CLOUD TYPE IDENTIFICATION OVER THE WESTERN GHATS FROM RADAR DATA USING MACHINE LEARNING MODELS: AN AI BASED APPROACH

Project Report

Submitted in partial fulfilment of the award of degree of

Master of Science in Meteorology

by

Mehzooz Nizar

(Reg No: 30221008)

Department of Atmospheric Sciences,

Cochin University of Science and Technology

Under the guidance of

Mr. Ambuj Kumar Jha & Dr. Manmeet Singh

INDIAN INSTITUTE OF TROPICAL METEOROLOGY PUNE,

MAHARASHTRA June 2023

भारतीय उष्णदेशीय मौसम विज्ञान संस्थान
(पृथ्वी विज्ञान मंत्रालय, भारत सरकार का एक स्वायत्त संस्थान)
डॉ. होमी भाभा मार्ग पाषाण, पुणे– ४११ ००८
INDIAN INSTITUTE OF TROPICAL METEOROLOGY
(An Autonomous Institute of the Ministry of Earth Sciences, Govt. of India)
Dr. Homi Bhabha Road, Pashan, Pune - 411 008. India

MoES/IITM/ May 19, 2023

CERTIFICATE

This is to certify that the project entitled "**Cloud type identification over the Western Ghats from radar data using machine learning models: an AI based approach**" submitted by Mehzooz Nizar (Register No:30221008) for the partial fulfillment of M.Sc. degree in Meteorology, Department of Atmospheric Sciences, Cochin University of Science and Technology has been carried out by him at Indian Institute of Tropical Meteorology (IITM), Pune under our guidance and supervision.

Date: 19/05/23

Place: Pune

Mr. Ambuj Kumar Jha

Scientist-D

Indian Institute of Tropical Meteorology, Pune

Dr. Manmeet Singh

Scientist-D

Indian Institute of Tropical Meteorology, Pune

DECLARATION

I hereby declare that this project work entitled "**CLOUD TYPE IDENTIFICATION OVER THE WESTERN GHATS FROM RADAR DATA USING MACHINE LEARNING MODELS: AN AI BASED APPROACH**" submitted at the Department of Atmospheric Sciences, Cochin University of Science and Technology for the partial fulfillment of the award of M.Sc. degree in Meteorology is an authentic record of my own work carried out under the guidance of Mr. Ambuj Kumar Jha & Dr. Manmeet Singh. I further declare that no part of this thesis had already been submitted or is currently been submitted for any other degree.

Date :25-05-2023

Place: Kochi

Mehzooz Nizar

Register No: 30221008

Department of Atmospheric Sciences

Cochin University of Science and Technology

ACKNOWLEDGEMENTS

It is my immense pleasure to express my sincere gratitude to my guides Mr. Ambuj Kumar Jha, Scientist-D, IITM , Pune and Dr.Manmeet Singh, Scientist-D, IITM, Pune for providing me an opportunity to carry out my project work at IITM Pune. I am really grateful for their strong support, valuable suggestions and guidance which helped me in completing the project successfully. I am privileged that I got an opportunity to work with them and was able to learn from them about the field.

A Special Thanks to Mr. Hamid Ali Syed, Purdue University-USA and Max Grover**,**Atmospheric Science Software Engineer at Argonne National Laboratory for helping me through the Open Radar Science Community.
I also take this opportunity to thank Ms.Meenu R Nair, Research Scholar IITM Pune and Mr. Vaisakh S B, Project Scientist-1,IITM, Pune and my friend Adithiy R for their timely-help and suggestions at various stages of the work.

I am also grateful to Dr.Satheesan K, Head of the Department of Atmospheric Sciences, Cochin University of Science and Technology, Dr.C.A. Babu, Dr. Manoj M G, Prof. B. Chakrapani, Dr.Abhilash.S, Dr.Madhu.V, Dr. LekshmyP.R., Dr.Midhun M, Dr. Sreekala P.P and office staffs Department of Atmospheric Sciences, CUSAT and all other professors from every departments who imparted knowledge in me.

