection Sensor, Chi-squared test, ANOVA and Spearman-Rho, Extra Trees, bagging, Random Forest classifier

## 1. Introduction

Lightning is among the most powerful and dangerous phenomena of the atmosphere. According to MONGABAY, between 2010 and 2020, lightning killed 3,273 people in Bangladesh, or about four people a week (Islam R, 2022). This aligns in with a subtle alert on how fatal the phenomenon could be and how serious a concern it is to human safety. Yet, despite the leaps and bounds achieved in technology, forecasting and mitigating the effect of lightening remains a stiff challenge for the developing nations. The high mortality indeed points to the urgent need for more effective early warning systems and comprehensive public awareness campaigns to reduce casualties due to lightening.

In the backdrop of above events, development of early warning system for lightning prediction is of most importance specially for the under developed countries, where it causes maximum life loss. Traditional statistical methods and early-warning systems for lightning prediction have shown limitations in effectiveness. Basic statistical analyses and simplistic models often fail to capture the complex, non-linear relationships in lightning activity, leading to issues with data latency and coverage. These methods struggle with real-time accuracy and adaptability. In contrast, advanced techniques like machine learning offer more precise predictions and improved reliability by handling complex data more effectively

Towards this, excellent efforts have predominantly been focused on identifying correlations between various atmospheric and meteorological parameters and lightning activity (Zhang, 2024). This can be further enhanced by incorporating AI/ML for improved predictive accuracy. Based on the correlation of lightning and atmospheric parameters, this study focuses on correlating atmospheric parameters dataset provided by MERRA-2 and developing a model for predicting lightning events by integrating MERRA-2 dataset with LDS dataset using AI/ML for early warning system.

Although the extent of the Lightning Detection Sensor network in India is covering all the regions (Taori et al. 2022, Venkatesh, Degala et al. 2023), ensuring maximum coverage in lightning detection, but there are countries with no or limited LDS coverage. The LDS is prone to electromagnetic interference, there are gaps in coverage in nations where there is limited LDS coverage also it works on the real time detection of the occurrence of lightning rather than early prediction. Our model, therefore, tries to fill these gaps by complementing the nations with some existing LDS network or no LDS network by integrating predictive capabilities using the power of artificial intelligence and machine learning. We can improve the lightning forecast accuracy and reliability by advance prediction of lightning, providing an improved early warning system to the meteorological department, disaster management agencies and public for enhanced safety.

This study aims to enhance the capabilities of lightning prediction and detection by leveraging the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) dataset (GMAO, 2015), along with the LDS data, this research focuses on developing correlation between atmospheric parameters and lightning events, and developing robust model to predict lightning events before they occur. The study area, encompassing regions in Bihar, Jharkhand and West Bengal, was selected for its varied climatic zones, which offer a unique opportunity to analyse lightning patterns across different environmental conditions in tropical regions.

This research employs a combination of statistical analyses and machine learning techniques to identify key meteorological parameters correlated with lightning occurrences. Parameters such as humidity, pressure, and precipitation are found to be the most predictive, and are used to train models like Extra Trees and Random Forest classifiers. The integration of these models with refined meteorological data holds the potential to significantly advance early warning systems, thereby enhancing resilience to lightning-related disasters in under developed nations.

## 2. Study Area

Part of Bihar, Jharkhand and West Bengal states of India is chosen as study area for lightning analysis, covering 86.68°E to 88.52°E in longitude to 24.70°N to 26.42°N in latitude as shown in figure 1. The study area encompasses of districts with diverse rainfall and humidity patterns to offer a study of lightning activity across varied climatic zones and under the buffer zones of LDS locations Patna, Ranchi, Medinapur, Gangtok and Dhubri districts.

These regions are among the most lightning-prone areas in India, with an average of 1,700 cloud-to-ground lightning strikes per day according to LDS data. Bihar, in particular, ranks as one of the most vulnerable states in terms of lightning-related casualties and injuries (Shankar, Kumar, & Sinha, 2024), with West Bengal and Jharkhand also experiencing significant impacts (Paul & Maity, 2023, Mondal, et al., 2023).

