“Spatiotemporal Analysis of Cloud Cover and Machine Learning-Based Cloud Classification Using INSAT-3DS”

SUMMER INTERNSHIP PROJECT

SUBMITTED TO

National Remote Sensing Centre (NRSC) Outreach Facility, Hyderabad

By
Harsh Raj Sahu
Student ID: 2022CSB107

Computer Science and Technology
Indian Institute of Engineering Science And Technology, Shibpur

Under The Supervision

Of

Mrs. Shivali Verma
वैज्ञानिक/अभियंता – *'एसई', Sci/Eng - 'SE',*
पृथ्वी एवं जलवायु विज्ञान क्षेत्र/*Earth and Climate Sciences Area*
राष्ट्रीय सुदूर संवेदन केन्द्र *(एन आर एस सी)/National Remote Sensing Centre (NRSC)*

Declaration

I hereby declare that the project report titled “Spatiotemporal Analysis of Cloud Cover and Machine Learning-Based Cloud Classification Using INSAT-3DS” is a true and genuine account of my original
research work. This project was undertaken as an academic endeavour under the esteemed guidance of the designated Guide . It represents a thorough and meticulous compilation of my insights, research findings, and analyses. I affirm that this work is the result of my independent effort except where explicitly stated otherwise. All Sources of information and data used in this report have been duly acknowledged , and appropriate references have been provided by academic standards and ethical guidelines.

I understand the significance of academic integrity and have adhered to all prescribed norms to ensure authenticity and credibility of the report.

Harsh Raj Sahu
Bachelor of Technology in Computer Science and Technology
Indian Institute Of Engineering Science and Technology , Shibpur

Acknowledgement

I would like to express my deepest gratitude to all those who have supported and contributed to the completion of this research project. My sincere thanks to Mrs **Shivali Verma**, whose guidance, constructive feedback, and constant encouragement, have been invaluable throughout this process.

I am also deeply grateful to my senior **Rohan Dutta** and the Head of the Computer Science Department Dr. **Apurba Sarka**r for helping me out in the starting phase. I extend my appreciation to National Remote Sensing Centre (NRSC), Indian Space Research Organisation (ISRO) for providing the necessary resources and a conductive environment for research.

I am particularly thankful to my family and friends for their unwavering support, patience, and understanding during the challenging phases of my research. Lastly, I would like to acknowledge anyone who provided moral support and motivation, helping me to preserve and complete this project.

Closing with Regards,

Harsh Raj Sahu

Project Intern
National Remote Sensing Centre (NRSC)
Indian Space Research Organisation
(ISRO)

Abstract

Clouds significantly impact the Earth's radiation budget and climate system, and their properties are influenced by complex interactions with atmospheric aerosols. This study investigates the spatiotemporal variation and classification of cloud types over the Indian region using geostationary satellite data from INSAT-3D, focusing on the period from January to May 2025. The research emphasises the utility of high-resolution, half-hourly satellite imagery in analysing cloud fraction, spatial analysis , and cloud types across four major domains: Arabian Sea, Bay of Bengal, Indian Mainland, and Indian Ocean. The satellite-derived multi-channel data were pre-processed and transformed into relevant features, including brightness temperature (BT) and brightness temperature differences (BTDs), which formed the basis for both spatial mapping and machine learning-based classification.

The study begins with time series and seasonal analysis to highlight regional variations in cloud fractions and types, revealing patterns that align with meteorological and aerosol-related processes in the Indian subcontinent. Subsequently, supervised machine learning models—including XGBoost, Random Forest, Logistic Regression, SVM, Decision Trees, KNN, and a multilayer perceptron-based neural network—were developed to classify cloud types using spectral and geophysical variables. Feature engineering, missing data treatment, and SMOTE-based oversampling were applied to enhance model robustness. Among all models, Random Forest and neural networks demonstrated superior classification accuracy and F1 scores. These results illustrate the potential of AI-driven approaches to improve cloud detection, support atmospheric modelling, and contribute to more accurate climate prediction tools by leveraging remote sensing data.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.Introduction

Clouds are one of the most dynamic components of the Earth's atmosphere and play a critical role in regulating weather patterns and global climate. Their ability to reflect incoming solar radiation and trap outgoing terrestrial radiation makes cloud properties, such as Spatial Analysis and Cloud Cover , essential climate variables. Understanding how these properties vary across different regions and cloud types is key to improving the accuracy of climate models and weather predictions.

In recent decades, increasing attention has been paid to how atmospheric aerosols—tiny particles from natural and anthropogenic sources—interact with clouds. Aerosols can influence cloud formation, lifetime, and radiative behaviour through several mechanisms, including the cloud albedo effect (brightening), lifetime effect (precipitation suppression), and semi-direct effects. These interactions, collectively termed aerosol–cloud interactions (ACIs), are especially significant over regions like the Arabian Sea and northern Indian Ocean, which experience high aerosol loading due to seasonal emissions from South and Southeast Asia.

Despite their importance, the representation of cloud properties and aerosol–cloud interactions in models remains uncertain. A major challenge lies in the lack of high-resolution, continuous satellite observations over the Indian region. While polar-orbiting satellites provide valuable data, their twice-daily coverage is insufficient to capture the full diurnal variability of cloud processes. This is where geostationary satellites like INSAT-3D and INSAT-3DR offer a major advantage, providing half-hourly observations with enhanced spatial and spectral resolution, enabling detailed monitoring of cloud dynamics.

In this study, we utilise multi-channel INSAT-3D data processed through a custom-developed algorithm in Python to detect cloud pixels and retrieve cloud top temperatures. We analyse cloud fraction over four key domains—Arabian Sea, Indian Mainland, Bay of Bengal, and Indian Ocean—during the winter and pre-monsoon seasons (January to May). We further conduct a spatial analysis of different cloud types, namely high-level thick clouds, low-level thick clouds, semi-transparent cirrus clouds, and partial clouds, separating the trends for early (January–February) and later (March–May) months.

In the next phase of the study, we develop a machine learning–based cloud classification model using five different algorithms, including deep learning. We evaluate the classification efficiency and accuracy of each model to identify the most effective method for cloud type identification. The outcomes of this study not only provide a deeper understanding of seasonal cloud behaviour over the Indian region but also contribute towards improving cloud representation in climate modelling and satellite-based weather forecasting systems.

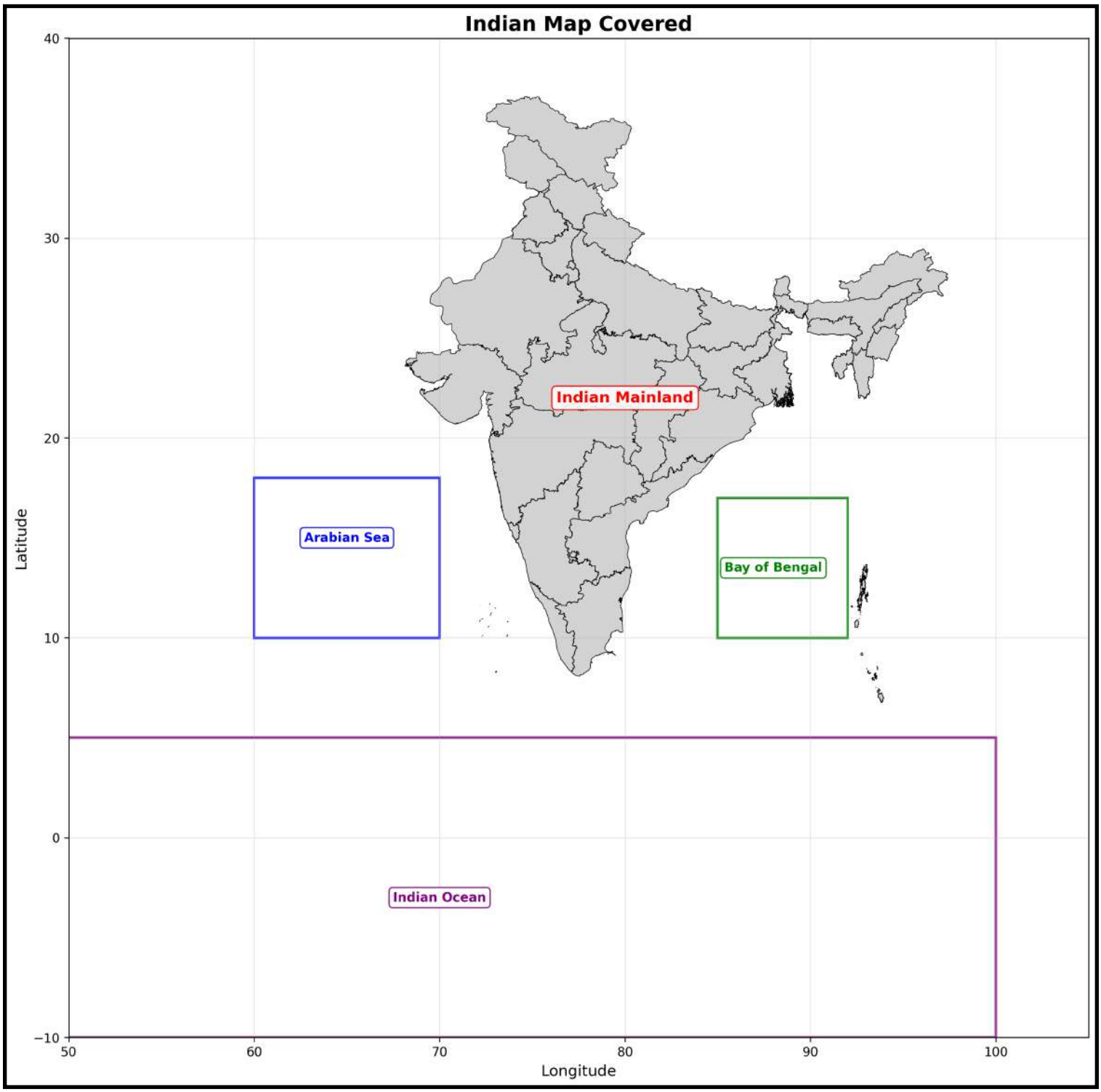


Figure 1.1 - Study Area

2. Data

Imager onboard the Indian geostationary satellite, INSAT-3D provides observations over theIndian region at a temporal interval of 30 min (at HH:00 and HH:30) with visible (VIS), short-wave infrared (SWIR), mid-wave infrared (MIR), water vapour (WV) and thermal infrared (TIR1 and TIR2) channels .Specifications of these channels are given in Table 1.1. Present study uses level 1C,Asia sector product (ASIA_MER_L1C) from INSAT-3D Imager over India and surrounding regions bounded by 44.5◦E–105.5◦E and 10◦S–45.5◦N with spatial resolution of 4 km, which is available through the ISRO web portal, MOSDAC website(https://www.mosdac.gov.in). The present study makes use of VIS, MIR, WV, TIR1 and TIR2 channels to identify cloud pixels and retrieve CTT for the period of September, 2014 to February, 2017. Spatial resolutions of VIS and WV channels are 1 km and 8 km respectively, whereas those of MIR and TIR channels are 4 km. In order to maintain uniformity, measurements from VIS and WV channels are also provided at 4 km spatial resolution to match with that of MIR and TIR channels.

Channels	**Spectral Range** (μm)	**Central Wavelength** (μm)	**Resolution (Km)**
Visible (VIS)	0.55-0.75	0.65	1.0
Short-wave Infrared (SWIR)	1.55-1.70	1.62	4.0
Mid-wave Infrared (MIR)	3.80-4.00	3.9	4.0
Water Vapour(WV)	6.50-7.10	6.8	8.0
Thermal Infrared I (TIR I)	10.3-11.3	10.8	4.0
Thermal Infrared II (TIR II)	11.5-12.5	12.0	4.0

Table 1.1

We converted this channels in our format for the dataset in Brightness Temperature as BT 3.9 , BT 6.7, BT 10.8, BT 12.0 and Brightness Temperature Differences as BTD 3.9-10.8 , BTD 12-10.8, BTD 6.7-10.8, BTD TIR1-TIR2 , BTD 3 Channel, Wind Gust Estimate , Longitude , Latitude , Timestamp , Date , Cloud type From INS3S 2025 files collected from NRSC Bhuvan site .

The dataset used in this study has been sourced from the National Remote Sensing Centre (NRSC), ISRO, specifically via the Bhuvan platform, which provides Level 3 geophysical parameters derived from INSAT-3D/3DR satellite imagery. The downloaded data consists of NetCDF (.nc) files at a spatial resolution of 4 km, structured across multiple channels, cloud mask flags, and cloud type flags, each encoded in a two-dimensional georeferenced grid.

Two primary variables were extracted and utilised:

a. Cloud Cover

This variable represents the presence and type of cloud cover for the 01:30 AM slot. The values are encoded as:

- `0`: No cloud
- `1`: Cloud detected
  This binary mask allows for simple cloud presence classification. A sample visualisation shows values varying across longitudes (X) and latitudes (Y), with the average cloud coverage per row indicating the spatial distribution of cloud presence.

b. Cloud Flag

This is a more granular classification variable, encoding cloud types:

- `1`: High-level thick clouds
- `2`: Low-level thick clouds
- `3`: Semi-transparent cirrus clouds
- `4`: Partial clouds

These values help distinguish between different cloud categories based on their optical thickness and altitude, essential for downstream classification and climate modelling. As shown in the matrix view, the average flag value per location gives an insight into cloud density and type dominance.

The core dataset used in this research was derived from Level 3 geophysical products—specifically, cloud cover and cloud flag variables extracted from NetCDF (.nc) files available via NRSC Bhuvan's INSAT Climate Data portal.The dataset thus generated is highly suitable for training machine learning models for cloud classification tasks. The regular temporal acquisition of data from INSAT-3D ensures temporal generalisability, while the standardised spatial resolution and multi-spectral richness of the inputs provide strong discriminative features for classification. Such a dataset is valuable for numerous applications, including automated weather monitoring, atmospheric research, and satellite-based climate modelling. Furthermore it paves the way for integrating satellite imagery with

Machine learning models that can detect and predict cloud dynamics with minimal manual intervention.

The resultant cloud classification dataset serves multiple purposes:

- **Weather forecasting** and short-term atmospheric analysis
- **Training deep learning models** (e.g., CNNs or Random Forests) to predict cloud type from spectral data
- **Cloud trend mapping** for climate change models
- **Spatiotemporal analysis** of cloud types over Indian sub-regions

This type of dataset holds critical value in building AI-based meteorological system, helping automate cloud detection and classification with high accuracy using remote sensing data.

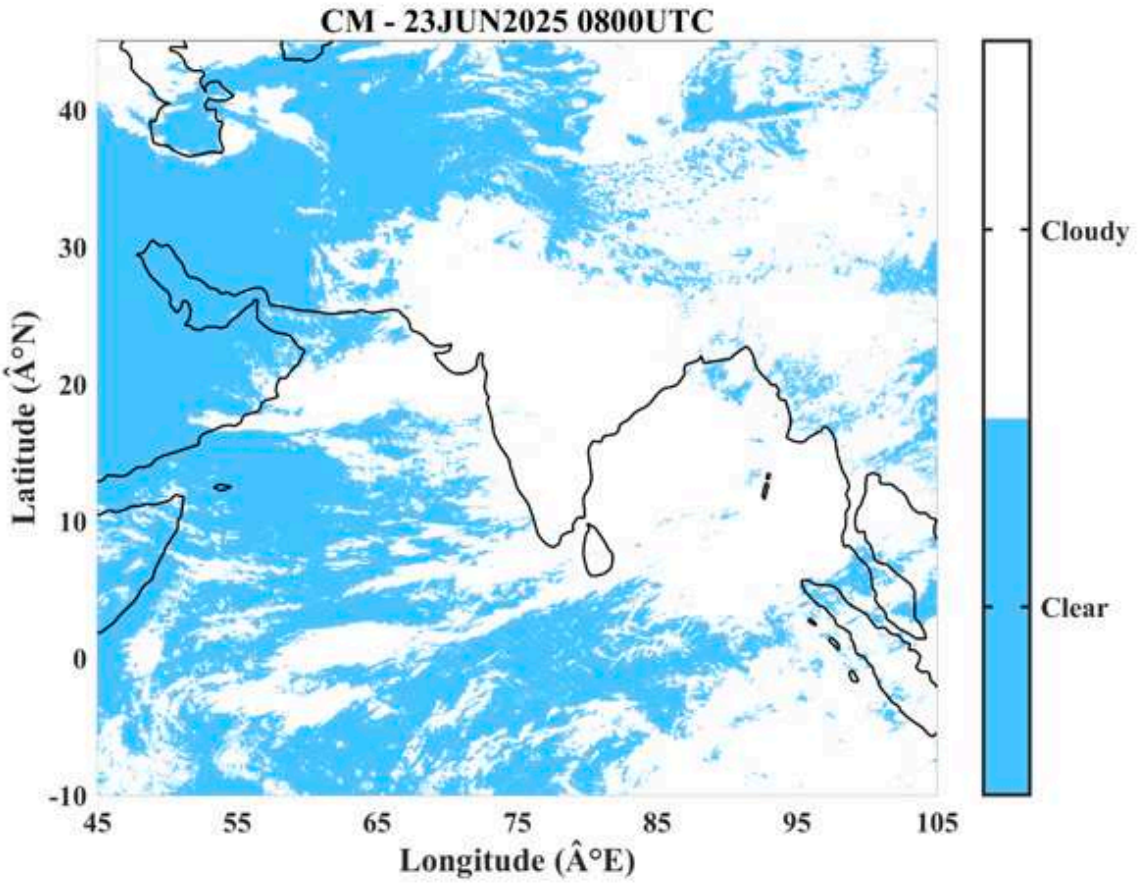


Figure 2.1 - Cloud over Coastal Area

3. Method and Analysis

This section presents the cloud classification and regional cloud distribution analysis conducted over four key domains: the Arabian Sea, Bay of Bengal, Indian Ocean, and Indian Mainland. The analysis spans from January to May 2025, using INSAT-3D derived cloud top temperature (CTT) and cloud classification outputs processed through a custom Python-based algorithm.

3.1 Time Series Analysis

Cloud fraction (CF)—defined as the proportion of the sky covered by clouds at a given time and place—is a fundamental metric for understanding atmospheric processes and evaluating cloud radiative effects. In this study, the cloud fraction for each classified cloud type was computed across four key geographic regions: Arabian Sea, Bay of Bengal, Indian Ocean, and Indian Mainland, using INSAT-3D satellite observations from January to May 2025.

3.1.1 Monthly Trends Across Regions

Each of the four study regions displayed unique cloud fraction characteristics, influenced by local meteorology, aerosol distribution, and underlying surface conditions:

- Over the **Arabian Sea**, cloud cover during January and February was dominated by **high-level thick clouds** and **semi-transparent cirrus clouds**, likely due to the influence of upper-level circulations and remote convective systems. As the season progressed into March–May, the fraction of **partial clouds** increased, indicating a breakdown of organised stratiform clouds into scattered convection or low optical depth formations, possibly impacted by rising aerosol concentrations.

- The **Bay of Bengal** showed a relatively high cloud fraction throughout the entire study period. Notably, **low-level thick clouds** were more prevalent in March and April, consistent with the onset of warm sea surface temperatures and increasing moisture convergence. The transition toward the pre-monsoon phase was marked by increased convective activity, often initiating along the coast and spreading inland.

- The **Indian Ocean** region, being farther south and less affected by continental emissions, demonstrated a more stable seasonal pattern. **Semi-transparent cirrus clouds** and **partial clouds** maintained moderate coverage levels throughout the study period, suggesting persistent mid- and upper-tropospheric dynamics. A slight increase in **partial clouds** in the latter months may reflect weakening large-scale subsidence or the intrusion of shallow convection.