I would also like to thank all my friends for their help and support that they have offered throughout the period. There are no words to express my thanks to my family for comforting and encouraging me as always.

Table of Contents

ABSTRACT 7
1.INTRODUCTION 8
2. Measurement site and Instruments used 11
2.1. Measurement site 11
2.2 Instruments used 11
2.2.1 X-band radar 11
2.2.2 Joss-Woldvogel Disdrometer 12
3.DATA AND METHODOLOGY 13
3.1DATA USED 13
3.2METHODOLOGY 14
3.2.1 Prepration of Data 14
3.2.2 Pre Processing of data for Machine Learning 16
3.2.3 Steps in Modelling 17
4.RESULTS AND DISCUSSIONS 22
4.1 The Evolution 22
4.2 Determining the reference structures for each cloud type and their frequency of occurrence during the period 28
4.3 Selection of features 31
4.4 Model Results 31
5. SUMMARY AND CONCLUSIONS 38
REFERENCES 40

LIST OF FIGURES

FIGURE 1(A)THE BLACK CIRCLES REPRESENT THE MANDHARDEV(MDV) WHERE THE X-BAND RADAR IS LOCATED AND HACPL.THE DOTTED CIRCLE SHOWS THE MAXIMUM RANGE (125 KM) OF X-BAND RADAR. RHI SCAN IS TAKEN ALONG THE DASH-DOTTED LINE (JHA K AMBUJ ET AL,METEOROLOGY AND ATMOSPHERIC PHYSICS,2022)(B)THE X-BAND AND KA-BAND RADAR SITUATED AT MANDHARDEV. 11
FIGURE 2. RHI PLOT OF EQUIVALENT REFLECTIVITY FACTOR. THE BLACK LINES SHOWS THE AREA OF STUDY. 13
FIGURE 3.GRIDDED(100×100) VERTICAL PROFILE OF DOPPLER VELOCITY FROM A PPI SCAN(85^{o} ELEVATION) 13
FIGURE 4(A)VERTICAL PROFILES OF REFLECTIVITY FROM RHI SCANS (B)DOPPLER VELOCITY OF HYDROMETEORS FROM PPI SCANS. ALL TIMES ARE IN UTC 23
FIGURE 5.(A) VPR OF A STRATIFORM CLOUD. THE 0,1,2,3,4 POINTS ARE USED TO INDICATE IMPORTANT MICROPHYSICAL PROCESSES.(B) DV PROFILE OF A STRATIFORM CLOUD 25
FIGURE 6. REFLECTIVITY PROFILE OF CLOUDS OVER HACPL 27
FIGURE 7. LOGICAL FLOW DIAGRAM FOR CLASSIFYING CLOUDS 28
FIGURE 8. IDEAL STRUCTURES THAT ARE USED AS REFERENCE FOR LABELLING CLOUDS 29
FIGURE 9. (A) CFAD PLOT OF SHALLOW CLOUDS. THE COLORMAP SHOWS THE NORMALISED NUMBER OF OCCURRENCES OF EACH REFLECTIVITY BINS AT A PARTICULAR HEIGHT. (B) CFADS OF STRATIFORM, TRANSITION AND CONVECTIVE CLOUDS. 30
FIGURE 10. HISTOGRAM DEPICITING THE PERCENTAGE OCCURRENCE OF EACH TYPE OF CLOUD IN THE PERIOD JUNE TO AUGUST 2018 OVER HACPL 31
FIGURE 11. THE CONFUSION MATRIX FOR A MULTICLASS CLASSIFICATION MODEL 33
FIGURE 12.CONFUSION MATRIX FOR THE (A) STRATIFORM MODEL (B) CONVECTIVE MODEL 34
FIGURE 13. FLOWCHART FOR THE ENSEMBLE BINARY CLASSIFICATION MODELS 35
FIGURE 14. STRATIFORM-CONVECTIVE MODEL 35
FIGURE 15. FINAL OUTPUT FROM THE 3 COMBINED ENSEMBLE BINARY MODEL 36
FIGURE 16. THE MODEL RUN ON A DATA OVER THE WHOLE RADAR RANGE TAKEN ON (A)JUNE 3 2018 AT 14:27 UTC. (B) JULY 15 2018 23:28 UTC.THE VPR IS TAKEN AVERAGING OVER 1 KM IN THE HORIZONTAL 37