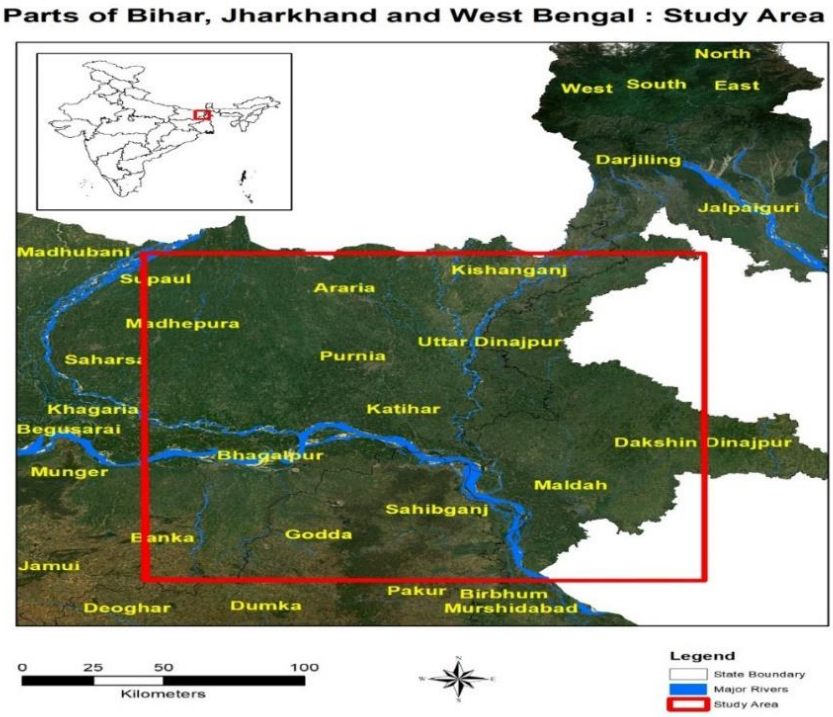


**Figure 1. Targeted area (86.68°E to 88.52°E in longitude to 24.70°N to 26.42°N in latitude (Zoomed)**

The study area encompasses mostly varied climatic zones. For example, in Bihar, it includes Bhagalpur, Kishanganj, and Purnia districts having humid climate (Sattar, Khan, & Banerjee, 2021). In Jharkhand, Godda district experiences a tropical climate with hot summers, a monsoon season with moderate to heavy rainfall and mild winters. The climate is generally humid, particularly during the monsoon. The bounding box choice is strategic in tapping into the variability of lightning across these zones, given the stark contrast in climatic zones across the study area. As such, the variability between these zones will influence lightning pattern in marked ways.

The analysis uses reanalysis atmospheric data to study lightning in these regions in order to understand the potential variables causing lightning strikes in these areas. The aim is to offer insight from these districts to serve as a baseline for risk assessment and management strategies for states in consideration. This could be used in understanding the dynamics of lightning in different climate zones and provides a method for observing this in most lightning affected zones in tropical region.

## 3. Database and Methods

### 3.1. Dataset

The target area in this research spans from 86.68°E to 88.52°E in longitude to 24.70°N to 26.42°N in latitude. The research objective was to predict the lightning event using the correlated parameters for the area under consideration. To achieve the objective, dataset used in this study can be divided into two categories:

First, the dataset provided by LDS (Lightning Detection Sensor) gathered from the electromagnetic pulse emitted by a lightning strike. Parameters like time, Longitude, Latitude, Type and Current Id were sourced from Lightning Detection Sensors located in these areas as shown in Table 1 (Source: https://bhuvan-app1.nrsc.gov.in/lightning/). As mentioned earlier, India has 46 lightning detection sensors (LDS) with over 98% confidence within a 300 km detection range for each sensor. The network is designed with a 50% overlap to ensure high geo-location accuracy and maintain redundancy in regions with a strong LDS presence (Venkatesh et al. 2023).

Second the atmospheric reanalysis dataset, Modern- Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) obtained from the official website of Global Modeling and Assimilation Office (GMAO) maintained by National Aeronautics and Space Administration (NASA). Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA-2) is a significantly improved version of the original MERRA dataset which begins in 1980. MERRA-2 takes advantage of advancements in the assimilation system to include newer observations, such as hyperspectral radiance, microwave observations, and GPS-Radio Occultation measurements, resulting in a more detailed and accurate atmospheric reanalysis. For example, the MERRA-2 atmospheric reanalysis dataset, specifically the M2I1NXASM (or inst1_2d_asm_Nx) collection provides 2-dimensional hourly instantaneous data which is very important for meteorological conditions analysis. This dataset has a total of 18 parameters which include temperature at 2 meters and at 10 meters, wind components at 2 meters, 10 meters, and 50 meters, sea level pressure, total precipitable water, total cloud ice, specific humidity, and sea level pressure shown in Table 2. These parameters give a good view of the conditions in the near-surface and upper-atmosphere, specifically wind patterns, water vapor distribution, pressure variances, etc., and are very important for weather and climate studies. The improvements in MERRA-2 mean that the representation of the atmosphere is more reliable and there are so many potential scientific and practical applications that rely on the continual use and availability of the MERRA-2 ((GMAO), 2015).

LDS's location, time, and type of lightning strikes are highly accurate due to India' s well-positioned 46 sensor networks, which provide real-time data. Reliable monitoring of lightning is facilitated by this. By utilizing cutting-edge technology, MERRA-2 offers comprehensive atmospheric data, including temperature, wind, and humidity, to provide context. The identification of lightning is facilitated by this.

By combining them, they provide a comprehensive understanding of the lightning events and the surrounding atmospheric conditions, making them ideal for forecasting nearby storms.

### 3.2. Statistical Analyses Techniques

**Chi-Squared Test**- The Chi squared test is a statistical method for testing independence between categorical variables such as lightning data and meteorological parameters and also provides insight into the impact of those variables. The Chi-Square test does not explain to the strength of the relationship between the variables, but it does inform if there is a relationship. (Singh, Lavakush., 2022)

**ANOVA and Spearman-Rho Correlation**- In addition to the correlational statistics, we used ANOVA and Spearman- Rho approach to examine the relationship between lightning events and selected parameters as a group, in order to have a complete knowledge of the possible relationships.

### 3.3. Machine Learning Techniques

**Extra Trees (Extremely Randomized Trees)**- Extra Trees offer a variation to the traditional Random Forests through the introduction of additional levels of randomness in how it determines the split points for decision trees. This deliberate introduction of randomness in Extra Trees is beneficial because it reduces the model's ability to adapt to the training data (overfit) and especially in cases where the data itself is noisy or the patterns are intricate. By preventing the model from adapting or fitting too closely to the training data examples, Extra Trees supports the generalization of the model when it is confronted with new data instances.

(Abdelkader, Jaziri, & Bernard, 2019).