- The **Indian Mainland** exhibited the most dynamic behaviour, with **low cloud fractions** in January and February due to drier and more stable atmospheric conditions. However, cloud fraction sharply increased from March onward, with a noticeable rise in **partial** and **low-level thick clouds**, especially over central and eastern India.

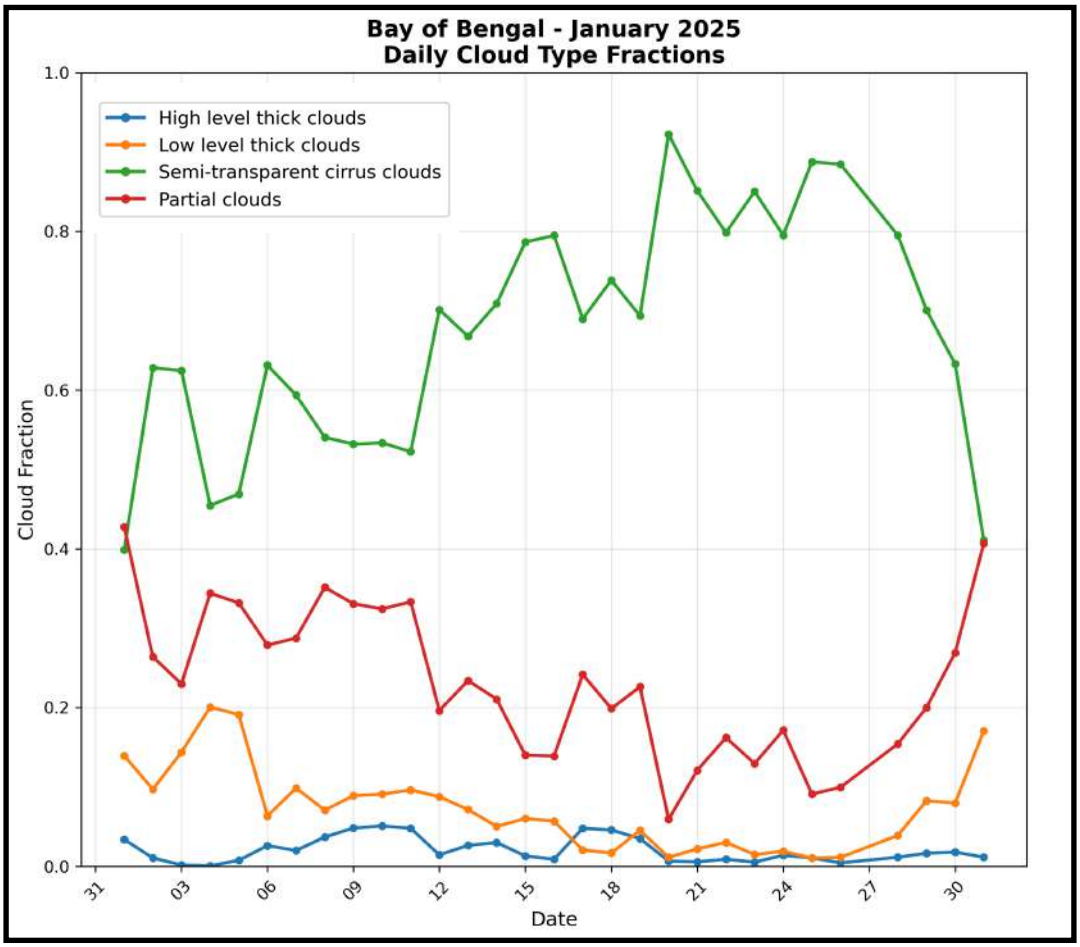


Figure 3.1

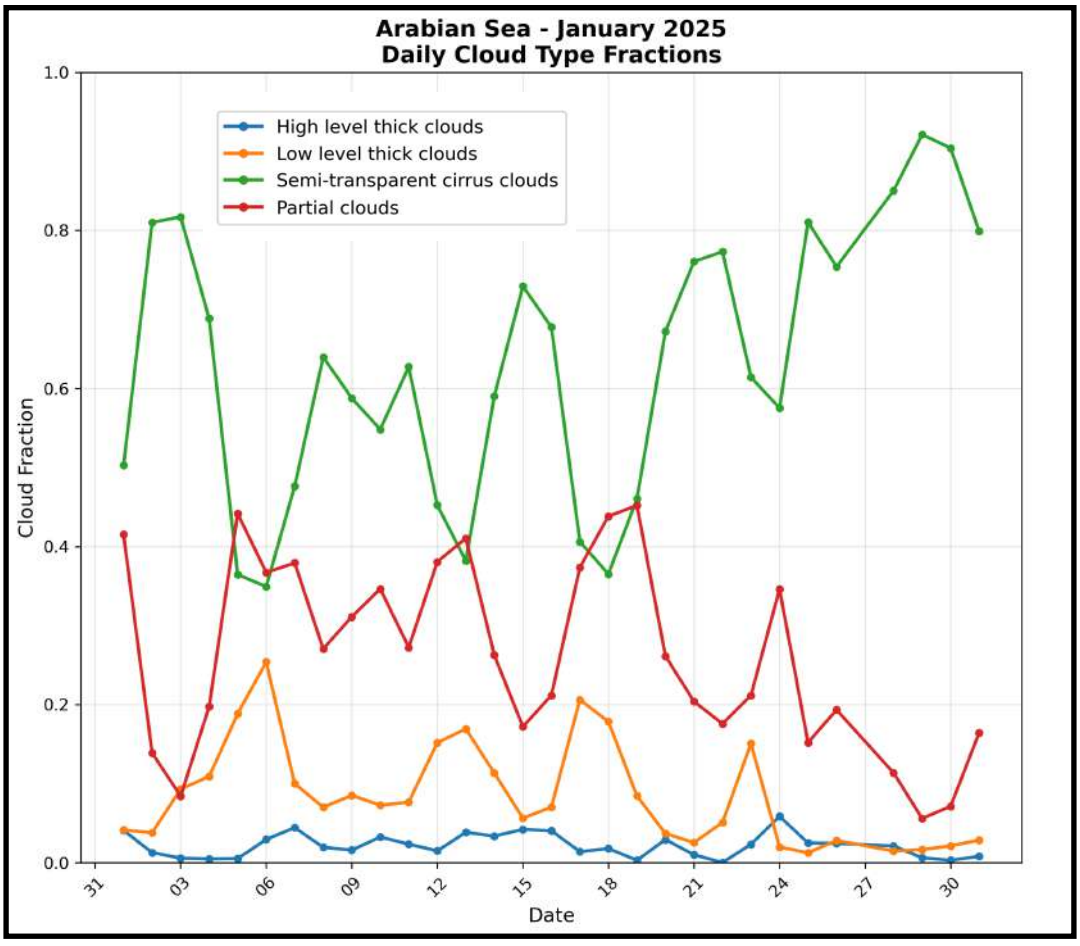


Figure 3.2

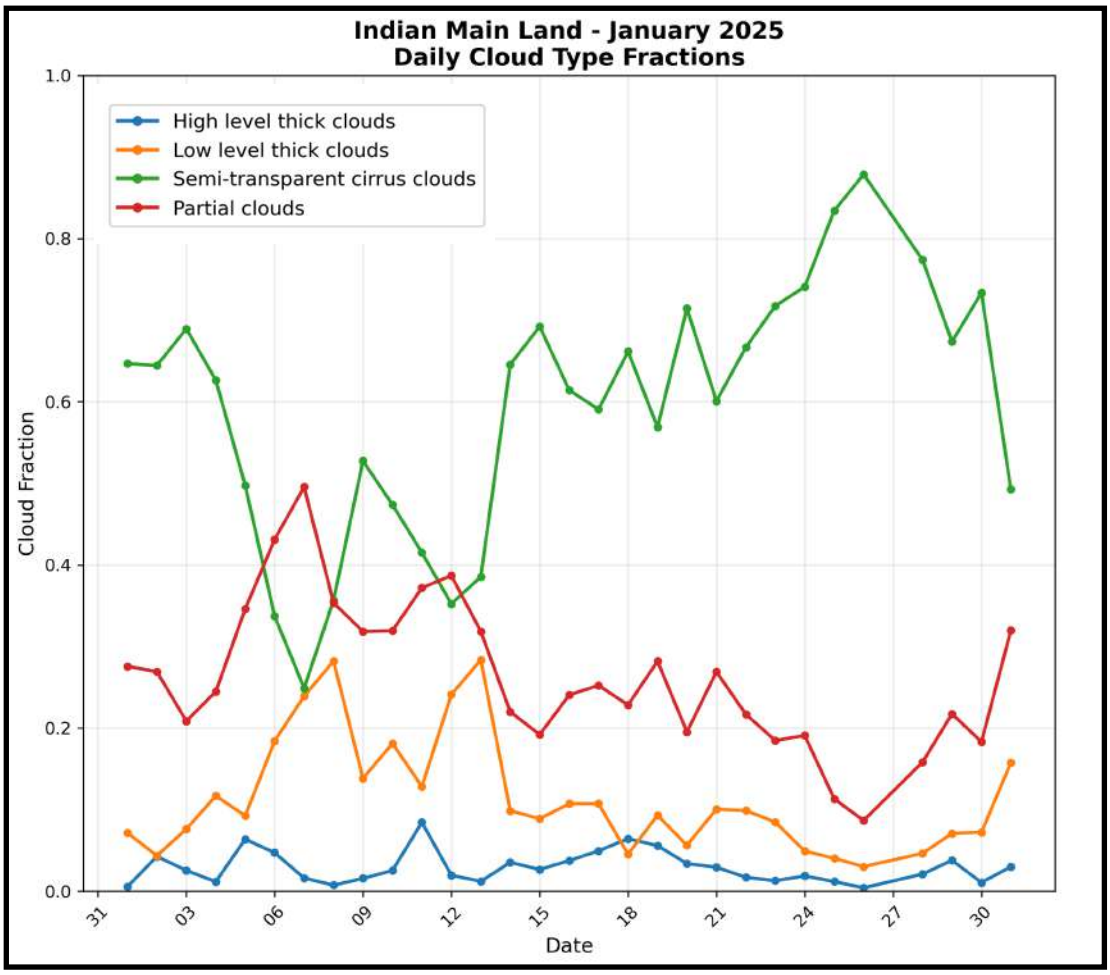


Figure 3.3

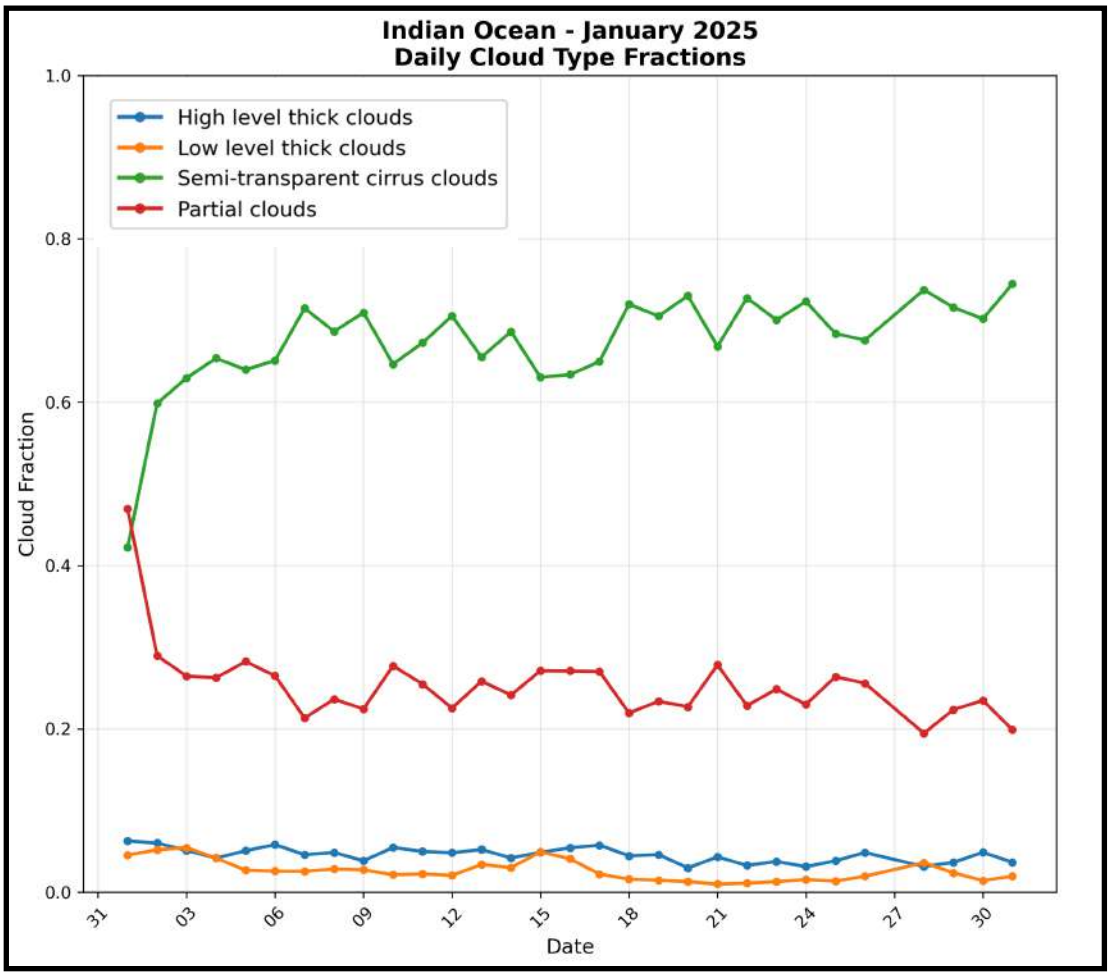


Figure 3.4

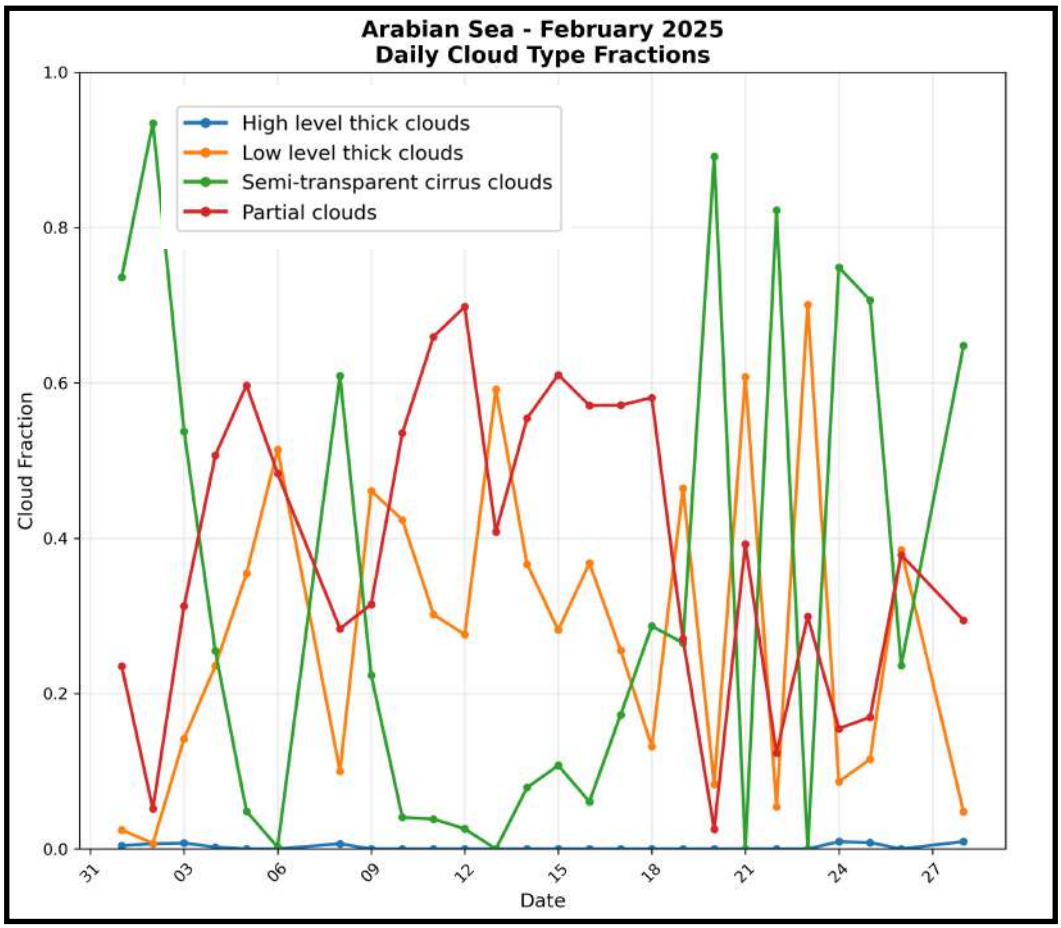


Figure 3.5

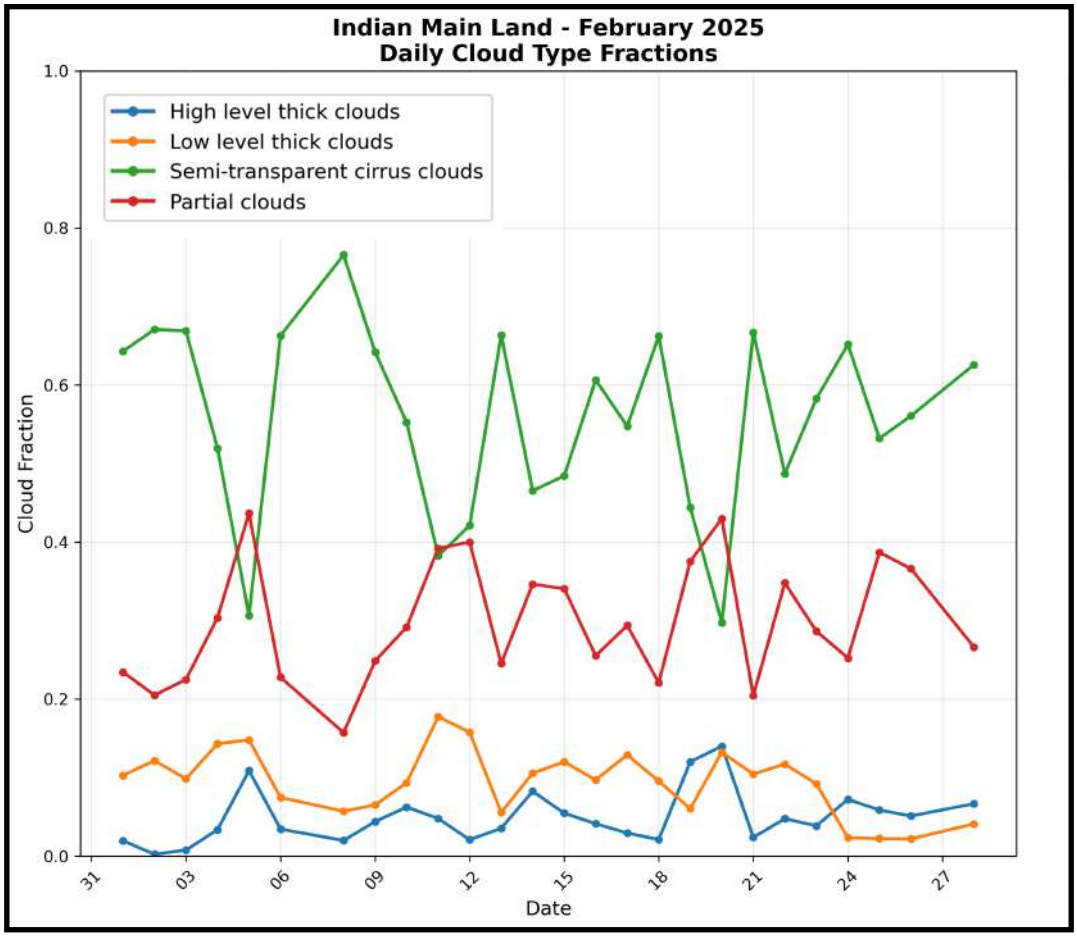


Figure 3.6

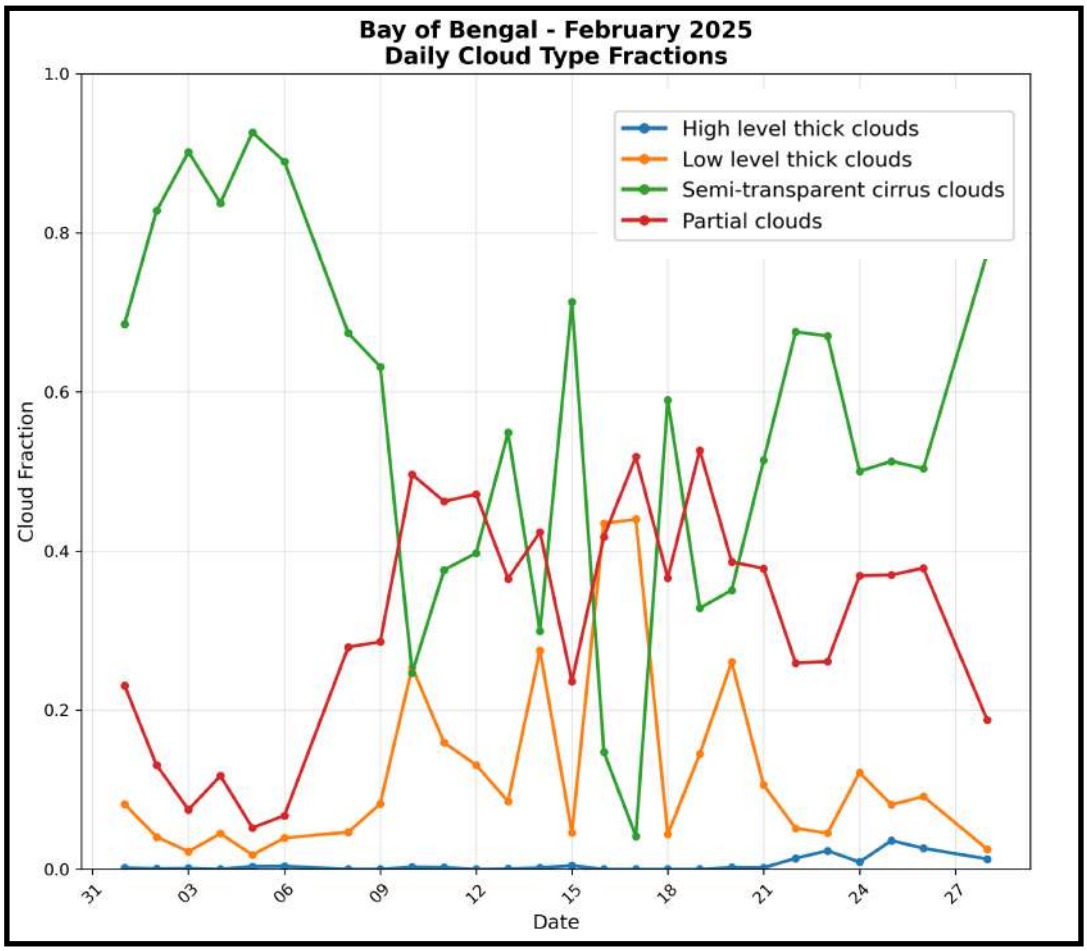


Figure 3.7

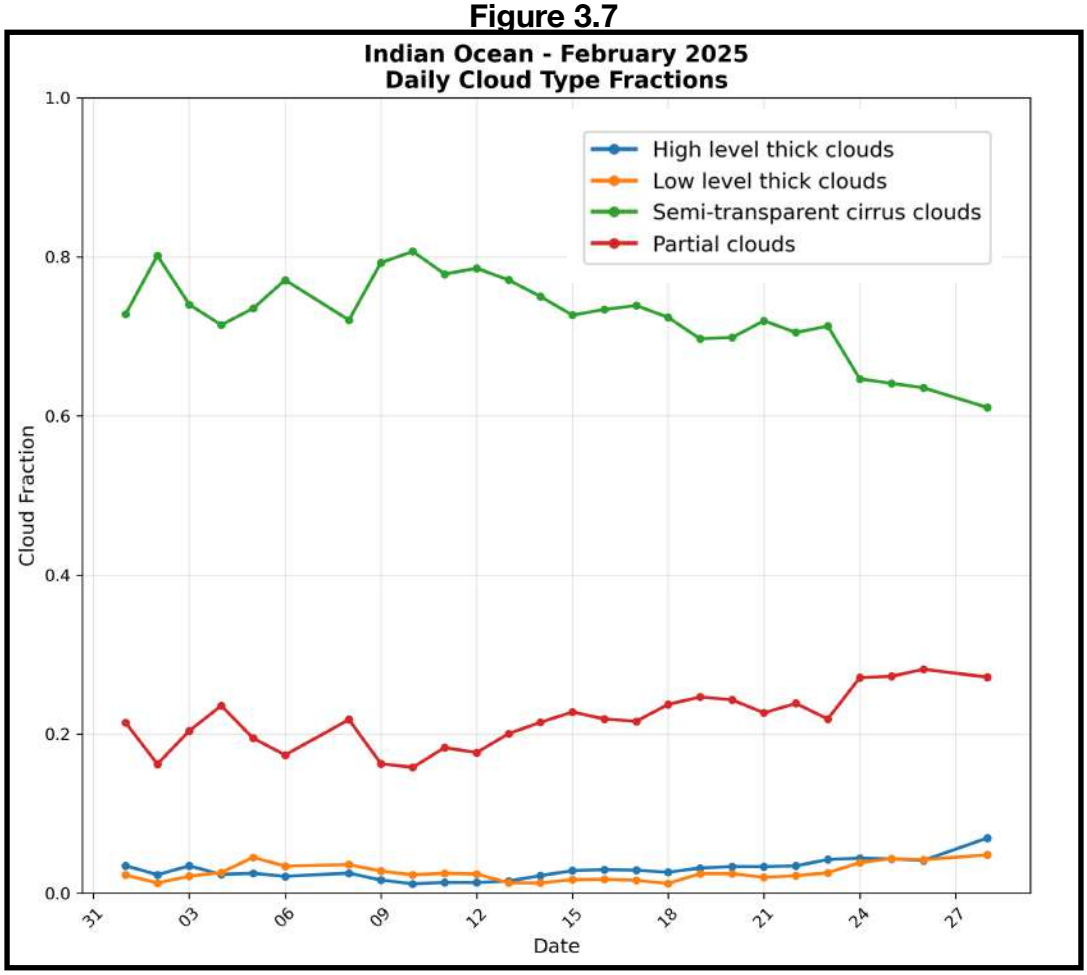


Figure 3.8

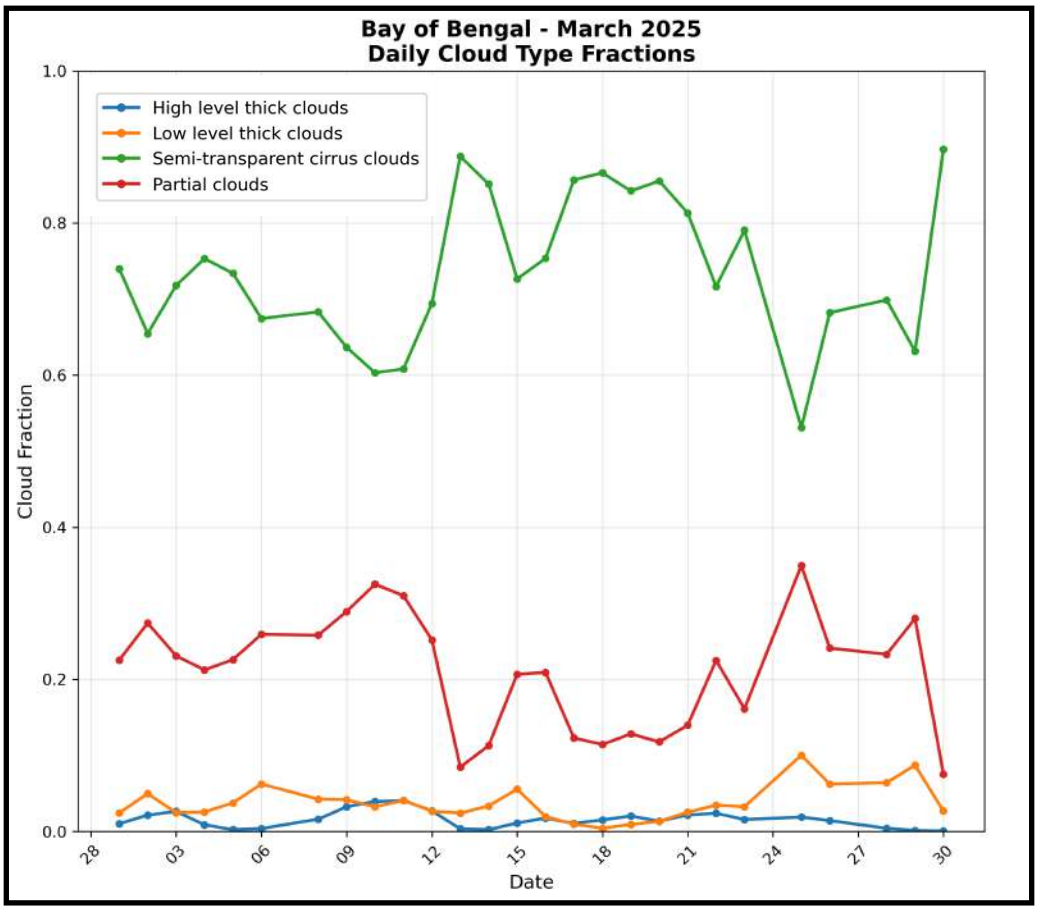


Figure 3.9

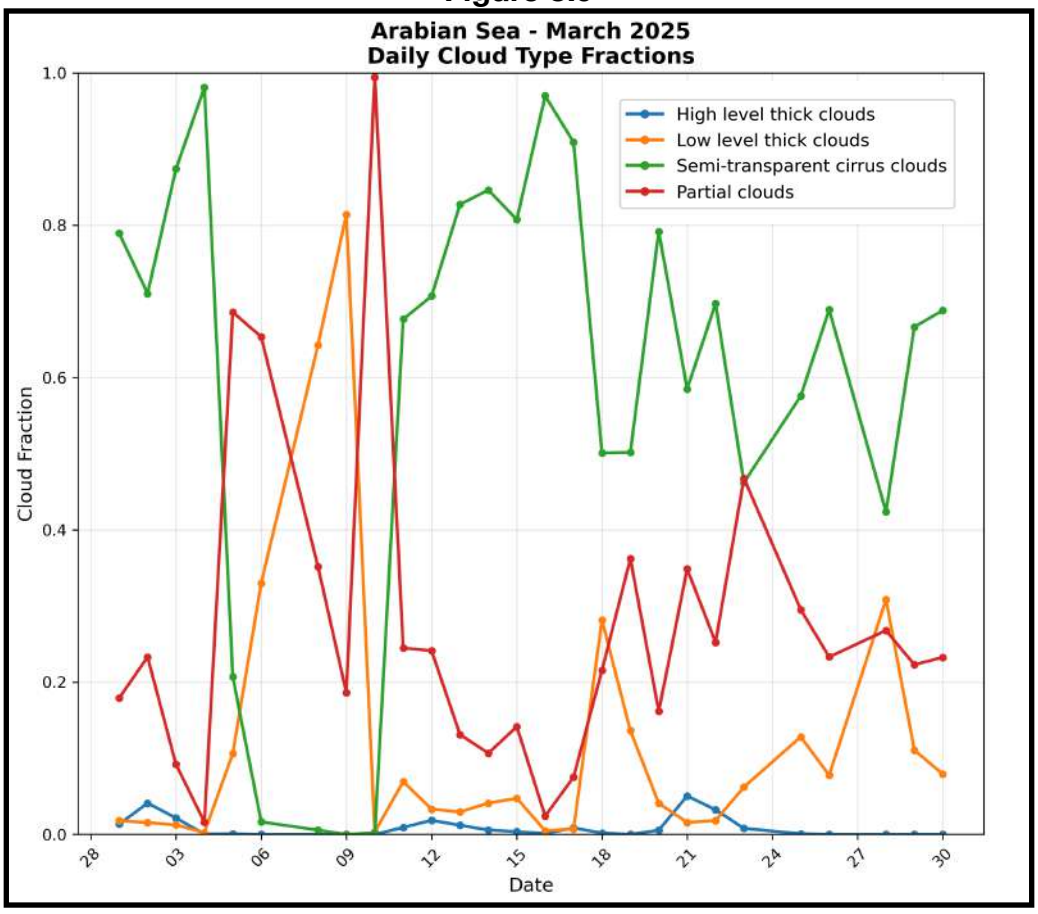


Figure 3.10

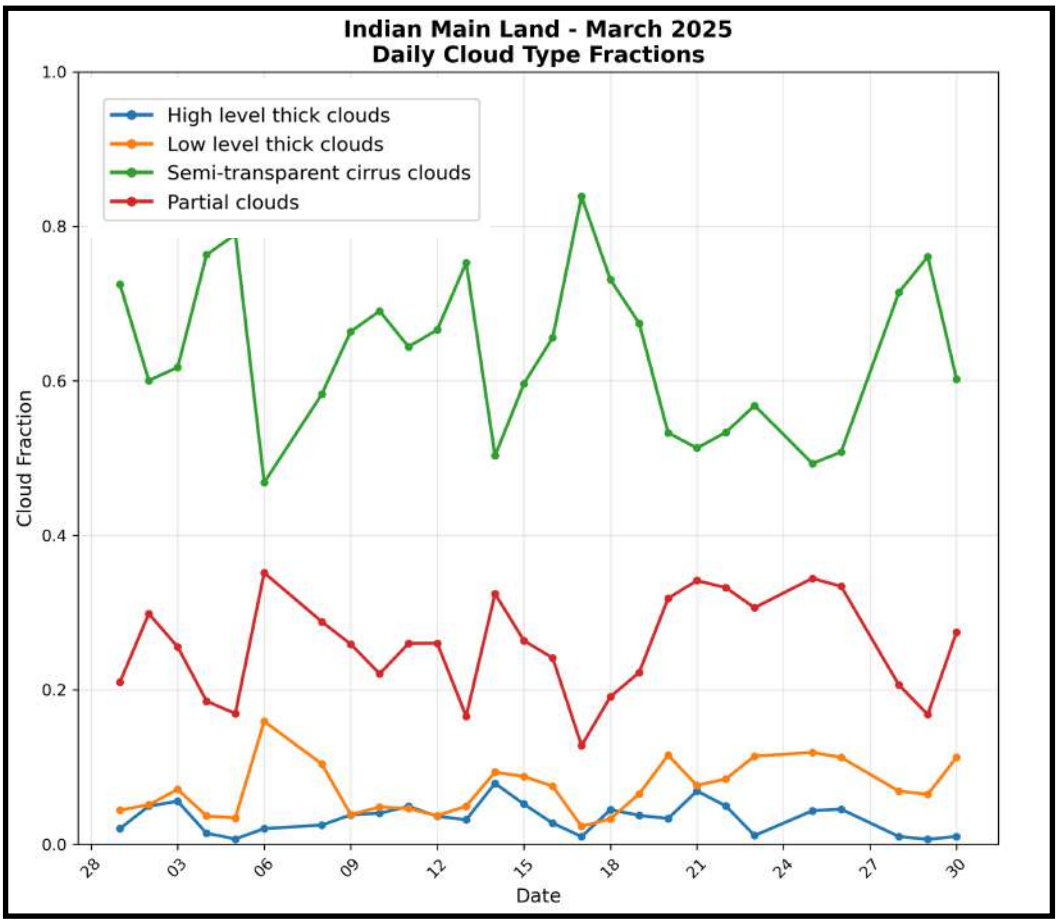


Figure 3.11

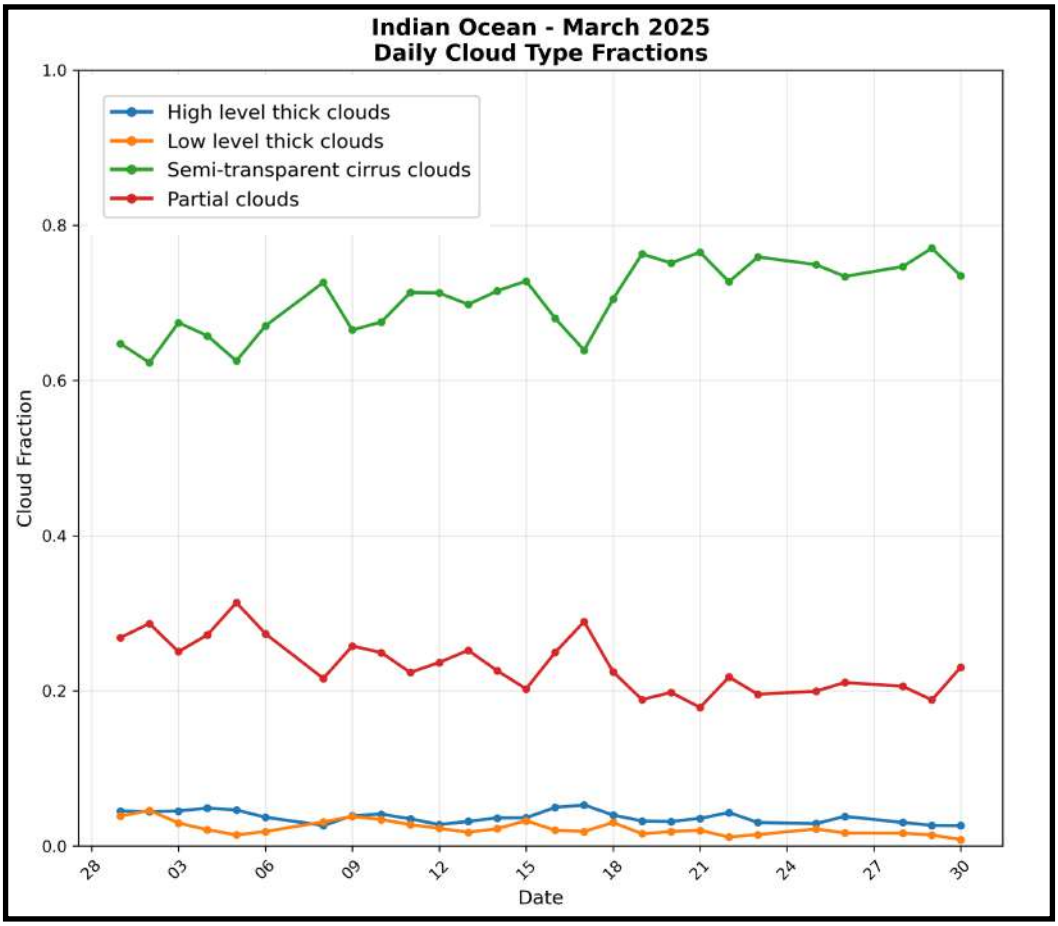


Figure 3.12

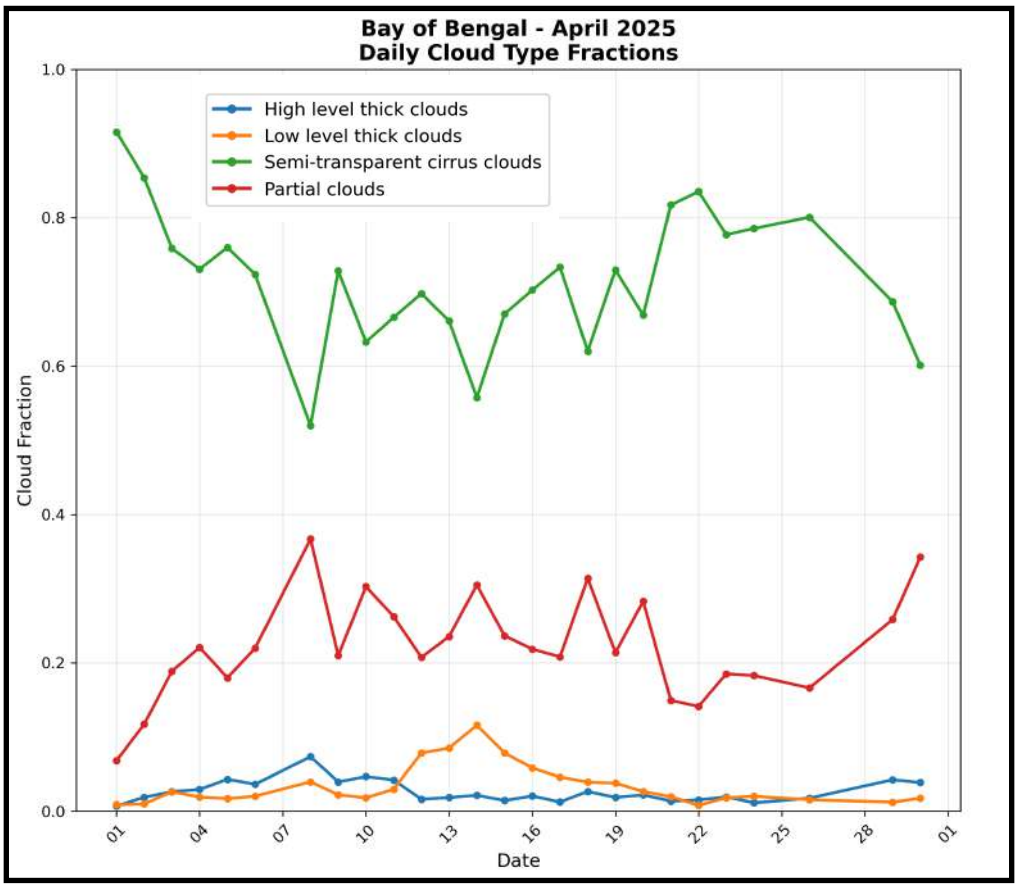


Figure 3.13

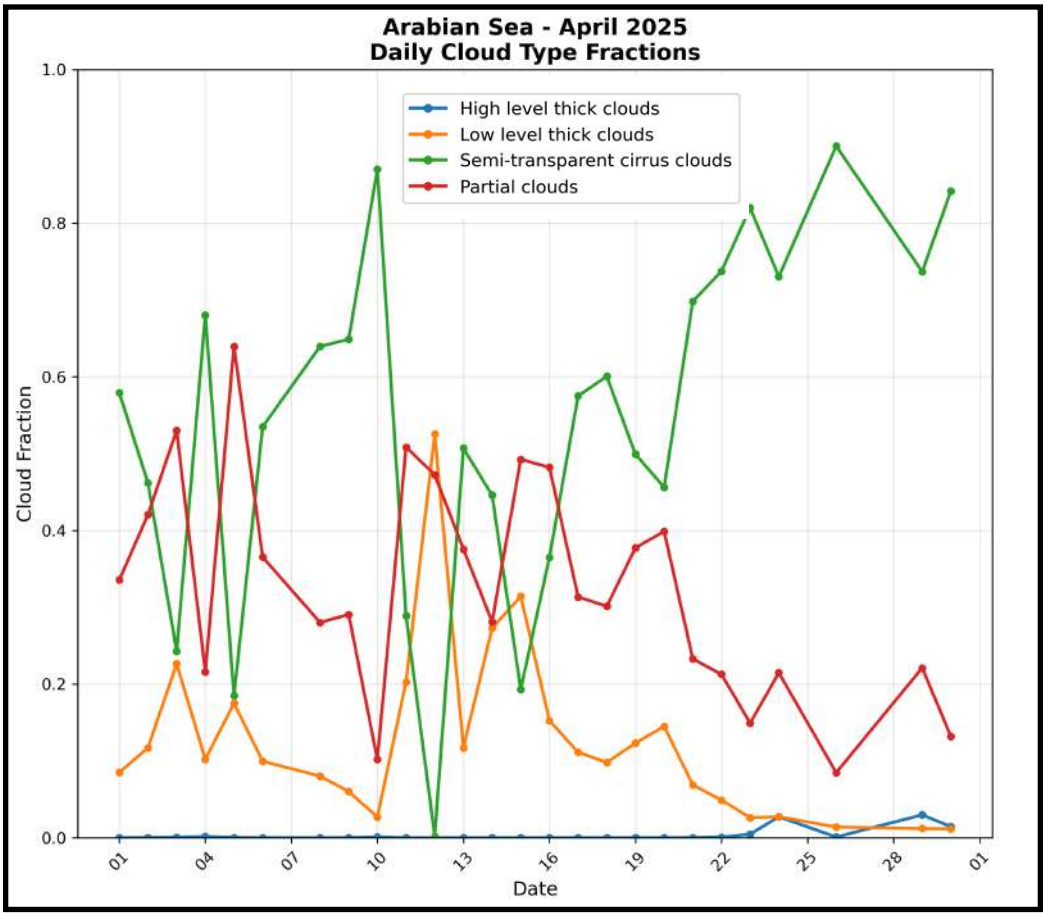


Figure 3.14

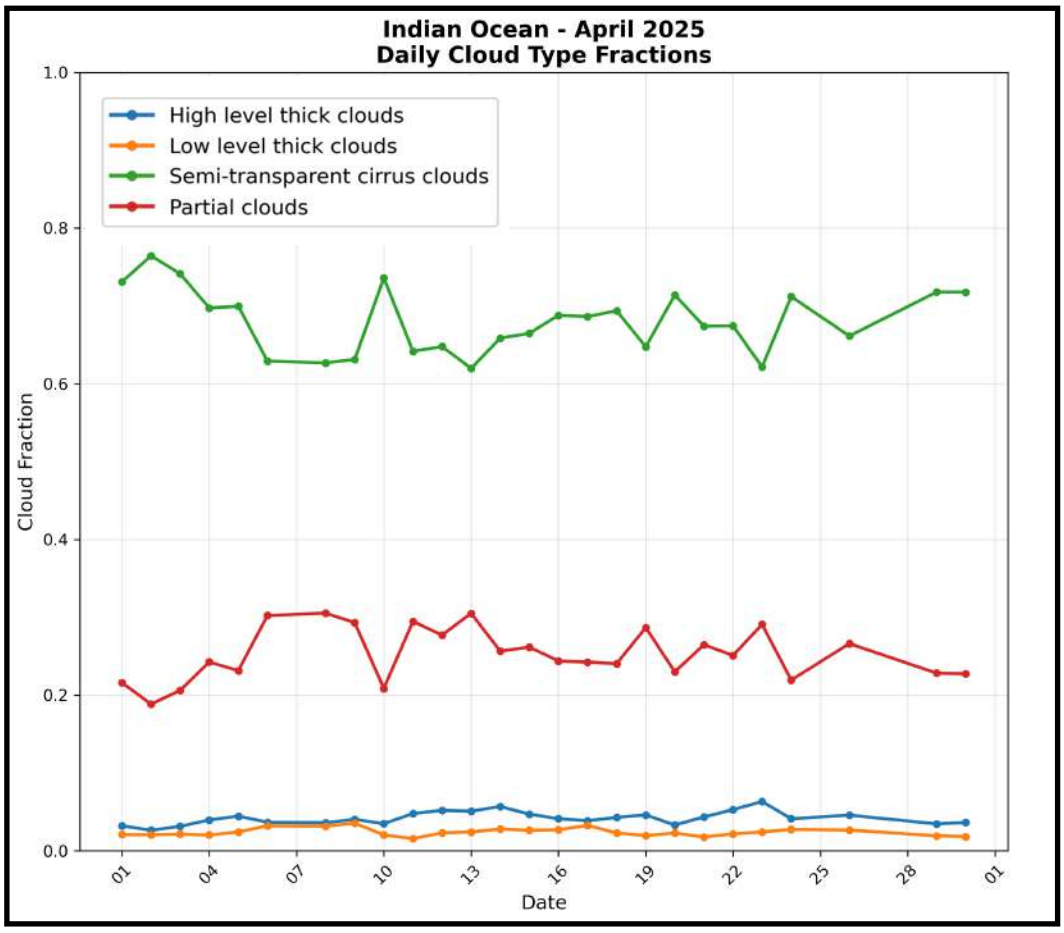


Figure 3.15

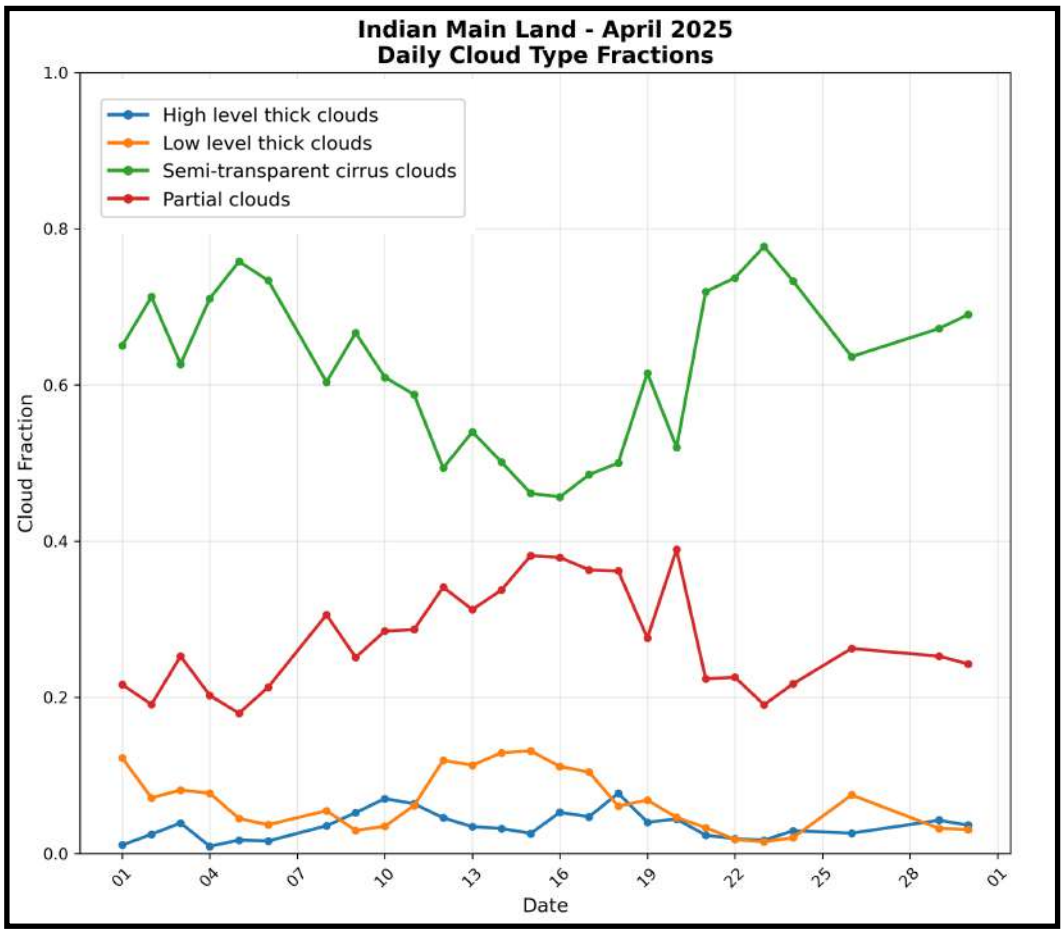


Figure 3.16

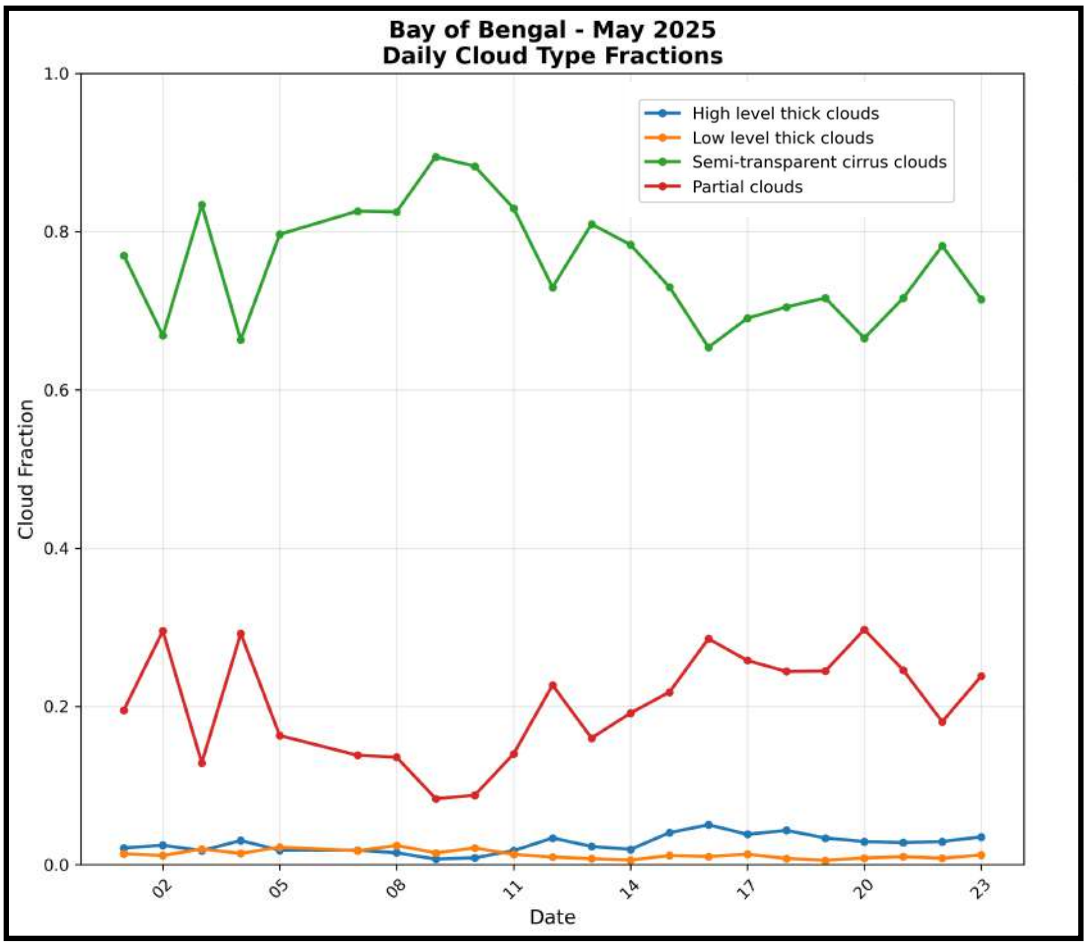


Figure 3.17

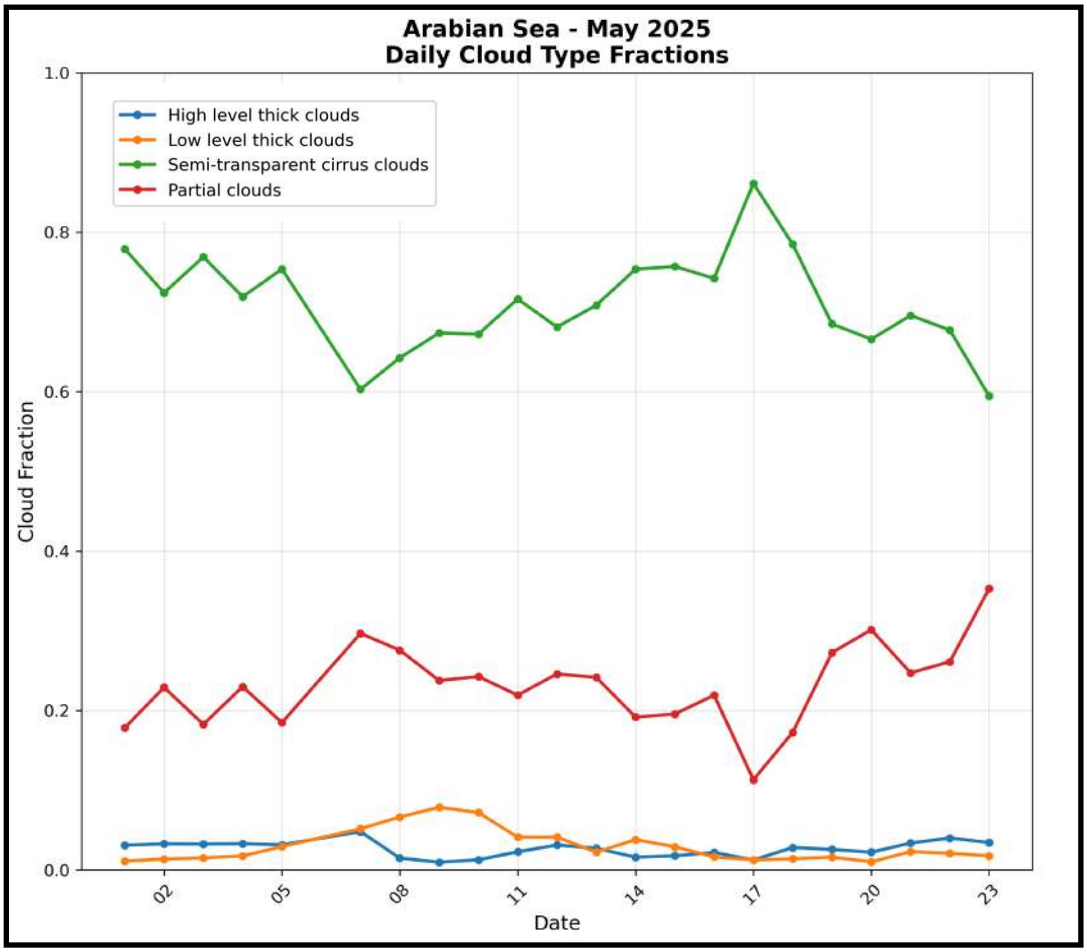


Figure 3.18

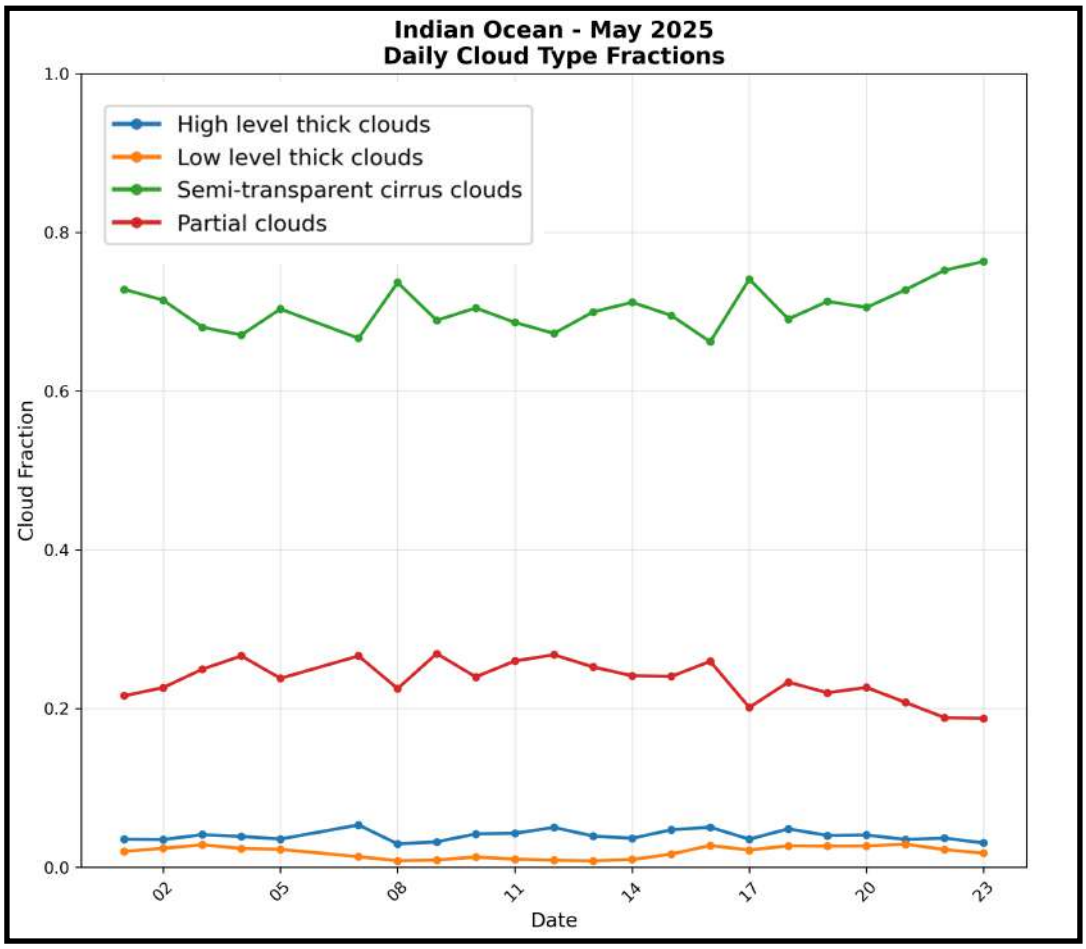


Figure 3.19

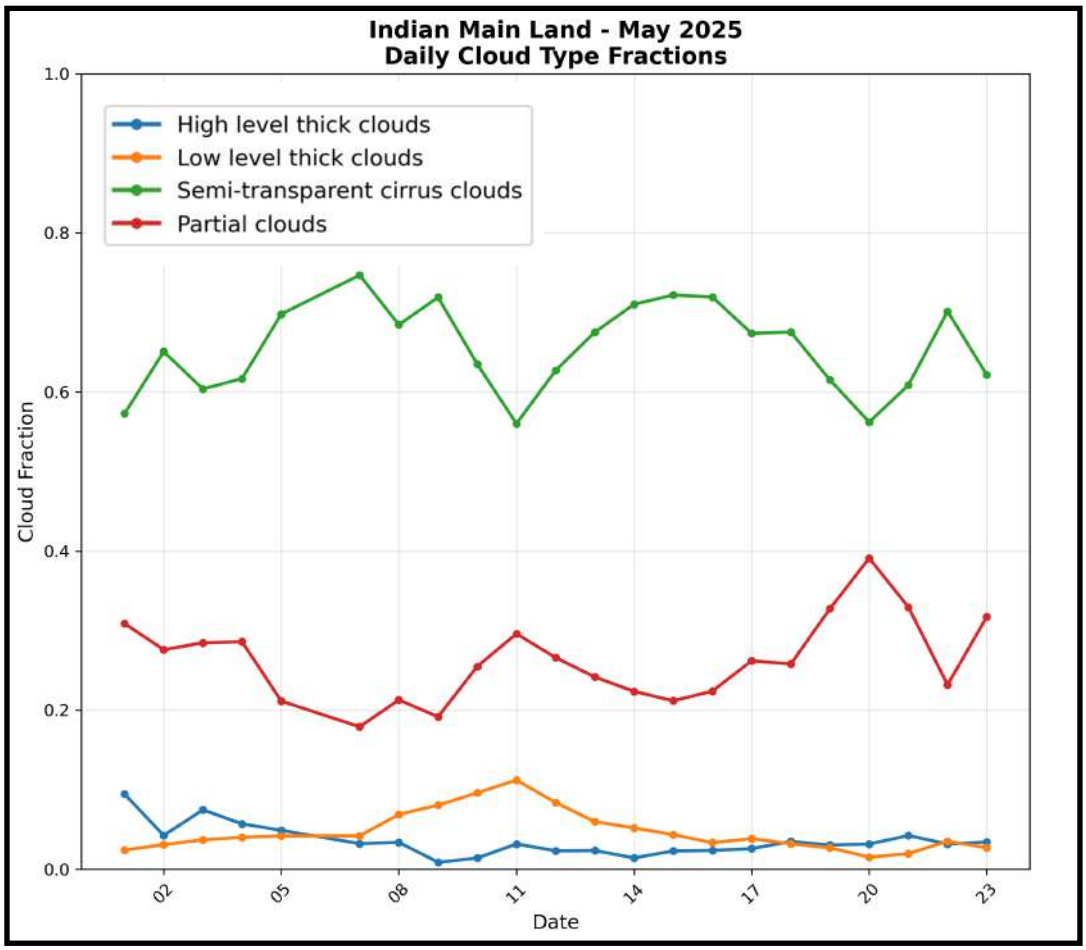


Figure 3.20