LIST OF TABLES

TABLE1.TECHNICAL SPECIFICATIONS OF X-BAND DOPPLER WEATHER RADAR 12
TABLE2.TECHNICAL SPECIFICATIONS OF JOSS- WOLDVOGEL DISDROMETER 12
TABLE 3. HYPERPARAMETER TUNING OF MACHINE LEARNING MODELS 20
TABLE 4 DISDROMETER READINGS ON JUNE 26 OVER MDV 27
TABLE5 DISDROMETER READINGS ON JUNE 1 OVER HACPL 27
TABLE 6. OVERALL RESULTS OF BASELINE AND TUNED MODELS RUN ON THE TEST SETS. BAC INDICATES BALANCED ACCURACY SCORE. THE HIGHLIGHTED RESULTS ARE THE BEST RESULTS OBTAINED.ALL THE VALUES ARE WEIGHTED ACCORDING TO THE NUMBER OF TEST SAMPLES FOR EACH CLASS. 32
TABLE 7. CLASSIFICATION REPORT FOR EACH CLOUD. SUPPORT IS THE NUMBER OF TEST SAMPLES OF EACH CLOUD TYPE. ALL THE VALUES ARE AVERAGED FOR EACH CLASS BY WEIGHING WITH THE NUMBER OF SAMPLES. 32
TABLE 8. CLASSIFICATION REPORT FROM THE RESULTS OF THE BINARY MODELS. 36

ABSTRACT

Doppler weather radars uses an empirical relationship between reflectivity and rainfall rate given by $Z=AR^b$ where A and b are constants that varies with geographical locations and precipitating cloud systems. Classification of precipitating clouds is an important procedure for improving the rainfall estimation using Doppler weather radars. The types of precipitating clouds that are observed over WGs are shallow convective, stratiform , mixed stratiform convective(transition) and convective clouds. Identification of these cloud types has been done over the Western Ghats using machine learning model and an X-band Doppler weather radar situated at Mandhardev. A proper study about the characteristics of the precipitating clouds which are observed over the WGs has been done using case studies on the evolution of a mesoscale convective system. From this study reference structures of vertical profiles of reflectivity for each cloud type is obtained which is used to label the clouds for training and testing the machine learning model. The Light GBM algorithm is showing the best results out of the 7 models that was tested. 3 binary classification models are constructed using Light GBM which gives a test result of **0.88 accuracy and 0.91 F1-Score** which is better than a single multiclass classification model. The individual F1-scores of stratiform, transition and convective clouds are **0.97, 0.77 , 0.84** respectively while the shallow clouds are classified based on the condition that the cloud top height does not exceed above the melting layer.

1.INTRODUCTION

The Western Ghats of India acts as a barrier to the south–west monsoon clouds and influence the distribution of rainfall in the region.The Western Ghats (WG) is one of the heavy rainfall regions in India. WG receives a large amount of rainfall (~6000 mm) during the Indian summer monsoon (ISM) period**(**Nandargi S, Mulye S 2012**).**The moisture laden air brought by the LLJ from the Arabian sea is orographically forced to lift upward by the mountains of WGs causing heavy rainfall majorly on the windward side(west) thereby suppressing the rainfall over the leeward side(east). This results in the rain shadow region (east) of WGs resulting in droughts. The topography of the WGs makes it complex and challenging for the rainfall estimation as its height varies spatially in great extent.