**Bagging (Bootstrap Aggregating)-** It is an ensemble machine learning method specifically designed to increase the stabilityand accuracy of predictive models. The approach is based on creating a series of different subsets from the original dataset by means of bootstrapping (random sampling with replacement). Each subset is then used to train a different decision tree classifier or machine learning model that is typically a high-variance mode. (Rocca, J. 2019)

**Table 1. Lightning Detection Sensor (LDS) data (Source: https://bhuvan-app1.nrsc.gov.in/lightning/)**

| Time | Strike Type | Current CLS | Current ID | cx | cy | gid | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|
| 4:51:07 AM | CG | C1 | 13412 | 72.25 | 28.55 | G067542 | 25.229 | 88.50429 |
| 4:54:56 AM | CG | C1 | 13589 | 72.25 | 28.55 | G067542 | 26.10299 | 87.77047 |
| 5:03:51 AM | CG | C1 | 14264 | 72.25 | 28.55 | G067542 | 25.08398 | 88.09991 |
| 6:25:32 AM | CG | C2 | 21198 | 72.25 | 28.55 | G067542 | 26.20184 | 87.41916 |
| 6:39:48 AM | CG | C1 | 18073 | 72.25 | 28.55 | G067542 | 26.33062 | 87.62166 |
| 7:00:20 AM | CG | C1 | 16747 | 72.25 | 28.55 | G067542 | 26.30667 | 87.10836 |
| 7:52:39 AM | CG | C2 | 23675 | 72.25 | 28.55 | G067542 | 25.92526 | 87.29363 |
| 7:54:09 AM | CG | C1 | 19577 | 72.25 | 28.55 | G067542 | 24.83792 | 86.76564 |
| 7:54:12 AM | CG | C2 | 27821 | 72.25 | 28.55 | G067542 | 24.81592 | 86.7205 |
| 7:56:48 AM | CG | C1 | 15237 | 72.25 | 28.55 | G067542 | 24.95233 | 86.92459 |
| 7:57:09 AM | CG | C2 | 28092 | 72.25 | 28.55 | G067542 | 26.32156 | 87.05604 |
| 8:04:05 AM | CG | C2 | 32092 | 72.25 | 28.55 | G067542 | 26.30038 | 86.77194 |
| 8:04:32 AM | CG | C1 | 19674 | 72.25 | 28.55 | G067542 | 26.3275 | 86.87326 |

**Table 2. M2I1NXASM (or inst1_2d_asm_Nx) is an instantaneous 2-dimensional hourly data ((GMAO), 2015)**

| Name | Dim | Description | Units |
|---|---|---|---|
| DISPH | tyx | Zero plane displacement height | m |
| PS | tyx | Surface pressure | Pa |
| QV10M | tyx | 10-meter specific humidity | kg $kg^{-1}$ |
| QV2M | tyx | 2-meter specific humidity | kg $kg^{-1}$ |
| SLP | tyx | Sea level pressure | Pa |
| T10M | tyx | 10-meter air temperature | K |
| T2M | tyx | 2-meter air temperature | K |
| TO3 | tyx | Total column ozone | Dobsons |
| TOX | tyx | Total column odd oxygen | kg $m^{-2}$ |
| TQI | tyx | Total precipitable ice water | kg $m^{-2}$ |
| TQL | tyx | Total precipitable liquid water | kg $m^{-2}$ |
| TQV | tyx | Total precipitable water vapor | kg $m^{-2}$ |
| TROPPB | tyx | Tropopause pressure based on blended estimate | Pa |
| TROPPT | tyx | Tropopause pressure based on thermal estimate | Pa |
| TROPPV | tyx | Tropopause pressure based on EPV estimate | Pa |
| TROPQ | tyx | Tropopause specific humidity using blended TROPP estimate | kg $kg^{-1}$ |
| TROPT | tyx | Tropopause temperature using blended TROPP estimate | K |
| TS | tyx | Surface skin temperature | K |
| U10M | tyx | 10-meter eastward wind | m $s^{-1}$ |
| U2M | tyx | 2-meter eastward wind | m $s^{-1}$ |
| U50M | tyx | Eastward wind at 50 meters | m $s^{-1}$ |
| V10M | tyx | 10-meter northward wind | m $s^{-1}$ |
| V2M | tyx | 2-meter northward wind | m $s^{-1}$ |
| V50M | tyx | Northward wind at 50 meters | m $s^{-1}$ |

## 4. Methodology

### 4.1. Preprocessing Dataset

The LDS dataset provides lightning data a point information with a geo-locational accuracy of 100 m, at regular time interval, which necessitates interpolation for effective utilization in this study. On the other hand, the MERRA-2 dataset, having spatial resolution of 0.5 ° x 0.625 ° and a temporal resolution of 1 hour, also require processing.

**Temporal Alignment-** Unlike the LDS dataset, which provides point-based information about lightning strikes with an accuracy of 100 meters, the MERRA-2 dataset has a temporal resolution of 1 hour which is too coarse for capturing lightning activity. To make the datasets comparable, the temporal resolution of MERRA-2 was increased to 5 minutes using interpolation. This adjustment aligns the MERRA-2 dataset with the higher temporal resolution of the LDS data. The temporal resolution of the MERRA-2 dataset was adjusted for better compatibility with the LDS dataset.

**Spatial Alignment-** To align MERRA-2's lower spatial resolution with the higher resolution of the LDS dataset, spline interpolation was applied to MERRA-2 data. This method ensures smooth and accurate adjustments of spatial values to match the finer grid structure. The interpolation facilitates better compatibility and integration of the two datasets for spatial analysis.