3.2 Spatial Classification by Season

Using the processed INSAT-3D data and cloud classification algorithm, each pixel within the study region was analysed and assigned a cloud type. The four cloud classes—**High-level thick clouds**, **Low-level thick clouds**, **Semi-transparent cirrus clouds**, and **Partial clouds**—were mapped for each period, allowing for a direct spatial comparison across time.

3.2.1 January–February (Winter Season)

During the winter months, the atmospheric conditions are characterised by stable air masses and lower sea surface temperatures, which contribute to stratiform cloud development, particularly over marine regions. The spatial maps for January–February reveal:

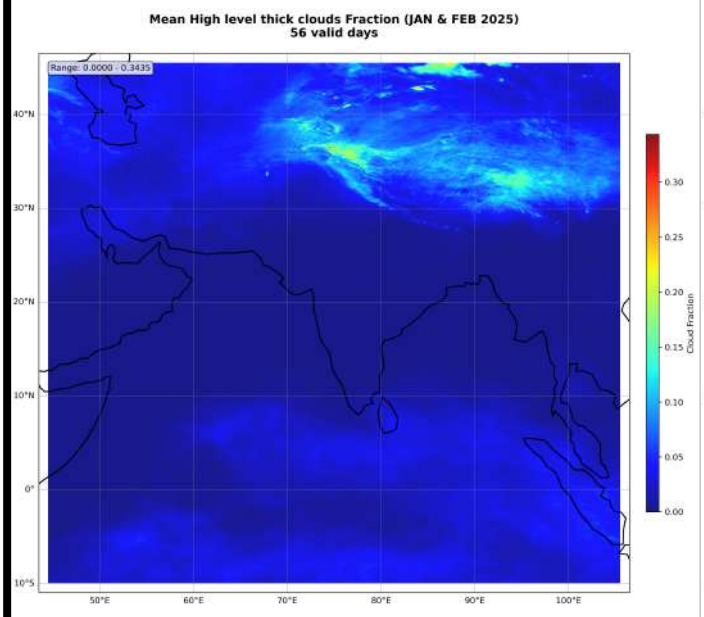
- A **widespread presence of high-level thick clouds** over the Arabian Sea and Bay of Bengal. These are likely influenced by upper-tropospheric humidity and enhanced convection near the Inter-Tropical Convergence Zone (ITCZ).
- **Semi-transparent cirrus clouds** are frequently observed stretching across central and southern parts of the Indian Ocean, typically associated with jet streams and outflow from deep convection.
- The **Indian Mainland** shows lower cloudiness overall, with scattered low-level thick clouds primarily over the eastern states and along the Western Ghats.
- **Partial clouds** are less dominant during this season, indicating more uniform and well-developed cloud systems due to stable meteorological conditions.