An accurate knowledge of the amount of rainfall falling on a catchment area is of high importance in flood forecasting and warning systems (Smith,1993; Arnaud et al,2002).Doppler weather radars(DWR) operating at microwave frequencies conventionally uses an empirical relationship between radar reflectivity factor and rainfall rate of the form $Z=AR^b$ where A and b are constants. Many studies have shown that A and b of Z-R relationship vary with geographical locations and precipitating cloud systems (Fujiwara,1965 ,Schuur et al,2001). Therefore to improve the quality of rainfall estimation over WGs proper classification of precipitating clouds is of much importance. In order to do this we must go deeper into the evolution of a mesoscale convective system and the inherent cloud and rain microphysical and dynamical processes associated with it. This could also help in improving the understanding of rain microphysics from remote sensing devices.

The rainfall in the western ghats is contributed by cloud systems such as convective, transition(mixed stratiform-convective) ,stratiform and shallow clouds(Das et al 2017, **Konwar et al. 2012a;Maheskumar et al. 2014; Das et al. 2015;Deshpande et al. 2015**) among which shallow convective clouds contributes majorly to the total precipitation and number of occurrences in this region during ISM(Konwar et al 2014).Although there have been many studies on the precipitating clouds over WGs based on their dynamical, thermodynamical and cloud microphysical properties(Konwar et al 2014**;Kumar et al 2013;Utsav et al 2017; Sarker, 1966; Grossman and Durran, 1984; Oguraand Yoshizaki, 1988; Konwar et al., 2012a; Kumar et al., 2013;Maheskumar et al., 2014)** studies that have done an accurate cloud type classification are very few.

Several studies in the past have classified clouds into convective and stratiform based on observations from ground based disdrometers , rain gauges and radars.(**Houze et al 1973**) proposed a

method to classify the precipitation cloud types by using rain gauge measurements. When the precipitation intensity exceeds a certain threshold, the precipitation cloud is considered as convective cloud; otherwise, it is stratiform cloud. The limitation of this method is that it is easy to misjudge the precipitation area of convective cloud with weak precipitation intensity nearby. (Tokay and Short 1996) classified the precipitating cloud systems into convective and stratiform using the surface-based disdrometer observations. Testud et al. (2001) classified rain into convective and stratiform using aircraft measurements keeping a rain rate threshold value of $R=10mmhr^{-1}$. **Lavanya S and Kirankumar NVP (2021)** used datasets of Joss-Woldvogel disdrometer (JWD) observations and classified into convective, transition and stratiform rain based on the ratio of mass-weighted mean diameter (D_m) and rain rate (R) over Thumba.

The advantage of ground based radars is that it gives the vertical structure of a cloud which will help in understanding the microphysical processes undergoing within the storm. **Williams et al 1995** used a 915 MHz wind profiler to classify precipitating clouds as shallow convective,stratiform,mixed stratiform-convective and deep convective clouds by analysing the vertical structure of reflectivity, velocity and spectrum width. Radar reflectivity is conventionally used as a proxy for the intensity of storms. Similarly, **Das et al 2017** used simultaneous measurements from the Micro Rain Radar (MRR) and JWD and from the MRR reflectivity factor and fall-velocity profiles, the observed precipitation systems are classified into four categories: shallow-convective, convective, stratiform and mixed convective-stratiform. Both these studies made use of the phenomena of radar bright band which is a distinctive property of stratiform clouds that will be made use in this study too.

Usually reflectivity based algorithms are used which is based on some threshold values for classifying precipitating clouds. **Churchill and Houze 1984** used a 40 dBZ peak reflectivity factor threshold to determine convective echo cells and a 20-23 dBZ for stratiform precipitation. But several echoes that were in the range of 20-30 dBZ were classified as convective even though they still showed a definite bright band character. Later **Steiner et al 1995** modified this separation criteria for convective clouds by using Intensity, peakedness and surrounding area. Any grid point in the radar reflectivity field with reflectivity greater than 40 dBZ is automatically labelled as convective cloud since stratiform could practically never be this intense. If this condition is not met, then the reflectivity difference between grid point and mean background should be greater than a threshold to be labelled convective. Also, each grid point identified as convective by one of the above two criteria, all surrounding grid points within an intensity dependent convective radius around the point is also included in the convective area. **Biggerstaff and Listemaa 2000** took the three dimensional structure of radar reflectivity field. They noted two main misclassifications in

Steiner's algorithm: heavy stratiform rain, originally classified as convective, and the periphery of convective cores, originally classified as stratiform, were both reclassified by this modified algorithm. Here the vertical profile of reflectivity (VPR) and the altitude of the 0°C isotherm are applied to improve the algorithm.