To align MERRA-2's lower spatial resolution with the higher resolution of the LDS dataset, spline interpolation was applied to MERRA-2 data. This method ensures smooth and accurate adjustments of spatial values to match the finer grid structure. The interpolation facilitates better compatibility and integration of the two datasets for spatial analysis. (Laipelt et al. 2020),

**Critical Adjustments:**

**Temporal Interpolation:** Linear interpolation was employed to increase the temporal resolution of MERRA-2 from 1 hour to 5 minutes, providing a closer match to the temporal accuracy of LDS data.

**Spatial Interpolation:** Spline interpolation was used to enhance MERRA-2's spatial resolution to align with LDS's finer granularity.

These adjustments were essential to harmonize the datasets, accurately capture lightning activity, and avoid pitfalls like poor model accuracy or low correlation in the results. By aligning the datasets effectively, the analysis ensures a more reliable evaluation of atmospheric conditions and their influence on lightning events. (Zhang, 2024)

### 4.2. Analysis of the Correlation tests:

*4.2.1. Chi Squared Test:*

From June 1,2022 to June 10, 2022 a total of 15,334 lightning points was observed by LDS in the study area.

We created a grid of the study area to map LDS data with the corresponding MERRA-2 data to obtain the parameters of the particular strike. Also, we made a non-lightning points dataset to get an accurate correlation result. Lightning points obtained by mapping the LDS and MERRA-2 dataset had 15,334 rows and 22 columns including Longitude, Latitude, date, time and 18 parameters (TQV, TQI, QV10M, PS, SLP, TO3, TROPPT, TROPPV, TROPQ, U10M, U2M, V2M, V10M, V50M, TROPT, T2M, T10M). Table 2. Non-lightning dataset was prepared with the same layout because many results showed, as the sampling portion increases it leads to an increase in the sample size, and the result of the chi squared test also gradually increases and may lead to degraded results. For Chi Squared correlation analysis between LDS data and MERRA-2 dataset data was divided into 10 equally spaced intervals with 95% confidence as it produced better results because increasing the intervals lead to occurrence of empty intervals. This became more frequent as the sample division increased 14. When the number of divisions is too small, then also it can lead to inaccurate analysis.

Out of these 18 parameters, 10 best parameters were chosen on the basis of Chi Squared Test and their values as shown in Table 4 and figure 2. To avoid overfitting, a threshold was determined at the point where changes in the bar graph became minimal, indicating that further refinements would not significantly impact the results as shown with orange line in Figure 2. The parameters investigated included Total Perceptible Water (TQV), Total Ozone Column (TO3), Total Cloud Ice (TQI), Sea Level Pressure (SLP), Tropopause Temperature (TROPT), Surface Pressure (PS), Temperature at 10 meters (T10M), Temperature at 2 meters (T2M), and Wind Speed at 50 meters (V50M).

*4.2.2 ANOVA, and Spearman-Rho correlation:*

ANOVA and Spearman-Rho correlation test were performed to get the most suitable correlated parameters, and the best 4 parameters extracted were; Specific Humidity (QV10M), Total Precipitable Water (TQV), Total Cloud Ice (TQI) and Surface Pressure (PS) as shown in Table 3 below. (Akhter, Roy, & Midya, 2024) (Han, Luo, Wu, & et , 2021) (Chakraborty et all. 2021).

### 4.3. Finding the Best Model

According to studies, Extra Trees performs well on complex datasets, particularly its ensemble approach that makes it efficient and presents an added randomness in the selection of splits. By randomly sampling subsets of the data and collecting the predictions of multiple decision trees, the risk of overfitting to specific patterns in the MERRA-2 and LDS datasets is reduced. This is important to maintain the validation of the model and their accuracy in predicting lightning events based on correlated parameters.

**Table 3. Comparison between Selected Atmospheric Parameter**

| Parameter | Total Precipitable Water (TQV) | Total Cloud Ice (TQI) | Surface Pressure (PSFC) | Specific Humidity (QV10M) |
|---|---|---|---|---|
| **Represents** | Total water vapor in the column | Total ice in the cloud column | Atmospheric pressure at the surface | Water vapor at 10m height |
| **Phase** | Vapor | Solid (ice crystals) | N/A | Vapor |
| **Vertical Extent** | Entire atmospheric column | Entire atmospheric column | Near surface only | Near surface only |
| **Physical Relevance** | Precipitation potential | Cloud radiative properties | Weather forecasting and dynamics | Surface weather conditions |

**Table 4. Chi Squared Test**

| Parameter | Chi-squared |
|---|---|
| TQV | 12,594.63 |
| TO3 | 8,601.15 |
| TQI | 7,779.13 |
| QV10M | 6,825.60 |
| SLP | 6,614.18 |
| TROPT | 4,990.00 |
| PS | 4,732.94 |
| T10M | 4,613.85 |
| T2M | 4,058.93 |
| V50M | 3,981.77 |
| U2M | 3,342.45 |
| U10M | 3,314.40 |
| TROPPT | 3,134.42 |
| U50M | 2,988.22 |
| TROPQ | 2,911.65 |
| V10M | 2,365.18 |
| V2M | 2,281.47 |
| TROPPV | 735.87 |