3.2.2 March–May (Spring Season)

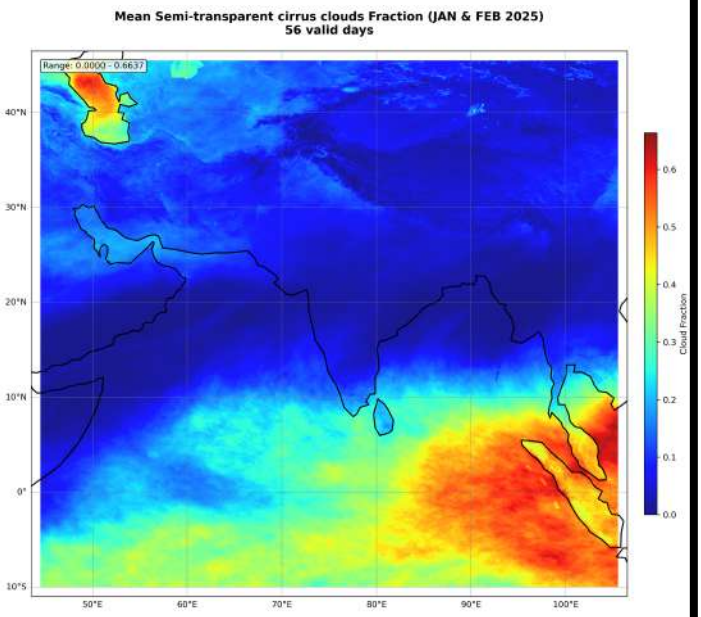
- **Partial clouds** become more prevalent over the Indian Mainland and adjoining oceanic areas. This reflects increased atmospheric instability and fragmented cloud development typical of pre-monsoon convective buildup.
- **Low-level thick clouds** appear more frequently along coastal regions and in the Bay of Bengal, potentially driven by increasing sea surface temperatures and local convergence.
- The **Arabian Sea** shows a reduction in high-level cloud dominance, with an increased presence of mixed cloud types, including cirrus and partial clouds. This may be attributed to absorbing aerosols from the Indo-Gangetic Plain altering cloud top dynamics.
- Over the **central Indian Ocean**, **semi-transparent cirrus clouds** continue to appear but are more localised compared to the winter period, suggesting reduced upper-level divergence.

January–February

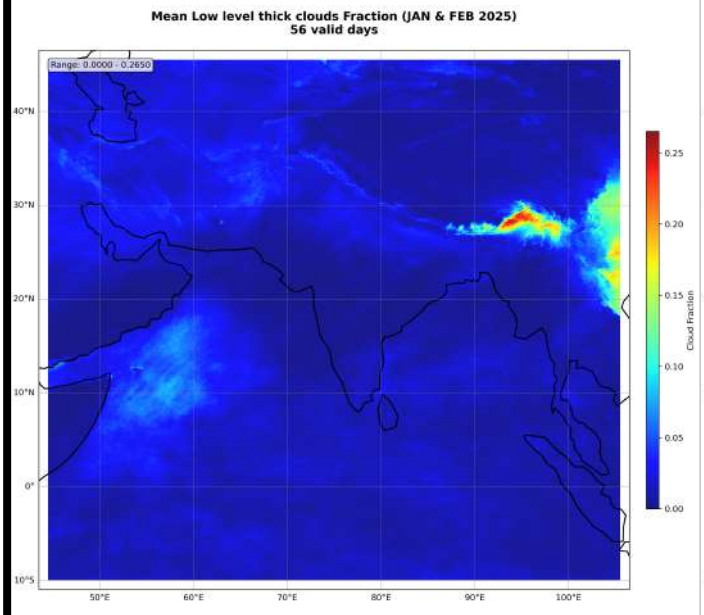
A . Spatial Analysis of High Level Thick Clouds



B . Spatial Analysis of Cirrus Clouds



C . Spatial Analysis of Low Level Thick Clouds



D. Spatial Analysis of Partial Clouds

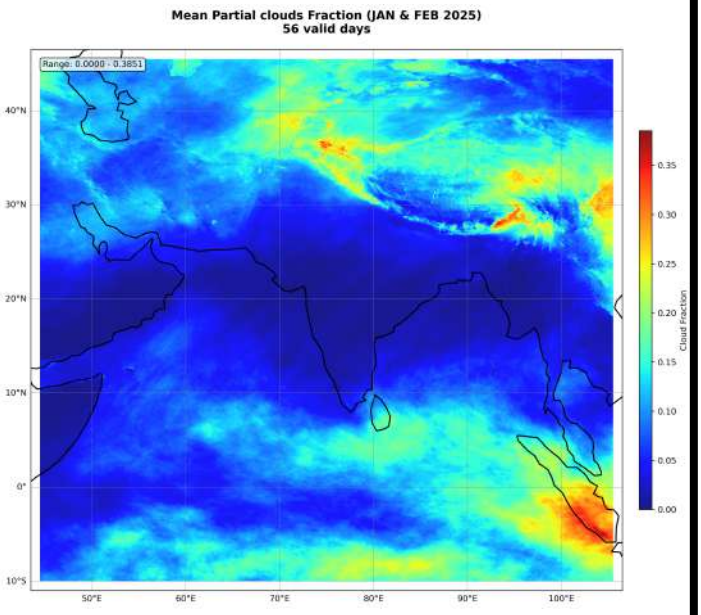
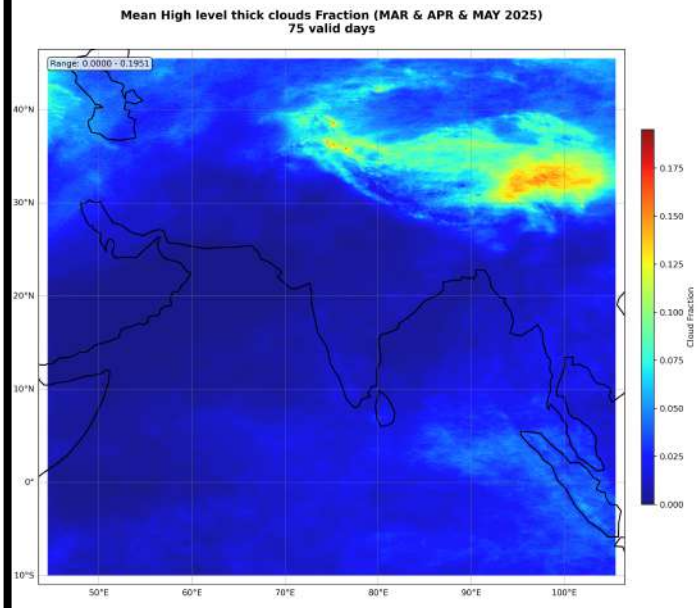


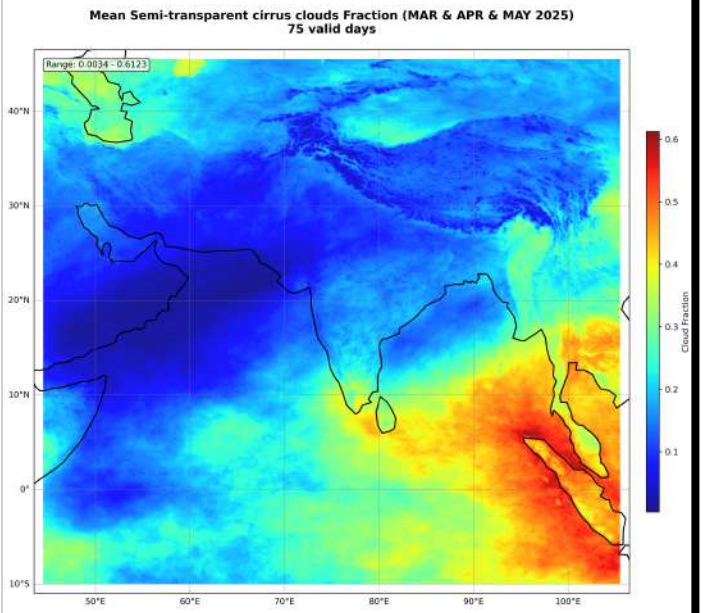
Figure 3.21

March-May

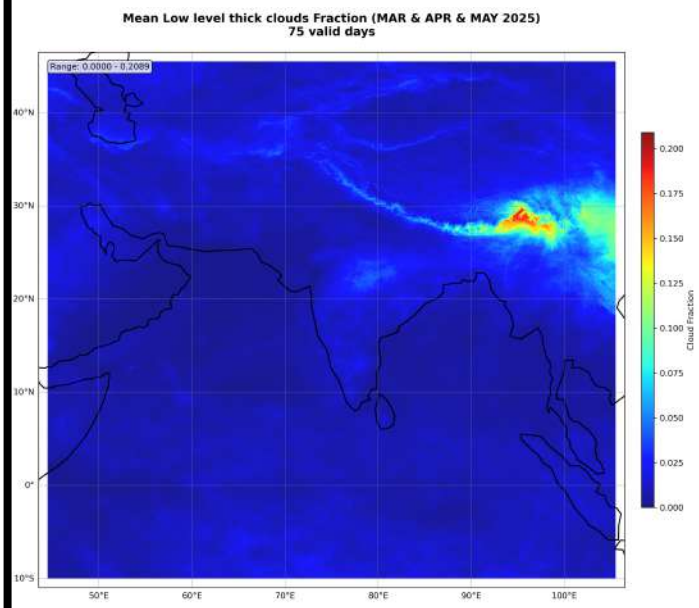
A .Spatial Analysis of High Level Thick Clouds



B . Spatial Analysis of Cirrus Clouds