The main drawback of threshold-based cloud-type classification is that a threshold over some situations may not be applicable for another one. In view of overcoming this challenge this study focuses on a machine learning based approach for the classification of precipitating clouds. There have been many studies in the past that have used machine learning algorithms for this task**(Anagnostou 2004 , Wang H et al 2018,Wang Y et al 2021, Ran Y et al 2021,Ghada W et al 2022). Anagnostou 2004** used a 3 layer(input,hidden,output) neural network approach radar reflectivity fields from 18 National Weather Service WSR-88D radars in the United States for the period December 1997 to October 1999.**Wang H et al 2018** used a deep learning approach for the classification of precipitating clouds using DWR. **Ghada W et al 2022** using radar reflectivity, doppler velocity and spectrum width provided a method to classify rain types based on MRR data.

This study aims at providing a classification method for convective, transition, stratiform and shallow convective clouds over the WGs using the radar reflectivity factor obtained from an X-band Doppler weather radar situated at Mandhardev, Maharashtra. The first section of the study focuses on the rain microphysics of these clouds and the evolution of a convective cell in a mesoscale convective system to trailing stratiform precipitation through two case studies on June 1st and June 26th 2018 at the High Altitude Cloud Physics Laboratory (HACPL) site situated at Mahabaleshwar and the X-band radar site respectively. We then classifies the 4 types of clouds based on the condition of bright band and determines the occurrence of each type of cloud during ISM. In the next section, we provides the vertical profile of reflectivity taken over the HACPL and radar site to 7 machine learning models which includes 3 baseline models and 4 tuned models which classifies them as stratiform, transition, convective and the results are compared.

2. Measurement site and Instruments used

2.1. Measurement site

The measurement sites, Mandhardev (18.04°N, 73.87°E, 1.3 km MSL) is a remote location in the hill top of Western Ghats, India and Mahabaleshwar (17.92°N, 73.6°E, 1.4 km above mean sea level) is located on the windward slope of the WGs. Mahabaleshwar is a hill station and it is located at about 50 km inland to the east of the Arabian Sea coast. It is one of the heavy rainfall regions(~6000 mm) during the ISM period. **Konwar et al 2014** proposed that the 850 hpa westerlies during ISM bring the giant size aerosols (like sea salt aerosols) which trigger the early onset of rainfall and due to the mountain slope, the cloud masses gets lifted and there is a forced condensation to form large cloud droplet (as it contains giant aerosols, acting as a cloud nuclei) which trigger the collision-coalescence process and finally leads to heavy rainfall amount in this region. Figure 1(a) shows the topographical map of the measurement sites.The distance between both the sites is nearly 26 km.