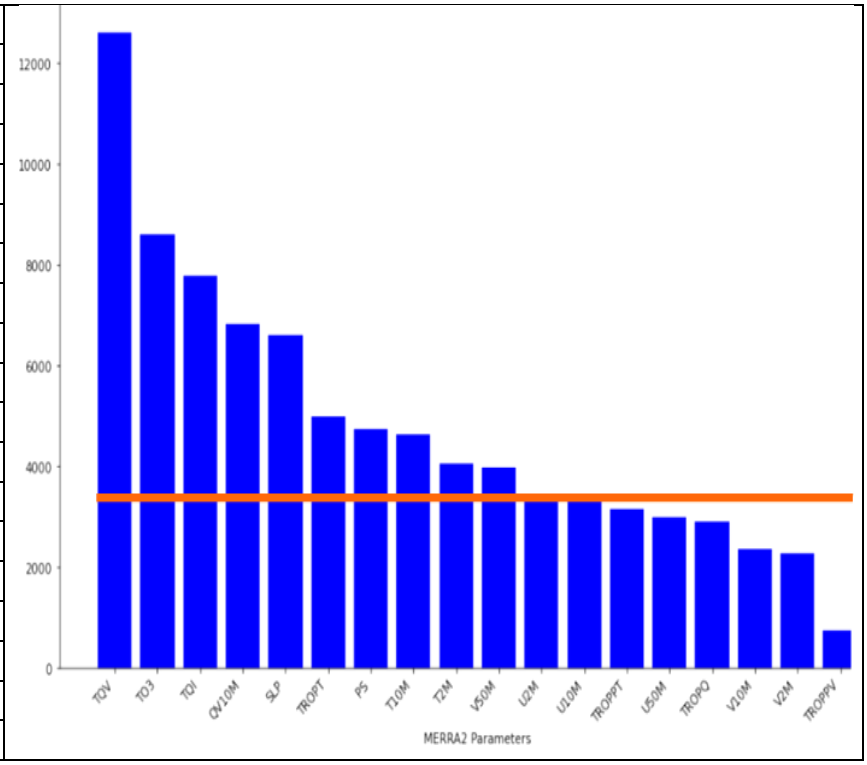


**Figure 2. Chi Squared Test**

To ensure the results, after finding the best parameters, we used PyCaret classifier with 5-fold cross-validation which is an open-source, low-code machine learning library in Python designed to simplify and speed up the machine learning process. By automating many of the tasks involved, it helps manage and experiment with models more efficiently, making work faster and more productive. (Ali, 2020).

To find the most suited model, the dataset used was the LDS dataset with corresponding atmospheric parameters from reanalysis MERRA-2 dataset and non-lightning points in target area for the targeted duration. Out of 15 models compared by PyCaret Classifier, Extra trees gave the best results.

However, upon examining the learning and validation curves, we observed that the model exhibits signs of over-fitting

### 4.4. Using Bagging to Mitigate Overfitting

To eliminate the signs of overfitting, Bagging was used.

Bagging (bootstrap aggregating) is an ensemble method that involves training several models separately on different random parts of the data and then combining their predictions by either voting or averaging. This approach helps to improve accuracy and robustness by leveraging the strengths of each model. It can reduce the chance of an overfit model, resulting in improved model accuracy on unseen data. Multiple models trained on different subsets of data average out their predictions, leading to lower variance than a single model.

Before applying Bagging, numerical features were preprocessed using Simple Imputer, and the dataset was split into training (80%) and testing sets (20%). The Extra Trees model was chosen as the base estimator for Bagging. After implementing Bagging with the Extra Trees model, the accuracy achieved was 0.8951. To ensure the reliability of the results, learning curves as shown in figure 3 and confusion matrix were analyzed and yielded satisfactory outcomes, confirming the effectiveness of the approach in managing over-fitting.

### 4.5. Model Validation

As model was trained for the duration of 1 June 2022 to 10 June to 2022, therefore to evaluate the performance of our Extra Trees model with Bagging for detecting lightning events, we conducted a validation test using a dataset of lightning occurrences provided by LDS network recorded on 8 June 2023. Our test dataset comprised a total of 350 data points within the targeted area covered by the bounding box in Bihar, Jharkhand and West Bengal, covering 86.68°E to 88.52°E in longitude to 24.70°N to 26.42°N in latitude

**Table 5- Results Overview**

| | |
|---|---|
| Total Data Points: | **350** |
| Correctly Marked Points (True Positive, TP): | **331** |
| Incorrectly Marked Points (False Negatives, FN): | **19** |

Shown in figure 4(a) and figure 4(b).
Based on these results, we calculated several performance metrics to assess the efficacy of our model. The metrics include accuracy and recall. These metrics are defined as follows:

**Accuracy:** Accuracy measures the proportion of correctly identified points (both true positives and true negatives) relative to the total number of points. It provides an overall indication of the model's correctness. Accuracy=Correctly Marked Points/Total Points=331/350

**Recall:** To calculate recall, we need to consider the following definitions:

**True Positives (TP):** Points correctly identified as lightning events.
**False Negatives (FN):** Points incorrectly identified as non-lightning events.

**Recall:** Recall measures the proportion of actual lightning events that were correctly identified by the model. In our case, since the total number of false negatives (incorrectly marked points) is 19, the recall can be calculated as under:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} = \frac{331}{331 + 19}$$