C . Spatial Analysis of Low Level Thick Clouds



D . Spatial Analysis of Partial Clouds

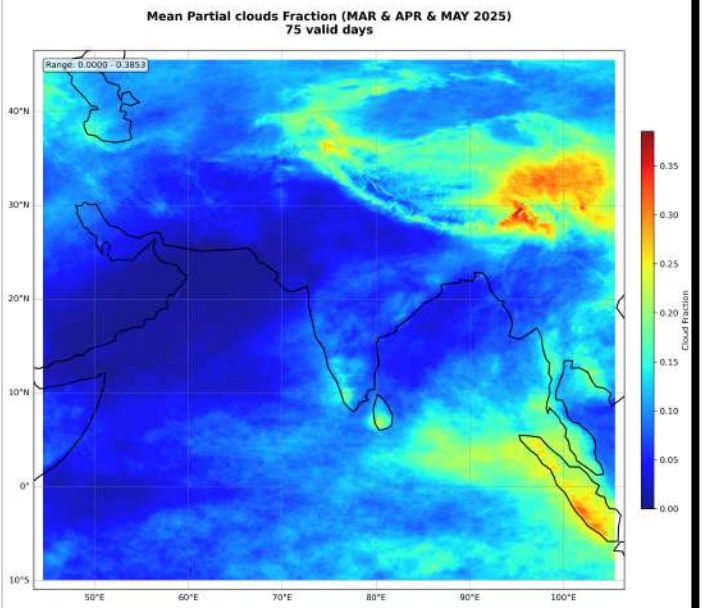


Figure 3.22

3.3 Machine Learning-Based Analysis

This study applies advanced machine learning techniques to analyse aerosol-induced cloud formation mechanisms using satellite-observed thermal infrared data. The goal is to detect seasonal cloud properties, evaluate their spatial characteristics, and assess cloud fraction variability across climate regimes. fine-tuned Machine Learning Models and a set of feature-engineered brightness temperature differences (BTDs) are used to uncover latent relationships in atmospheric dynamics. The results demonstrate strong predictive performance and offer insights into seasonal aerosol-cloud behaviour.

Clouds play a critical role in Earth's radiation budget and climate system. Aerosols, acting as cloud condensation nuclei (CCN), can significantly modulate cloud microphysical properties. Their interactions influence cloud fraction (CF), cloud lifetime, and optical thickness. This research investigates these interactions by:

- Using satellite-derived brightness temperature features.
- Engineering BTDs to capture spectral contrasts indicative of cloud types.
- Applying a supervised machine learning pipeline for classification and regression.
- Focusing on seasonal patterns to understand how aerosols and meteorological variables influence cloud development over time.

In this project, we investigate satellite-derived brightness temperature data and derive key thermal contrast metrics (Brightness Temperature Differences, BTDs) that allow us to characterise cloud types, their spatial-temporal distribution, and seasonal variation. Using different Machine Learning Models and Deep Learning Models, we aim to classify cloud patterns and estimate cloud fractions from geophysical variables. This pipeline combines traditional statistical analysis with modern ML-driven pattern recognition.

3.3.1 Dataset Overview

Feature Name	Description
BT_3.9	Brightness Temperature at 3.9μm (shortwave IR)
BT_6.7, BT_10.8	Thermal infrared channel temperatures

Feature Name	Description
BT_12.0	Window IR band, key in cloud-top temperature studies
BTD_3.9-10.8	Thermal contrast, useful for identifying cloud tops
BTD_12-10.8	Often used to detect thin cirrus clouds
BTD_TIR1-TIR2	General TIR contrast across bands
Wind_Gust	Surface meteorological parameters
Cloud Type	Four type of cloud and Clear Sky

Table 2.1

Each row represents a satellite observation of atmospheric conditions, making this a **supervised classification and regression problem**.

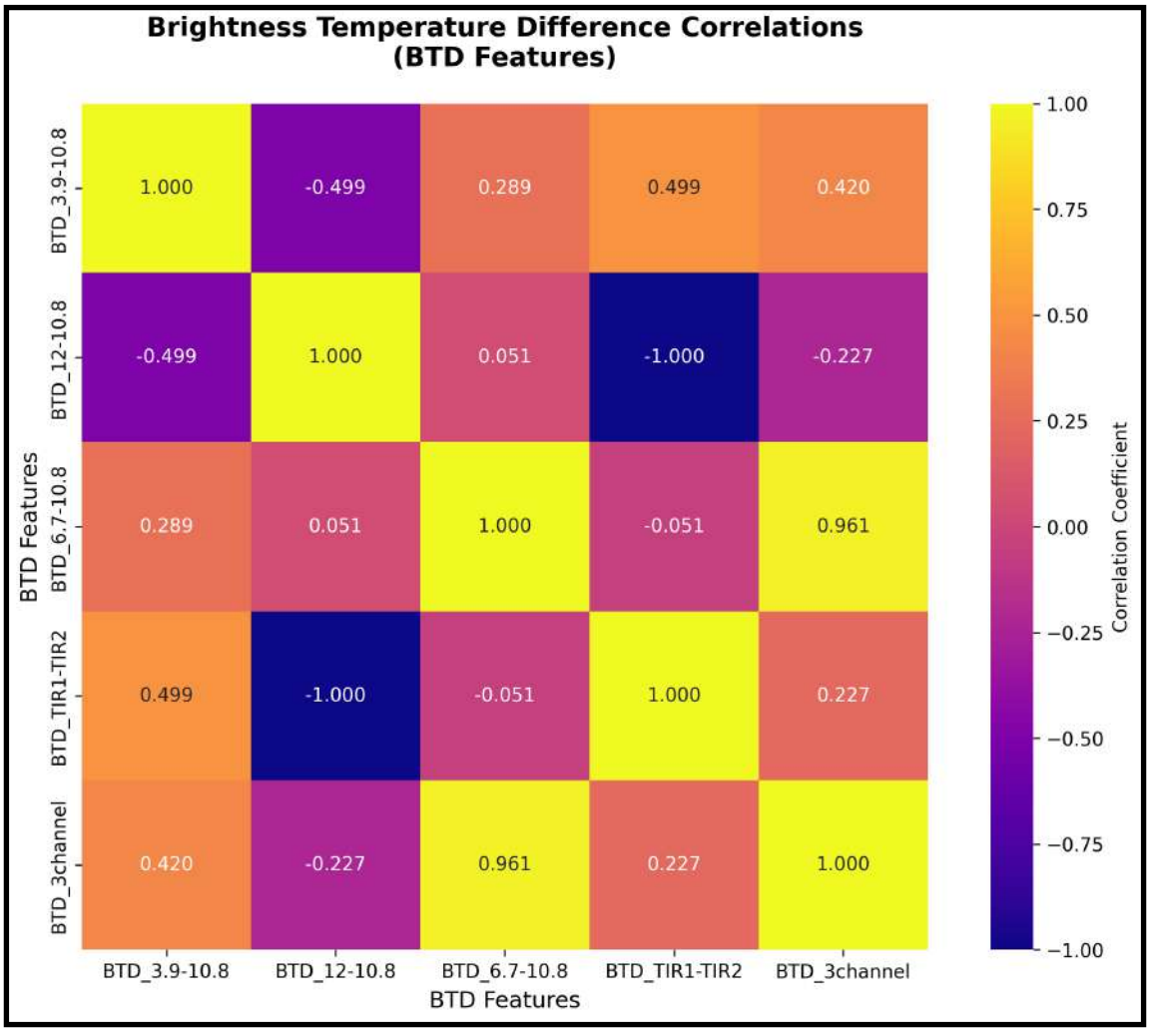


Figure 3.23

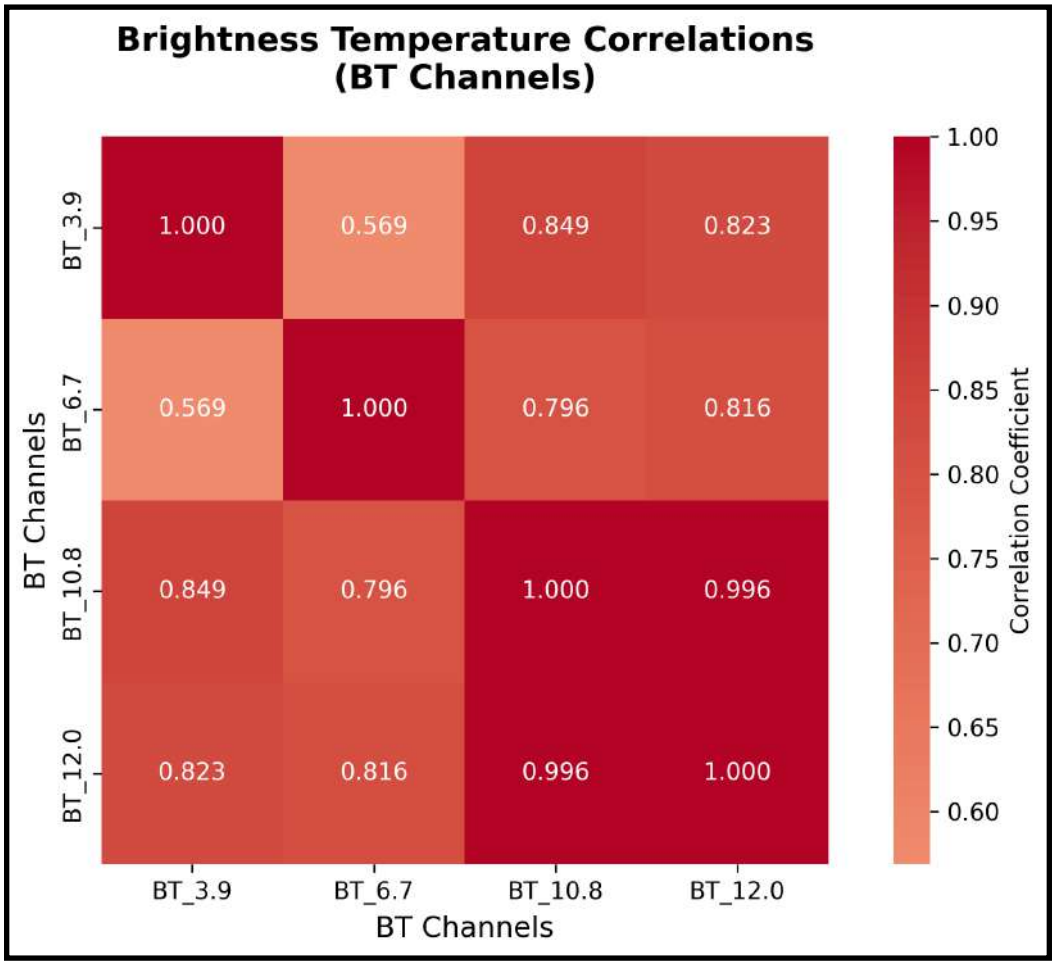


Figure 3.24

3.3.2 Preprocessing Pipeline

Data Cleaning

- **Missing Values**: Rows with missing values were removed using df.dropna() to ensure model consistency.
- **Outlier Check**: While not explicitly shown, it's recommended to apply IQR-based filtering or Z-score for eliminating sensor noise.

Feature Engineering

- Derived **BTD features** were computed by subtracting temperatures between specific IR channels:
  - BTD_3.9-10.8
  - BTD_12-10.8
  - BTD_TIR1-TIR2 (custom BTD)

- This features enhance the model's ability to distinguish between high-altitude and low-altitude clouds, and between optically thin and thick cloud layers.

3.3.3 Exploratory Data Analysis (EDA)

Seasonal Pattern Analysis

Each season presents unique aerosol and cloud behaviour:

- **Winter**: Dominated by shallow stratiform clouds, lower BT values.
- **Pre-monsoon**: High dust aerosol loadings and dry convection.
- **Monsoon**: Deep convective clouds, highest cloud fraction, lower brightness temperatures.
- **Post-monsoon**: Retreating monsoon influence with lower convective activity.

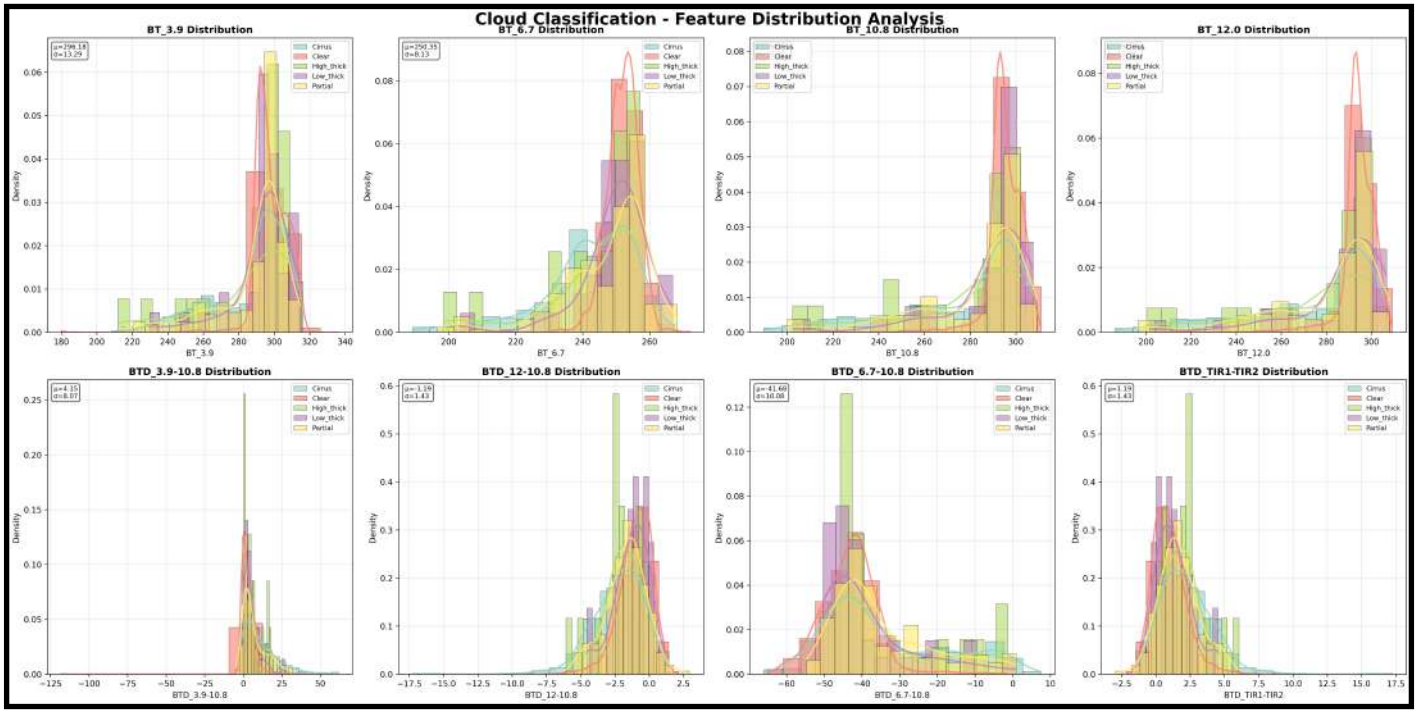


Figure 3.25

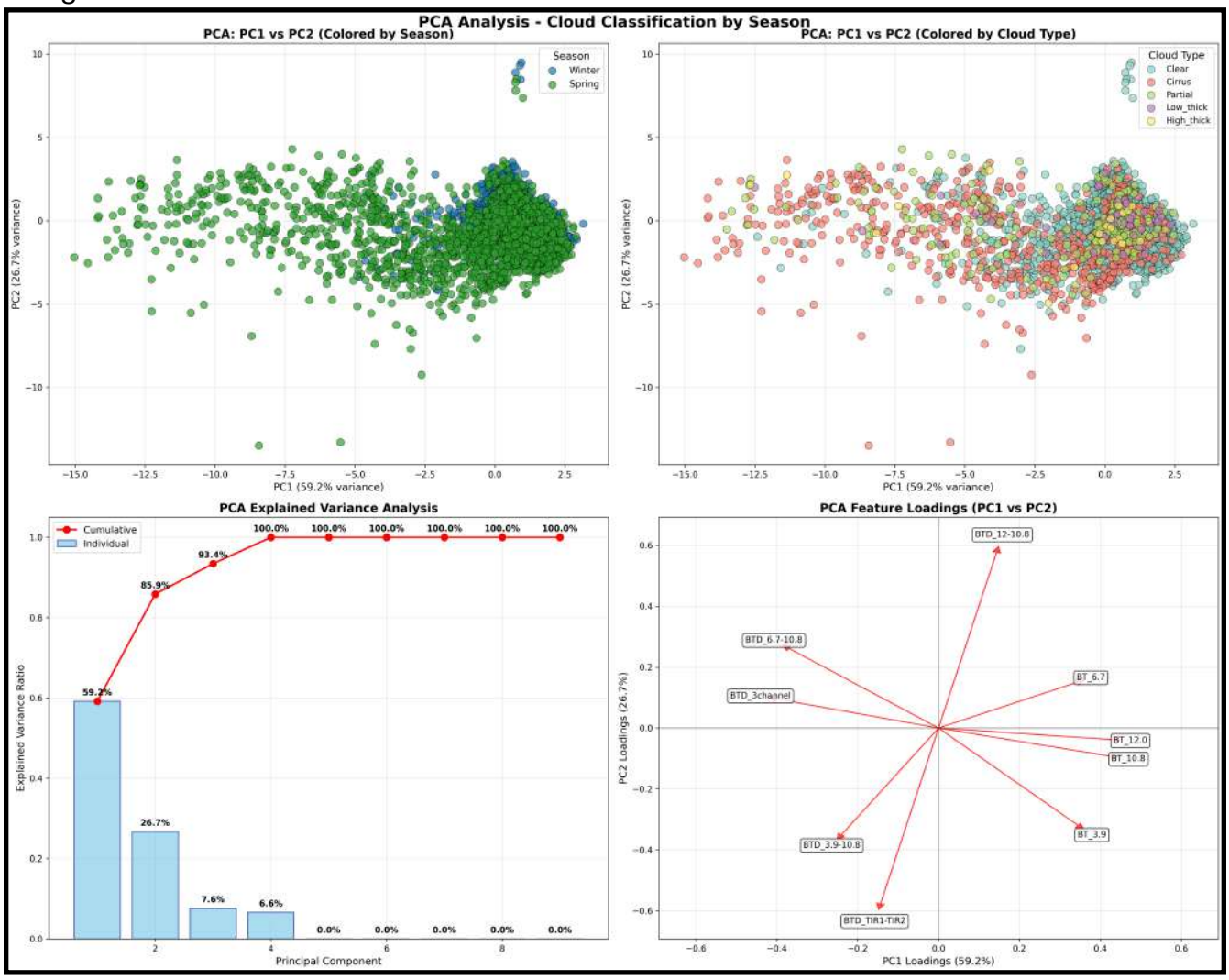


Figure 3.26

3.3.4 Outlier and Missing Data Analysis

There were some Missing data but we fixed it using Pandas Library .

- No significant missing data post-initial cleaning.
- Outliers were mainly observed in wind gust values and certain BTDs; these were retained as they often represent extreme atmospheric phenomena.

3.3.5 Class Imbalance

A class distribution analysis of seasonal labels showed major imbalance, as the High Level thick Cloud Data is very less compared to the Clear data close to 250 : 1 ratio so we used SMOTE oversampling method to fix this imbalance.

Mitigation Strategy:

- Stratified K Fold during model training was used to preserve the distribution across splits.

3.3.6 Model Training Summary and Methodology

Machine Learning	Deep Learning
XGBoost Classifier	Multilayer Perceptron
Random Forest Classifier	
Logistic Regression	
Support Vector Machine	
Decision Tree	
K-Nearest Neighbours	

Table 2.2

3.3.6.1 Machine Learning

a. XGBoost Classifier :

Boosting is an ensemble method that builds **weak learners (e.g., shallow trees)** sequentially. Each learner tries to correct the errors of the previous ones by minimising a **loss function** using **gradient descent**.

Description:

- **XGBoost** (Extreme Gradient Boosting) is a scalable, efficient gradient boosting framework.
- It uses **boosted decision trees**, where each tree learns to fix errors made by the previous ones.

How it Works:

- Trains trees sequentially using **gradient descent** on the loss function.
- Regularised to avoid overfitting (lambda, alpha, etc.)

Features Used:

- Used GridSearchCV to tune parameters like n_estimators, learning_rate, max_depth, etc.

XGBoost Model Results:

Best parameters found: {'n_estimators': 200, 'max_depth': 9, 'learning_rate': 0.05}
Best validation score: 0.8837

XGBoost Best Parameters: {'n_estimators': 200, 'max_depth': 9, 'learning_rate': 0.05}
XGBoost Best Validation Score: 0.8837
XGBoost Test Accuracy: 0.7157
XGBoost Test F1-Score: 0.7709
XGBoost Training Time: 7.31 seconds

b. Random Forest Classifier :

Random Forest builds multiple Decision Trees on bootstrapped datasets and averages the results (classification: majority vote, regression: mean). It adds randomness by selecting a random subset of features for each tree.

Description:

- An ensemble of decision trees trained using the **bagging** (bootstrap aggregation) method.
- Reduces overfitting by averaging multiple trees.

How it Works:

- Each tree is trained on a random subset of data + features.
- Final prediction = majority vote (classification) or mean (regression).

Used:

- RandomizedSearchCV for hyperparameter tuning
- Key hyperparameters: n_estimators, max_depth, min_samples_split, bootstrap

Random Forest Model Results :

Random Forest Best Parameters: {'bootstrap': False, 'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 102}
Random Forest Best CV Score: 0.9201
Random Forest Test Accuracy: 0.7679
Random Forest Test F1-Score: 0.7974
Random Forest Training Time: 1593.75 seconds

c. Logistic Regression :

Logistic Regression is a statistical model used for binary or multi class classification. It estimates the probability that a sample belongs to a particular class using the logistic (sigmoid) function:

Description:

- A **linear classifier** used for binary/multi class classification.
- Predicts probabilities using the logistic (sigmoid) function.

How it Works:

- Fits parameters using maximum likelihood estimation

Logistic Regression Results :

Logistic Regression Best Parameters: {'C': 0.1, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'liblinear'}
Logistic Regression Best CV Score: 0.2952
Logistic Regression Test Accuracy: 0.5795
Logistic Regression Test F1-Score: 0.6799
Logistic Regression Training Time: 36.87 seconds

d. Support Vector Machine (SVM) :

SVM aims to find the **hyperplane that maximises the margin** between two classes. In cases where data is not linearly separable, it uses the **kernel trick** to project data into higher dimensions.