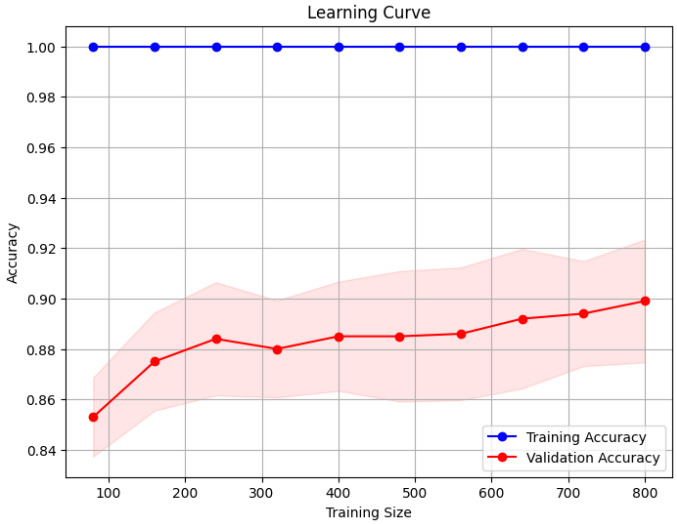


**Figure 3. Learning Curve of Extra Trees with Bagging**

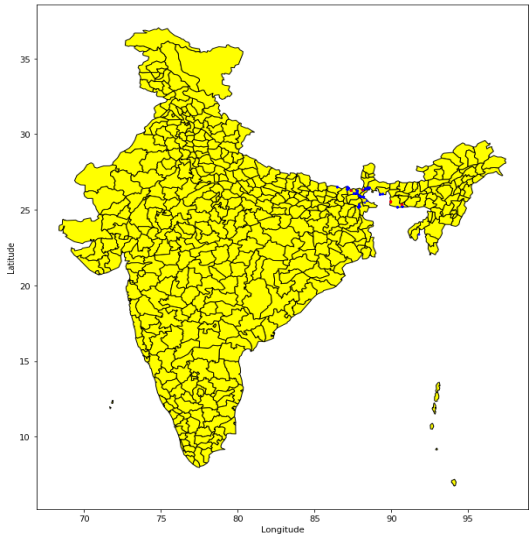


**Figure 4(a). Testing dataset where blue represents True Positive and purple represents False Negative with the buffer of 5.55 Km (Scale- 200 Km)**

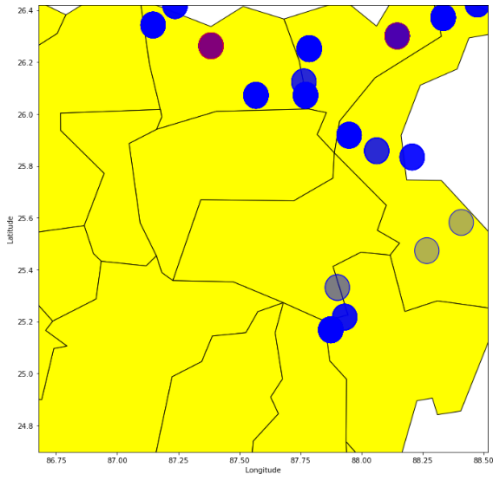


**Figure 4(b). Testing dataset where blue represents True Positive and purple represents False Negative with the buffer of 5.55 Km (Zoomed, Scale – 10 Km).**

### 4.6. Intensity Categories Based on Daily Lightning Strike Frequency

Data from Bhuvan ISRO was used to classify intensity which categorizes the Indian grid over 10 Km segments. The number of lightning marks in each grid is calculated for each day and categorize the intensity as follows Table 5.

To match it with our data we began by importing the lightning strike data. Following this, we established a grid of 10 km by 10 km over the area by calculating the boundaries of each cell within the grid using latitude and longitude increments.We then took a tally of the amount of lightning strikes in each of the grid cells on a daily basis and gave each level of lightning intensity a score. We applied specific threshold data to categorize the intensity as 'Very Heavy', 'Heavy', 'Moderate', 'Low', 'Very Low', or 'No Lightning'. Lastly, we joined this classified intensity data to the original dataset in order to provide a more extensive overview of the lightning activity.

**Table 6. Intensity Classification according to Bhuvan (Source: https://bhuvan-app1.nrsc.gov.in/lightning/)**

| Lightning Strikes per Day | Intensity Category |
|---|---|
| >500 | Very High |
| 126-500 | Heavy |
| 25-125 | Moderate |
| 6-124 | Low |
| 0-5 | Very Low |

### 4.7. Correlational Analysis and Training the Model

Various correlations tests were performed for example Kendall Tau and Pearson Correlation Test, leading to the conclusion that the parameters influencing lightning intensity differ somewhat from those affecting lightning/non-lightning conditions. From an initial set of 18 parameters, specific humidity and temperature at 2 meters were identified as the two most influential factors moderately affecting number of lightning strikes per day as shown in Figure 5(a) and Figure 5(b).

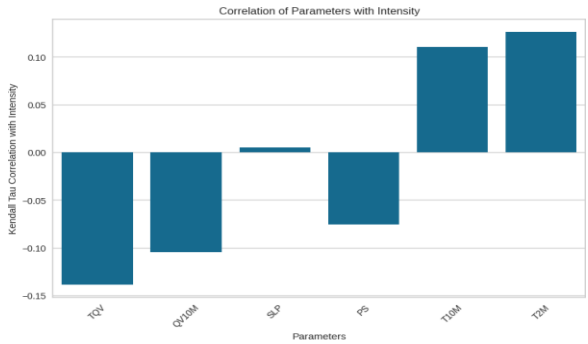


**Figure 5(a). Kendall Tau Correlation Test**

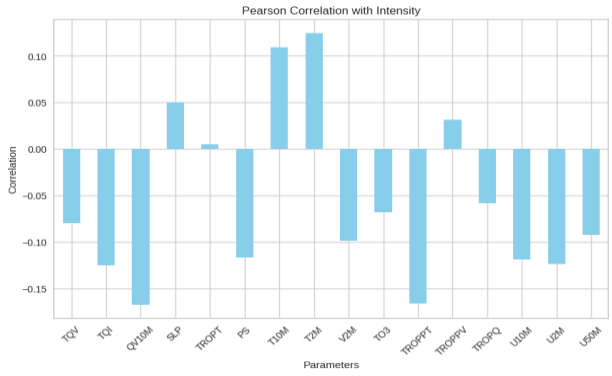


**Figure 5(b). Pearson Correlation Test**

A Random Forest model was selected for its ability to capture complex relationships within the data. However, the accuracy achieved for predicting the lightning intensity with this model was somewhat lower i.e. at 0.79. The corresponding confusion matrix is as follows:

**Confusion Matrix:**

$$\begin{bmatrix} 1208 & 15 & 139 & 4 \\ 41 & 127 & 90 & 24 \\ 182 & 51 & 1019 & 12 \\ 22 & 20 & 31 & 82 \end{bmatrix}$$

However, to generate real-time intensity predictor with collaboration with LDS dataset, data from the Indian Meteorological Department (IMD) were integrated using

API to predict the lightning intensity as IMD provides API for relative humidity and temperature at 2m.

Relative humidity can be converted in specific humidity using this formula

**$x = 0.622\ \varphi\ \rho ws / (\rho - \rho ws)\ 100\%$**

where,
x = specific humidity of air vapor mixture (kg/kg)
φ = relative humidity (%)
ρws = density of water vapor (kg/m3)
ρ = density of the moist or humid air (kg/m3)

The study underscores the distinct impact of these parameters on lightning occurrence and highlights the effectiveness of the model in forecasting intensity, providing valuable insights for meteorological prediction and risk mitigation strategies.

## 5. Results

### 5.1. Correlation Analysis

**Chi-Squared Test:** The Chi-Squared Test was employed to examine the correlation between lightning occurrences and various meteorological parameters. This analysis revealed significant correlations, identifying key parameters such as humidity, pressure, and precipitation as pivotal factors in predicting lightning events. The statistical significance of these parameters underscores their critical role in influencing lightning activity, highlighting their relevance for predictive modeling.

**ANOVA and Spearman-Rho Correlation:** Complementary analyses using ANOVA and Spearman-Rho Correlation further validated the importance of specific meteorological parameters. Among the parameters examined, specific humidity, total precipitable water, total cloud ice, and surface pressure emerged as the most influential. These findings reinforce the notion that these parameters are integral to understanding the relationship between meteorological conditions and lightning occurrences. The use of multiple analytical methods ensures a robust validation of the parameters crucial for lightning prediction.

### 5.2. Machine Learning Model Selection

**Extra Trees Model:** The Extra Trees model was selected for its robustness and ability to handle complex datasets effectively. This model excels in reducing overfitting, making it suitable for lightning prediction tasks. Through PyCaret's 5-fold cross-validation, the Extra Trees model demonstrated exceptional performance metrics.
These high-performance metrics indicate that the Extra Trees model is highly effective in predicting lightning occurrences, with minimal risk of overfitting. The model's accuracy and recall scores reflect its reliability and precision in predicting the lightning events in monsoon season of tropical region.

**Bagging Technique:** To further address overfitting and enhance model stability, the Bagging (Bootstrap Aggregating) technique was applied to the Extra Trees model. This approach improved the model's generalization capability, resulting in acceptable accuracy. The application of Bagging was validated through learning curves and confusion matrix analyses, which confirmed the model's enhanced ability to generalize across unseen data. This adjustment significantly improved the model's stability and predictive performance.

### 5.3. Intensity Prediction

**Intensity Classification:** As per Bhuvan ISRO (Bhuvan, NRSC, ISRO, Government of India, n.d.) lightning intensity is categorized into six levels ('Very Heavy', 'Heavy', 'Moderate', 'Low', 'VeryLow', 'No Lightning') based on daily lightning strike frequency. This classification framework provided a detailed and clear view of lightning activity within the study area. It enabled a more granular analysis of lightning intensity, offering insights into the varying intensities of lightning activity.

**Impact of Meteorological Parameters on Predicting Intensities:** Specific humidity and temperature at 2 meters were identified as significant predictors of lightning intensity. These parameters played a crucial role in differentiating between various levels of lightning activity. The relationship between these meteorological factors and lightning intensity underscores their importance in predicting and classifying lightning events.

### 5.4. Random Forest Model for Intensity Prediction

**Model Performance:** A Random Forest model was utilized to predict lightning intensity. This model integrated key meteorological variables such as relative humidity and temperature at 2 meters, achieving an accuracy of 79%. The Random Forest model's performance demonstrates its effectiveness in real-time lightning intensity prediction, providing valuable insights into lightning behavior.

## 6. Conclusion and Implications

Lightning Detection Sensors (LDS) are currently in use across India, capable of detecting lightning events with over 98% accuracy within a 300 km radius for each sensor. While LDS offers significant support, but for the nations where there is limited or no LDS coverage, the incorporation of Artificial Intelligence and Machine Learning techniques could help in generating an early warning system.
In conclusion, the integration of MERRA-2 data with Lightning Detection System (LDS) data, alongside the application of advanced predictive models like Extra Trees with Bagging, demonstrates significant potential for enhancing lightning prediction with 89.51% accuracy within the targeted area spans from 86.68°E to 88.52°E in longitude to 24.70°N to 26.42°N in latitude which can be further enhanced by taking LDS data for the different regions across the globe to develop an early warning lightning prediction system. This combined approach leverages the strengths of both data sources and modelling techniques to produce a robust forecasting system with high prediction accuracy.