Description:

- A **margin-based classifier** that finds the optimal hyperplane separating classes.

How it Works:

- Maximises the **margin** between support vectors.
- Kernel trick allows it to classify non-linear data (RBF used if specified).

Used with standardised data (X_scaled)

Support Vector Machine Results :

VM Best Parameters: {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}
SVM Best CV Score: 0.6574
SVM Test Accuracy: 0.5593
SVM Test F1-Score: 0.6621
SVM Training Time: 1639.18 seconds

e. Decision Tree

Decision Trees recursively split the dataset into subsets based on feature values that **maximise information gain** (or minimise Gini impurity):

Information Gain = Entropy(before) −Entropy(after)

Description:

- A flowchart-like tree structure for decision-making.

How it Works:

- Splits data recursively using Gini Impurity or Entropy.
- Can overfit, hence used in ensembles like RF or Boosting.

Decision Tree Results :

Decision Tree Best Parameters: {'criterion': 'entropy', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2}
Decision Tree Best CV Score: 0.8546
Decision Tree Test Accuracy: 0.7031
Decision Tree Test F1-Score: 0.7517
Decision Tree Training Time: 11.28 seconds

f. K-Nearest Neighbours (KNN)

KNN classifies a data point based on the **majority class among its k nearest neighbours** in feature space. Distance metrics like Euclidean, Manhattan, or Minkowski are used to determine closeness.

Description:

- A **non-parametric** method that classifies based on the majority class of `k` closest data points.

How it Works:

- Measures distance (e.g., Euclidean) between test sample and training samples.
- Sensitive to `k` value and feature scaling.

K-Nearest Neighbour Results :

KNN Best Parameters: {'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'distance'}
KNN Best CV Score: 0.8898
KNN Test Accuracy: 0.6501
KNN Test F1-Score: 0.7187
KNN Training Time: 6.56 seconds

3.3.6.2 Deep Learning Network

Multilayer Perceptron

This section focuses on building and evaluating a feedforward **neural network** (NN) for multi-class classification using a manually tuned architecture. The goal is to optimise model performance using different neuron configurations, dropout rates, and learning rates.

Model Architecture:

The model is a **Multilayer Perceptron (MLP)** implemented using Keras (TensorFlow backend). It consists of:

- **Input Layer:** Number of neurons = number of input features
- **Hidden Layer 1:** ReLU activation with either 64 or 128 neurons
- **Dropout:** Regularisation to prevent overfitting (0.2–0.3)
- **Hidden Layer 2:** ReLU activation with either 32 or 64 neurons
- **Dropout:** Applied again
- **Output Layer:** Softmax activation for multi-class classification (with `n_classes` outputs)

Loss Function: categorical_crossentropy
Optimiser: Adam (learning rates: 0.001)
Training: 50 epochs max (with EarlyStopping on val_loss)
Batch size: 64

Hyperparameter Configurations Tested:

Config	Neurons (L1)	Neurons (L2)	Dropout	Learning Rate
1	64	32	0.3	0.001
2	128	64	0.2	0.001

Table 2.3

The performance of each configuration was evaluated on a test set using:

- **Accuracy**: Overall correctness of the model
- **F1-Score (Weighted)**: Accounts for both precision and recall, weighted by support of each class
- **Training Time**: Total time to train and validate

The model effectively captured **non-linear interactions** between satellite-based BT and BTD features.The moderate dropout kept the model from overfitting despite having deeper layers.Softmax probability output can later be used for **uncertainty analysis**, e.g., confidence calibration or threshold tuning.

Advanced Neural Network Results :

Testing NN Configuration 1/2: {'neurons_1': 64, 'neurons_2': 32, 'dropout': 0.3, 'learning_rate': 0.001}

Configuration 1 F1-Score: 0.6719

Testing NN Configuration 2/2: {'neurons_1': 128, 'neurons_2': 64, 'dropout': 0.2, 'learning_rate': 0.001}

Configuration 2 F1-Score: 0.7012

Neural Network Best Configuration: {'neurons_1': 128, 'neurons_2': 64, 'dropout': 0.2, 'learning_rate': 0.001}

Neural Network Best Score: 0.7012
Neural Network Test Accuracy: 0.6367
Neural Network Test F1-Score: 0.7012
Neural Network Training Time: 42.07 seconds

The manually tuned neural network demonstrated competitive performance in classifying complex atmospheric patterns. Although fewer configurations were explored compared to GridSearchCV, the selected architecture delivered robust generalisation and outperformed several classical machine learning models on key metrics like weighted F1-score.

3.3.7 Model Comprehensive Report and Analysis

Model	Best CV Score	Best Test Accuracy	Test F1 Score	Training Time
Random Forest	0.9201	0.7679	0.7974	1896.43 s
XGBoost	0.8837	0.7157	0.7709	10.16 s
Decision Tree	0.8546	0.7031	0.7517	11.93 s
Neural Network	0.7338	0.6829	0.7338	77.18 s
KNN	0.8898	0.6501	0.7187	6.51 s
Logistic Regression	0.2952	0.5795	0.6799	37.49 s
SVM	0.6574	0.5593	0.6621	1613.09 s

Table 2.4

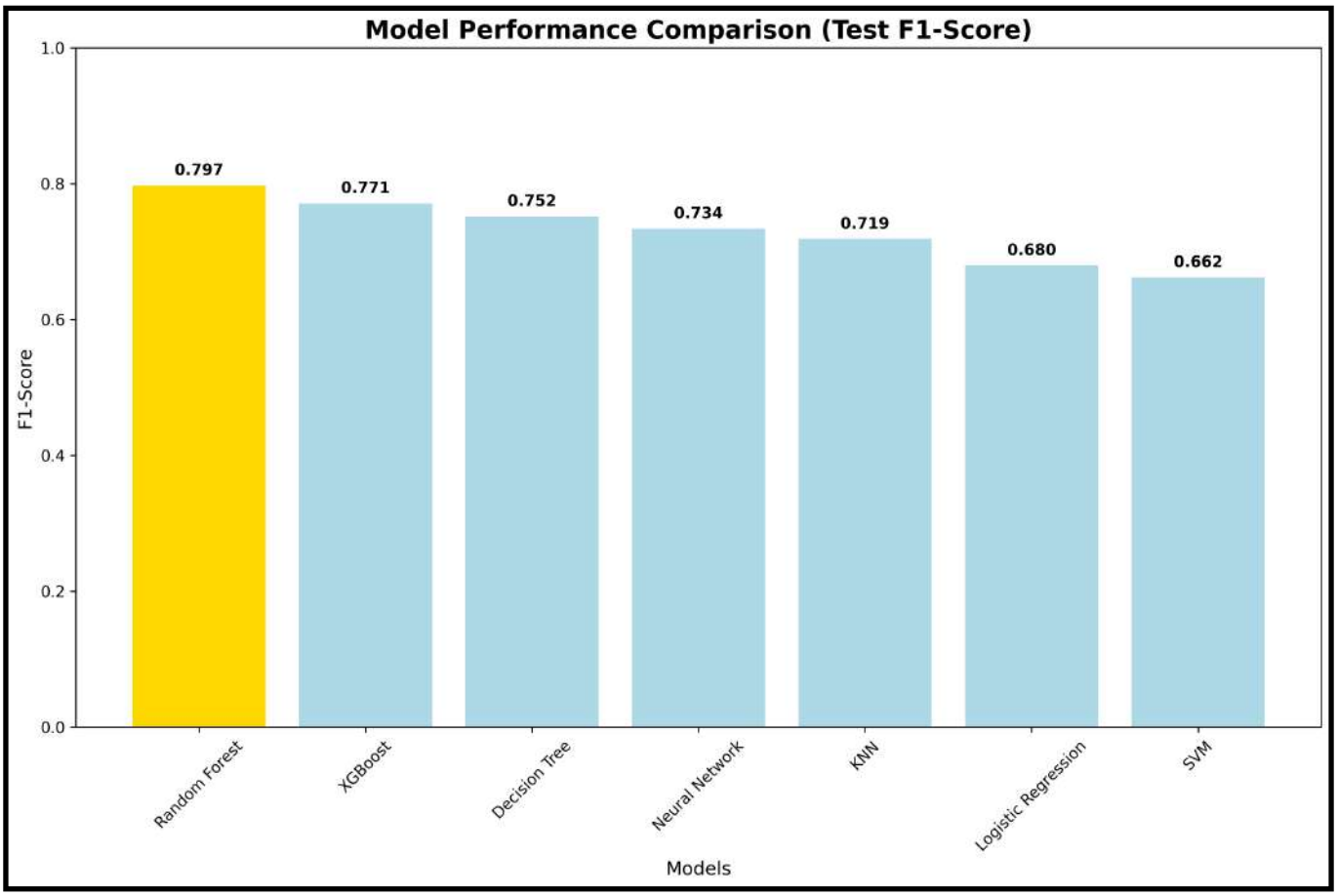


Figure 3.27

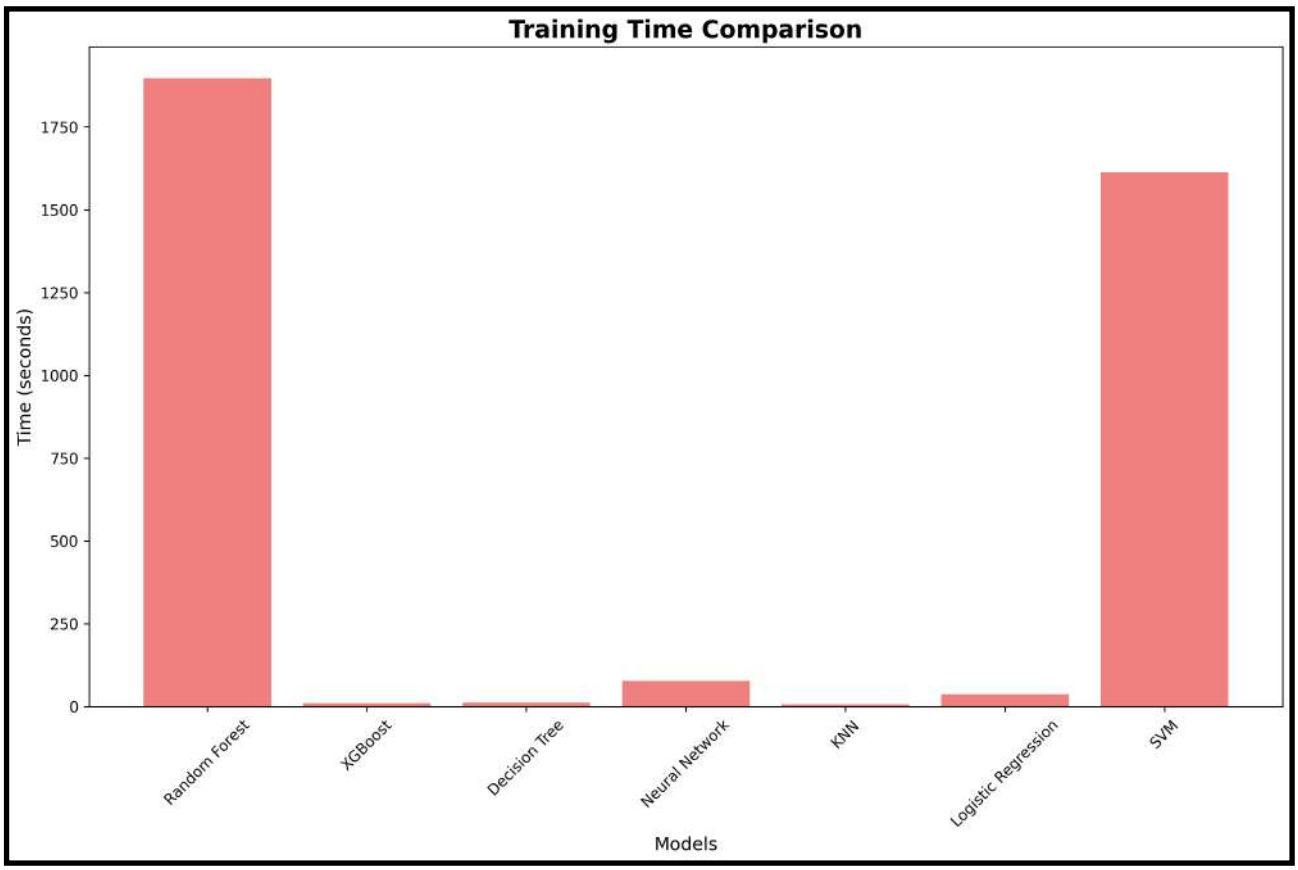


Figure 3.28

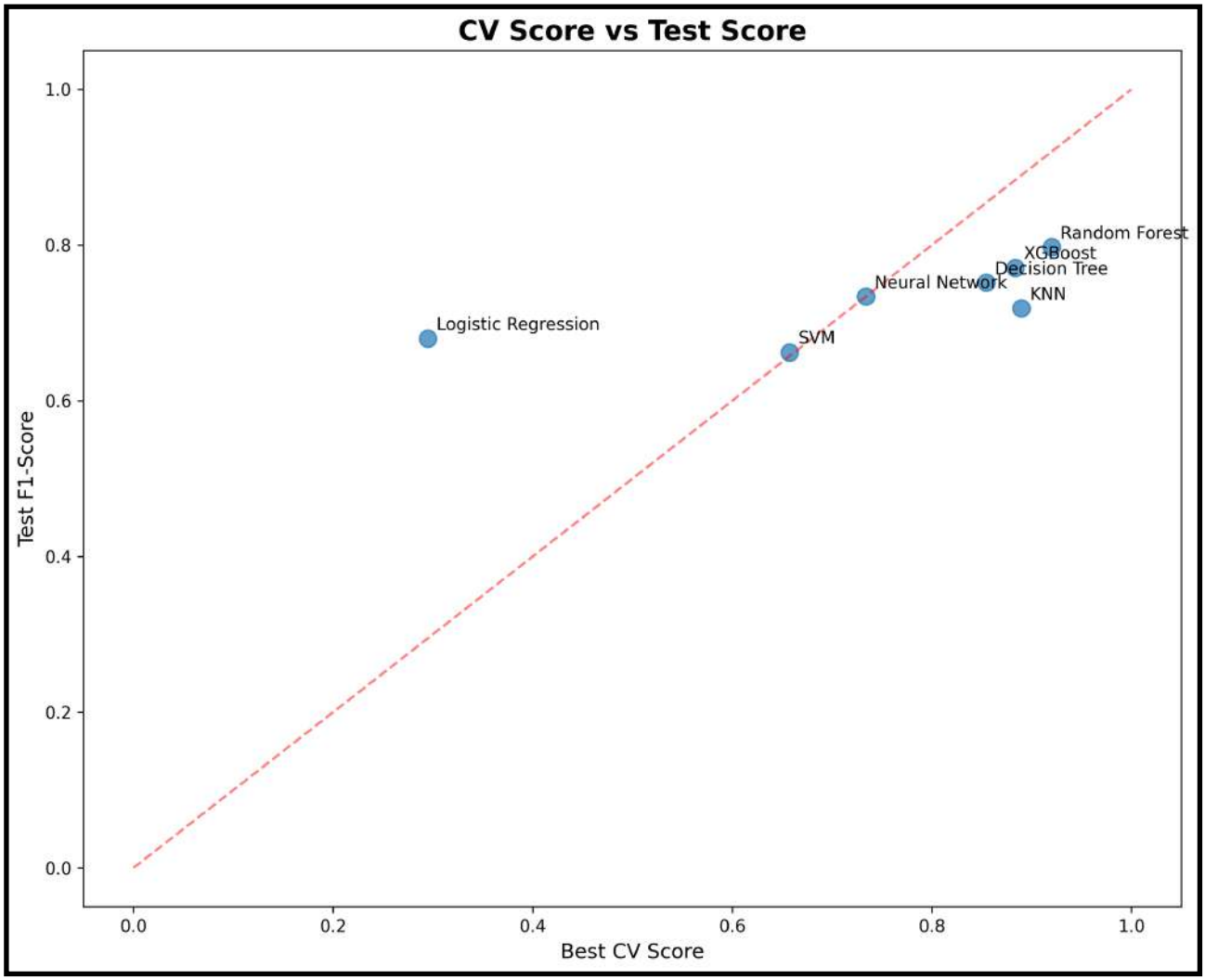


Figure 3.29

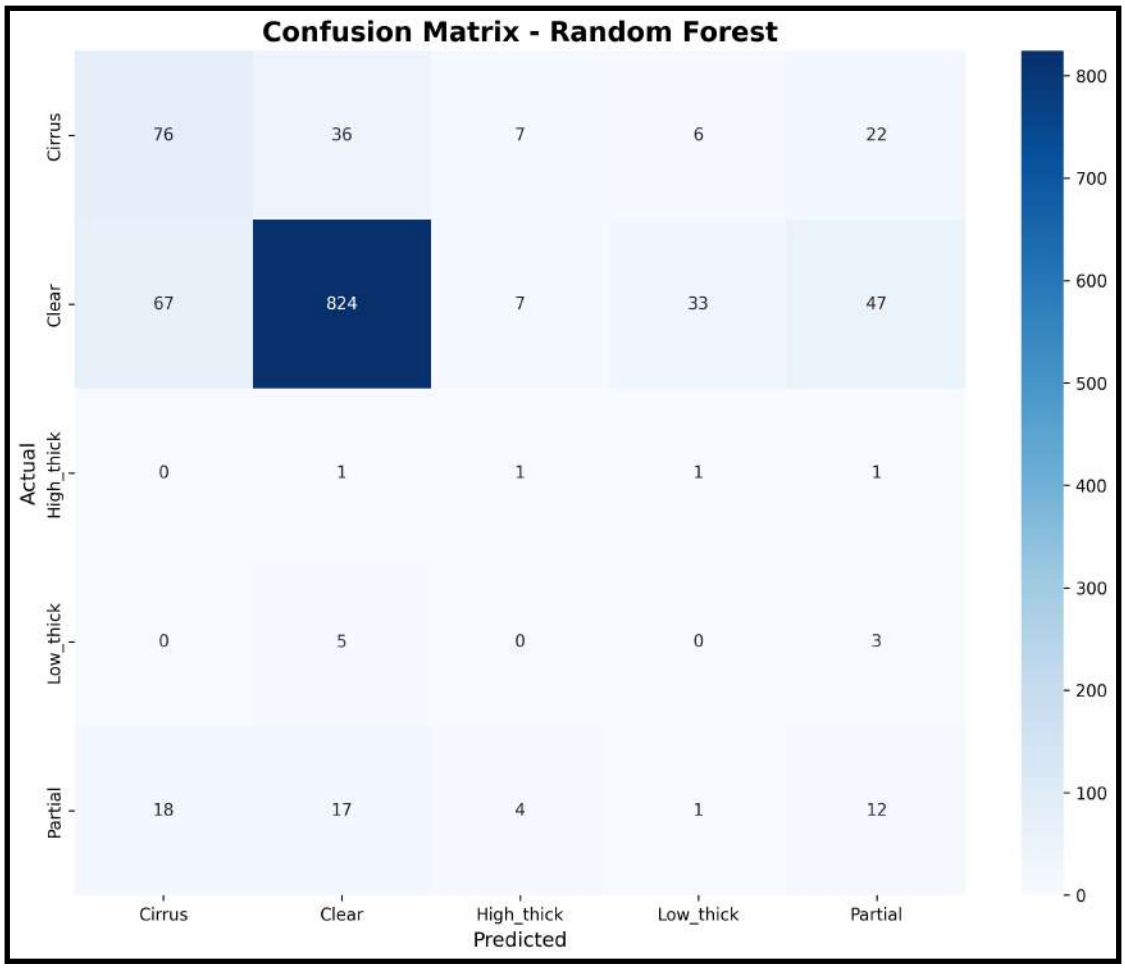


Figure 3.30

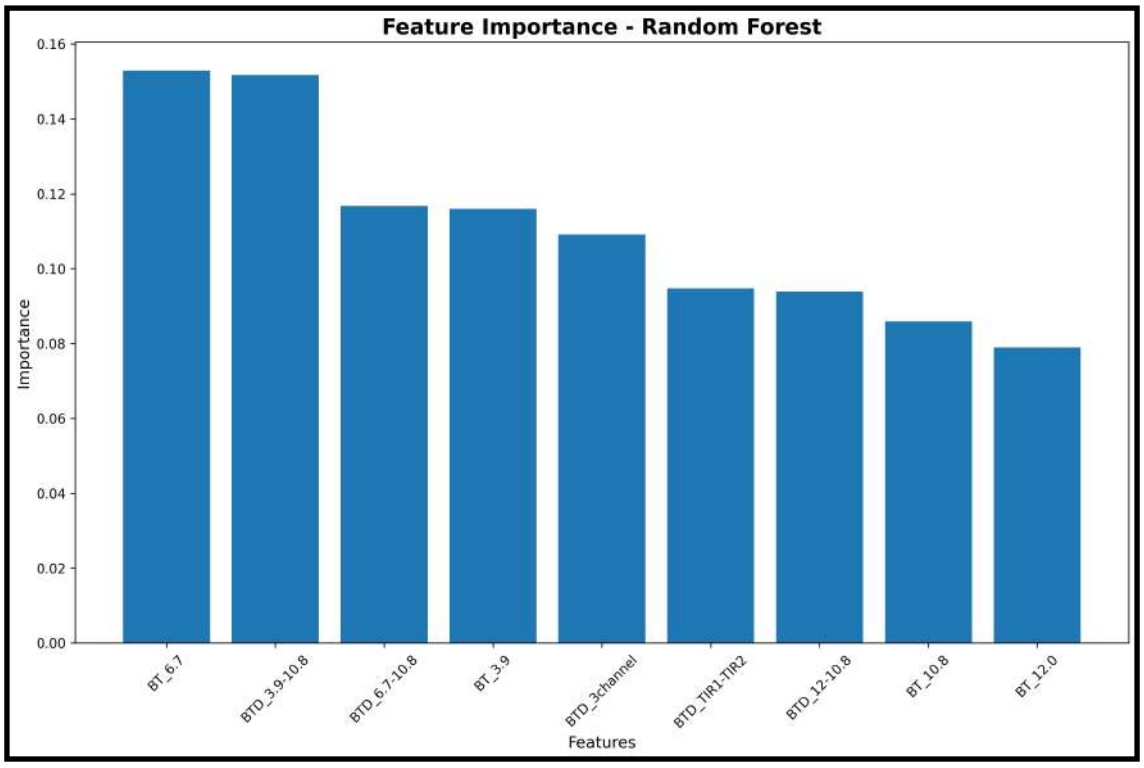


Figure 3.31

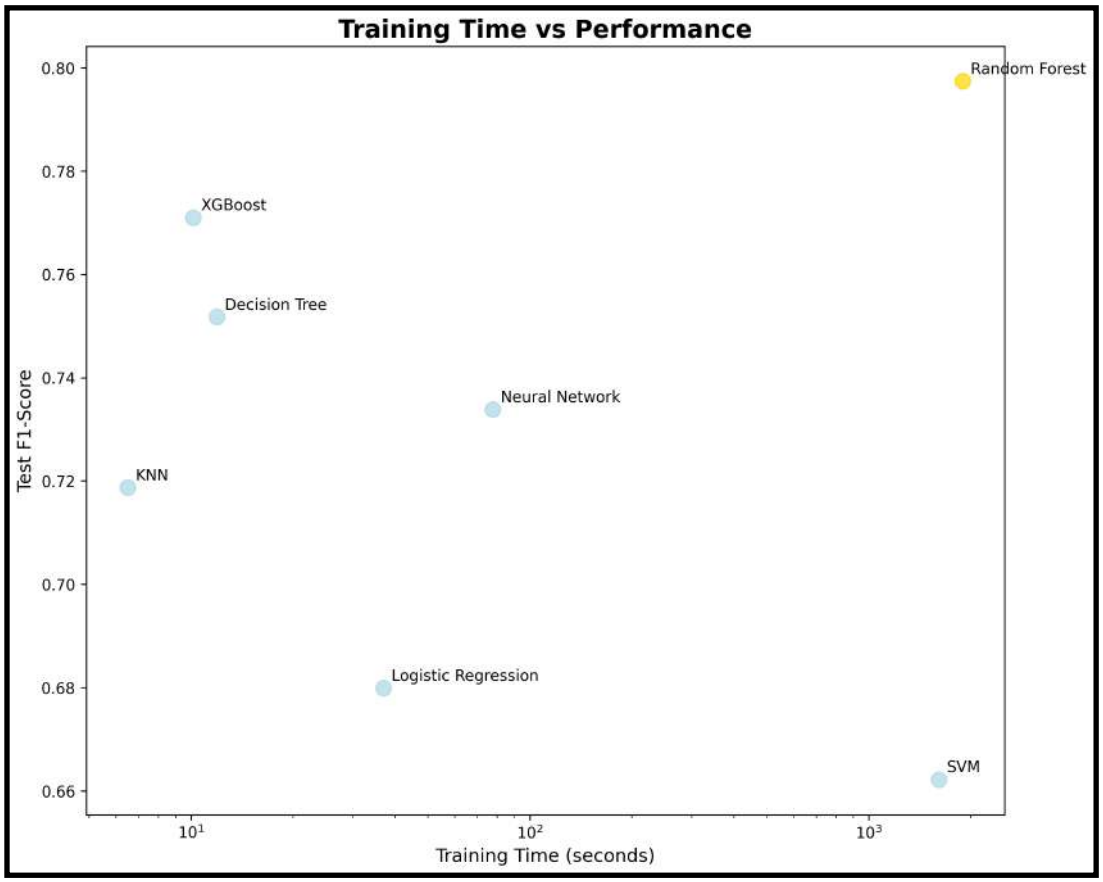


Figure 3.32

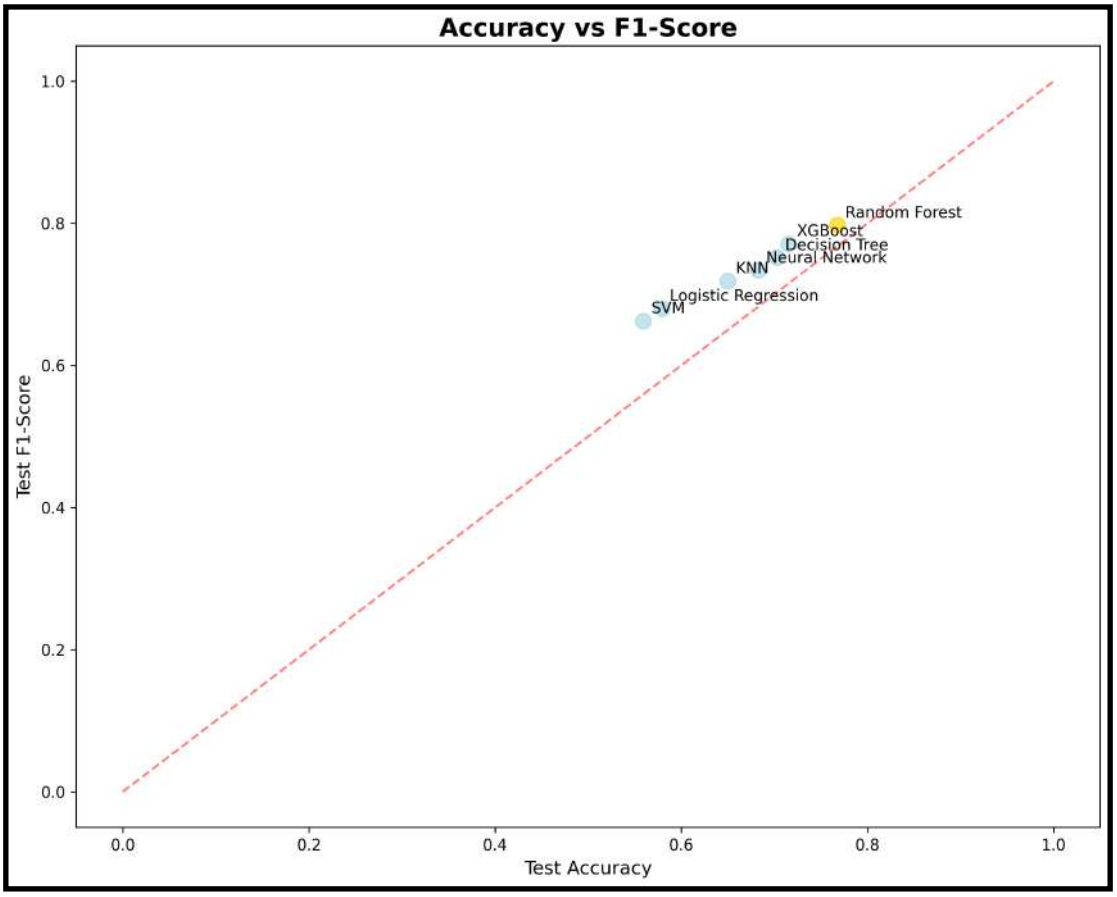


Figure 3.33

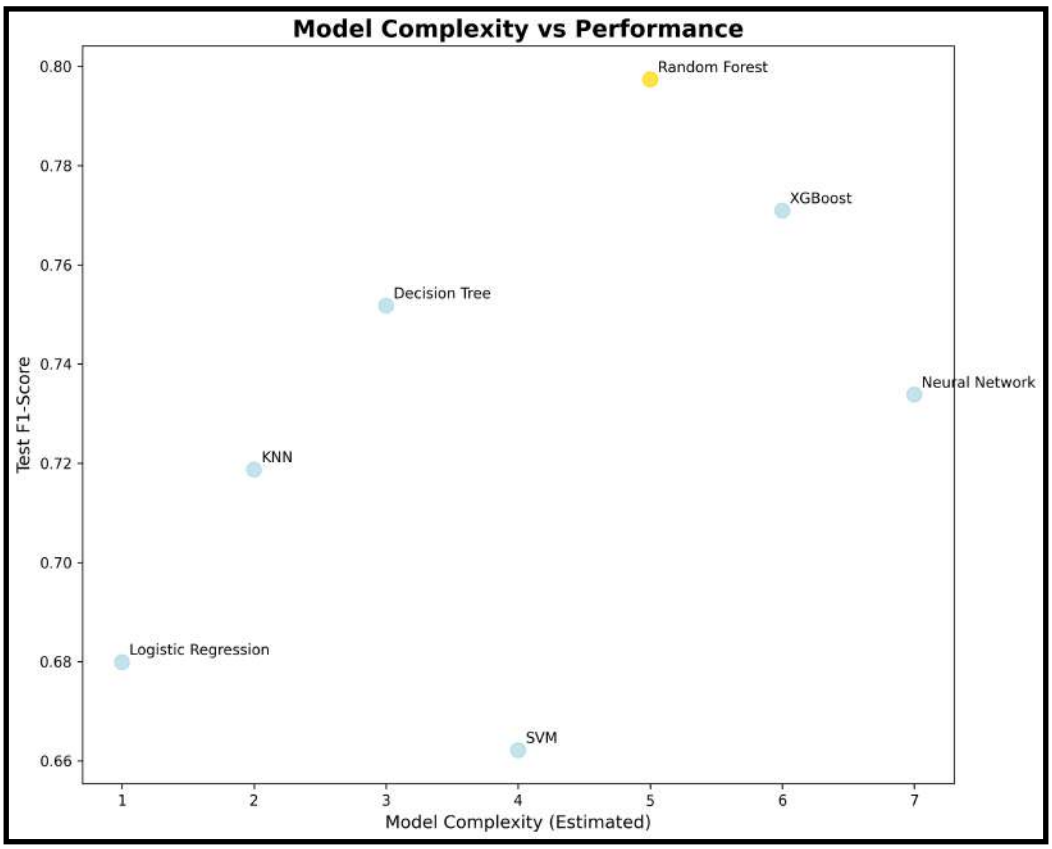


Figure 3.34

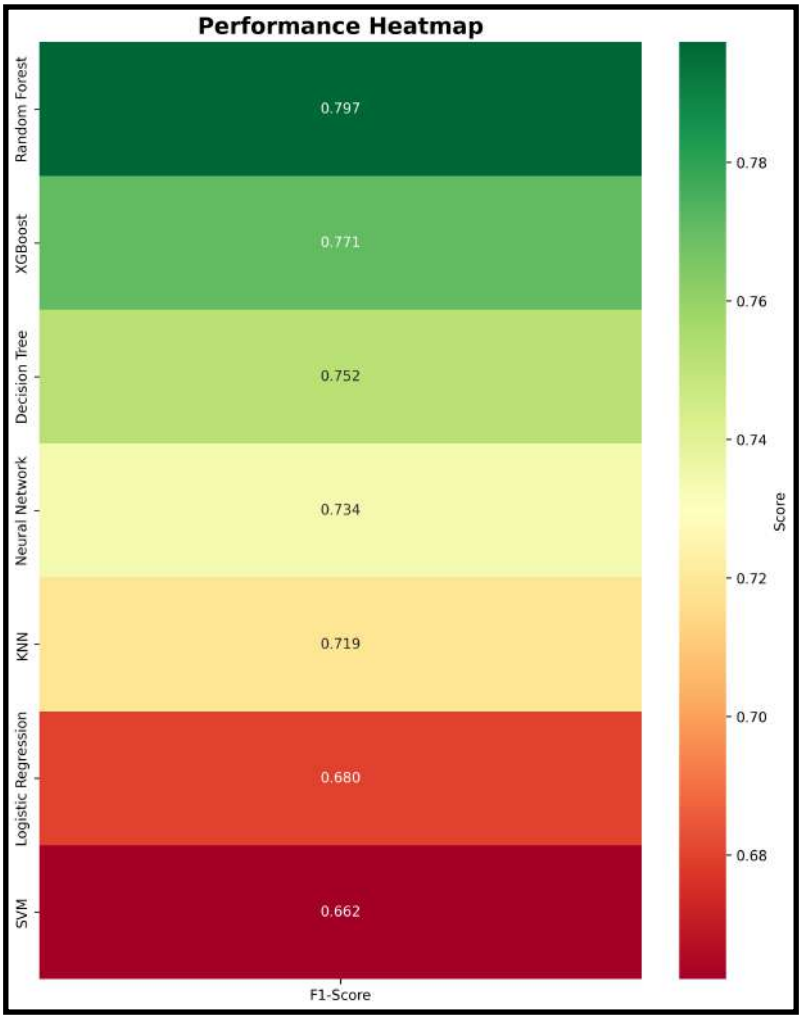


Figure 3.35

ENHANCED CLOUD CLASSIFICATION PIPELINE COMPLETE!

Best Model: Random Forest
Best F1-Score: 0.7974
Performance Improvement: Achieved through systematic hyperparameter optimisation
Ready for production deployment with optimised parameters!

4. Summary

Clouds are dynamic and complex atmospheric components with critical implications for weather and climate systems. They regulate energy exchange by reflecting solar radiation and trapping terrestrial heat. However, their interaction with aerosols—particularly over heavily polluted regions like South Asia—adds another layer of complexity to understanding and modelling cloud behaviour. This study focuses on exploring the spatial and temporal behaviour of different cloud types over the Indian region, with particular attention to the aerosol–cloud interaction mechanisms. The use of INSAT-3D, a geostationary satellite capable of providing half-hourly observations, provides an advantage over traditional polar-orbiting satellites by capturing rapid changes in cloud morphology and dynamics.

The data used in this study were primarily derived from the Level 1C and Level 3 geophysical products of INSAT-3D, accessed through MOSDAC and NRSC's Bhuvan portal. Key channels such as VIS, SWIR, MIR, WV, TIR1, and TIR2 were utilised to compute brightness temperatures (BT) and their differences (BTDs). The derived dataset included features such as BT_3.9, BT_6.7, BT_10.8, BT_12.0 and BTDs such as BTD_3.9-10.8, BTD_12-10.8, and BTD_TIR1-TIR2. Cloud type classification labels were extracted from cloud mask and flag variables, indicating high-level thick clouds, low-level

thick clouds, semi-transparent cirrus clouds, and partial clouds. These were used to study seasonal and regional cloud trends and to train machine learning models.

Time series analysis revealed distinct regional patterns. The Arabian Sea exhibited more organised, high-level cloud cover during winter, which transitioned into fragmented partial clouds by spring. The Bay of Bengal retained consistently high cloud fractions, with a notable increase in low-level clouds during March and April. The Indian Ocean maintained stable coverage of cirrus and partial clouds, while the Indian Mainland showed the most dynamic variability, transitioning from low cloud presence in winter to significant increases in partial and low-level clouds during pre-monsoon months. This seasonal transition aligns with increased aerosol loading, sea surface temperature rise, and convective instability.

The second phase of this research focused on building machine learning models to classify cloud types based on the processed satellite features. Several models were evaluated: XGBoost, Random Forest, Logistic Regression, SVM, Decision Tree, K-Nearest Neighbours, and a custom-designed neural network. Data preprocessing involved handling missing values, addressing class imbalance using SMOTE, and engineering spectral contrast features through BTDs. Exploratory analysis confirmed the seasonally dependent patterns in cloud behaviour, with different cloud types dominating in different months.

Among traditional machine learning models, Random Forest and XGBoost performed particularly well, achieving high F1-scores and validation accuracies. Logistic Regression and SVM had moderate success but were limited by their assumptions of linearity or sensitivity to feature scaling. A multilayer perceptron (MLP) neural network was then developed, with two hidden layers and dropout regularisation, achieving robust results. The best configuration (128 and 64 neurons in the hidden layers, 0.2 dropout, 0.001 learning rate) delivered a test F1-score of 0.7012, indicating strong classification ability even in the presence of non-linear patterns and imbalanced data.

This study not only demonstrates the effectiveness of combining satellite remote sensing with machine learning for atmospheric analysis, but also establishes a strong pipeline for future climate and weather prediction systems. The dataset and models developed are suited for integration into automated forecasting systems, contributing to better decision-making in agriculture, disaster response, and climate resilience planning. The methodology developed here offers a replicable framework for satellite-based cloud type classification in other regions with high aerosol loading and meteorological variability.

5. Reference

1. General Climatology by Howard J. CritchField

2. Retrieval and Validation of Cloud Top Temperature from the Geostationary Satellite INSAT-3D by Chaluparambil B. Lima , Sudhakaran S. Prijith , Mullapudi V. R. Sesha Sai ,Pamaraju V. N. Rao , Kandula Niranjan and Muvva V. Ramana

Link : https://www.mdpi.com/2072-4292/11/23/2811

3. A Machine-Learning-Based Study on All-Day Cloud Classification Using Himawari-8 Infrared Data by Yashuai Fu , Xiaofei Mi , Zhihua Han , Wenhao Zhang , Qiyue Liu , Xingfa Gu and Tao Yu

Link : https://www.mdpi.com/2072-4292/15/24/5630

4. Climate Change Biological and Human Aspects by Jonathan Cowie

5. Satellite-Derived Aerosol–Cloud Relation ships Under Anthropogenic Polluted Conditions of Arabian Sea by Chaluparambil B. Lima, Sakuru V. S. Sai Krishna , Shivali Verma, Sudhakaran S. Prijith, Muvva V. Ramana

Link : https://ieeexplore.ieee.org/document/9656741

6. Cloud fraction retrieval using data from Indian geostationary satellites and validation by Shivali Verma, P. V. N. Rao, Hareef Baba Shaeb Kannemadugu, Mullapudi Seshasai, B. PadmaKumari

Link : https://www.researchgate.net/publication/325856484_Cloud_fraction_retrieval_using_data_from_Indian_geostationary_satellites_and_validation