To enhance the reliability and effectiveness of lightning prediction models, incorporating insights into lightning intensity is crucial. The Random Forest classifier model, achieved 79% accuracy, underscores the importance of these insights for developing a powerful early warning system. By integrating MERRA-2 dataset with LDS dataset and implementation of this research, we can create highly effective early warning systems, especially in regions with sparse or no LDS coverage. This integrated approach represents a significant advancement in lightning prediction, leading to improved preparedness, more accurate alerts, and ultimately, greater public safety.

**Acknowledgment**
The authors express their sincere gratitude to Dr. Prakash Chauhan, Director of NRSC (National Remote Sensing Centre), and Dr. S.K. Srivastav, CGM of the RCs, NRSC, for their support and encouragement. Special thanks are extended to Shri G. Srinivasa Rao, Deputy Director of Earth & Climate Science Area (ECSA), Dr. M V Ramana (Group Director, ECSA), Dr. Alok Taori (Head, ECSSD) and team ECSA NRSC, Hyderabad, for their valuable guidance and for providing the datasets through ISRO geoportal Bhuvan. The authors are also grateful to people directly or indirectly involved helping in this research.

**References:**

Abdelkader B., R. Jaziri, R. and Bernard, G. (2019). *Deep extremely randomized trees*. In *Advances in Data Science and Artificial Intelligence* (pp. 843-850). Springer. https://doi.org/10.1007/978-3-030-36708-4_59

Akhter J., S. Roy and S. K. Midya (2024). Assessment of lightning climatology and trends over eastern India and its association with AOD and other meteorological parameters. *Journal of Earth System Science, 133*(36). https://doi.org/10.1007/s12040-023-02246-3 2021

Ali M. (2020). *PyCaret: An open source, low-code machine learning library in Python* (Version 1.0.0). PyCaret. https://www.pycaret.org

Bhuvan, NRSC, ISRO, Government of India. (n.d.). *Atmospheric lightning Essential Climate Variable*. https://bhuvan-pp1.nrsc.gov.in/lightning/

Chakraborty R., A. Chakraborty, G. Basha and M. V. Ratnam (2021). Lightning occurrences and intensity over the Indian region: Long-term trends and future projections. *Atmospheric Chemistry and Physics, 21*, 11161–11177. https://doi.org/10.5194/acp-21-11161-2021

GMAO M. (2015). *Modern-Era Retrospective analysis for Research and Applications, Version 2*. Greenbelt: Goddard Space Flight Center Distributed Active Archive Center (GSFC DAAC). doi:10.5067/VJAFPLI1CSIV

Han Y., H. Luo, Y. Wu (2021). Cloud ice fraction governs lightning rate at a global scale. *Communications Earth & Environment, 2*, 157. https://doi.org/10.1038/s43247-021-00233-4

Islam R. (2022) For lightning-prone communities in Bangladesh, new warning system may not be enough: Mongabay; [Available from: https://news.mongabay.com/2022/09/for-lightning-prone-communities-in-bangladesh-new-warning-system-may-not-be-enough/

Laipelt L., A. L. Ruhoff, A. S. Fleischmann, R. H. B. Kayser, E. D. M. Kich, H. R. da Rocha and C. M. U. Neale (2020). Assessment of an Automated Calibration of the SEBAL Algorithm to Estimate Dry-Season Surface-Energy Partitioning in a Forest–Savanna Transition in Brazil. *Remote Sensing*, *12*(7), 1108. https://doi.org/10.3390/rs12071108

Mondal U., S. Sreelekshmi, S. Panda, A. Kumar, S. Das and D. Sharma (2023). Diurnal variations in lightning over India and three lightning hotspots: A climatological study. *Journal of Atmospheric and Solar-Terrestrial Physics, 252*. doi:10.1016/j.jastp.2023.106149

Paul A. and R. Maity (2023). Future projection of climate extremes across contiguous northeast India and Bangladesh. *Scientific Reports, 13*(15616 ). doi:10.1038/s41598-023-42360-2

Rocca J. (2019) *Ensemble methods: bagging, boosting and stacking* . Retrieved from Medium: Ensemble methods: bagging, boosting and stacking | by Joseph Rocca | Towards Data Science

Sattar A., S. Khan and S. Banerjee (2021). Climatic water balance for assessment of growing season in the eastern Indian state of Bihar. *MAUSAM, 70*(3), 569-580. doi:10.54302/mausam.v70i3.269

Shankar A., A. Kumar and A. Sinha (2024). Incident of lightning-related casualties in Bihar, India: An analysis and vulnerability assessment. *J Earth Syst Sci, 133*(73 ). doi:10.1007/s12040-024-02277-4

Singh L.. (2022). A Practical Application of Chi-square Test in Hypothesis Testing. 1. 17-25.

Taori A., A. Suryavanshi, S. Pawar and M. V. R. Seshasai. (2022) "Establishment of lightning detection sensors network in India: generation of essential climate variable and characterization of cloud-to-ground lightning occurrences." *Natural Hazards* (2022): 1-14.

Venkatesh D., A. Taori, A. Suryavanshi, K. S. Rao, M. K M. Haridas, R.V. Bothale, and P. Chauhan (2023) "Comparison of ground-based lightning detection network data with WRF-Elec forecasting estimates over India–Initial results." *Remote Sensing Letters* 14, no. 10 (2023): 1009-1020.

Zhang H., Y. Deng, Y. Wang, L. Lan, X. Wen, C. Fang and J. Xu (2024). Extraction of factors strongly correlated with lightning activity based on remote sensing information. *Remote Sensing, 16*(11), 1921. https://doi.org/10.3390/rs16111921