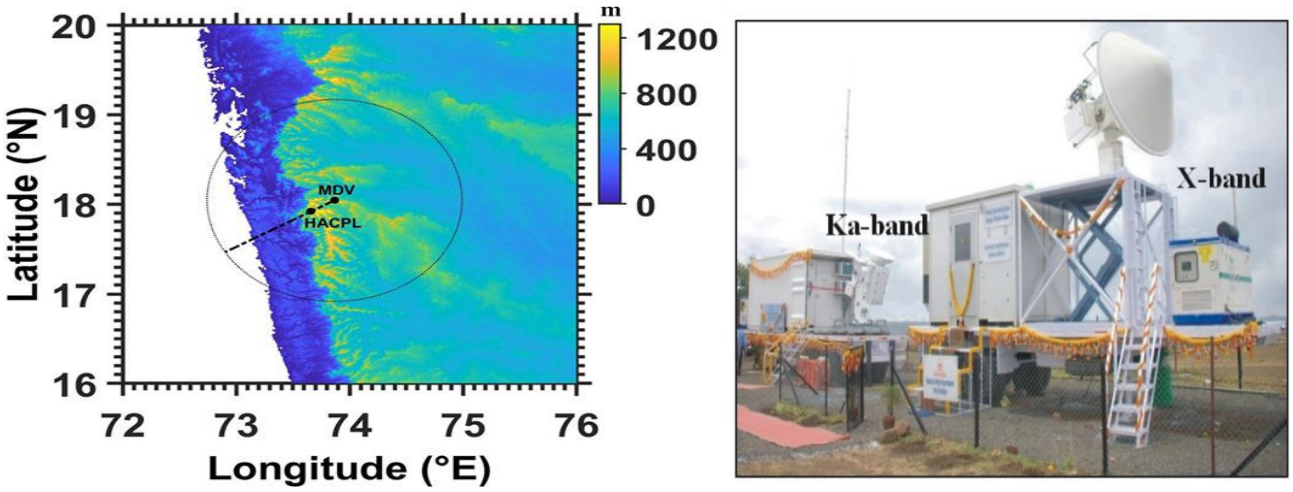


Figure 1(a)The black circles represent the mandhardev(MDV) where the X-band radar is located and HACPL.The dotted circle shows the maximum range (125 km) of X-band radar. RHI scan is taken along the dash-dotted line (Jha K Ambuj et al,Meteorology and Atmospheric Physics,2022)(b)The X-band and KA-band radar situated at Mandhardev.

2.2 Instruments used

2.2.1 X-band radar

Figure 1(b) shows the X-band radar in Mandhardev. It operates in a frequency of 9.53 GHz and a peak power of about 200 kW. The radar performs RHI scan every ~ 12 min, so oriented as to provide vertical cross-sections of precipitation over HACPL. The table shown below gives the technical specifications of the X-band radar.

Parameter	Value
Frequency (GHz)	9.535
Wavelength (cm)	3.14
Transmitter	Magnetron
Peak power (kW)	200
Pulse widths	0.8–2 μs
Beam width	0.97°
Antenna gain (dB)	44.3
Antenna diameter (m)	2.4
Cross-pol isolation (dB)	− 30
Minimum detectable signal	− 25 dBZ at 20 km

Table1.Technical specifications of X-band Doppler weather radar

2.2.2 Joss-Woldvogel Disdrometer

The used Joss-Waldvogel Disdrometer (JWD) is disdrometer RD-80,manufactured by Distromet Ltd., Switzerland. The sampling cross-section area of JWD is a styrofoam cone with an area of ~50 cm2. The JWD measures raindrops in 20 size intervals starting from0.3 to 5mm. The accuracy of measurement is about ±5% of the measured raindrop diameter. Table listed below shows the technical specifications of the JWD system.

Parameter	Specification
Range of drop diameter	0.3–5 mm
Sampling area	50 cm2
Accuracy	±5% of measured drop diameter
Resolution	127 size classes distributed exponentially over the range of drop diameters
Baud rate	9600 baud
Handshake	DCD and DTR signals
Dimensions of sensor	10 cm × 10 cm × 17 cm high
Dimensions of processor	12 cm × 26 cm × 27 cm deep
Operating temperature range	0–40 °C for processor; 0–50 °C for sensor

Table2.Technical specifications of Joss- Woldvogel Disdrometer

3.DATA AND METHODOLOGY

3.1DATA USED

For getting the equivalent reflectivity factor the **X-band radar**(9.53 GHz)is operated in the range-height indicator(RHI) mode by varying elevation angle of radar beam and maintaining a constant azimuth over the HACPL(239 degree) location at every ~12 minute intervals. We know that the HACPL is located at ~26 km from the radar site in the the southwest direction(239 degree) as shown in Figure 1(a).So we concentrate our study between 25.5 and 26.5 km such that we are taking the clouds over the HACPL region and 0 to 2 km for the radar site at mandhardev. An example of RHI plot of equivalent reflectivity factor and our area of study(MDV and HACPL) is shown in the Figure 2 below.

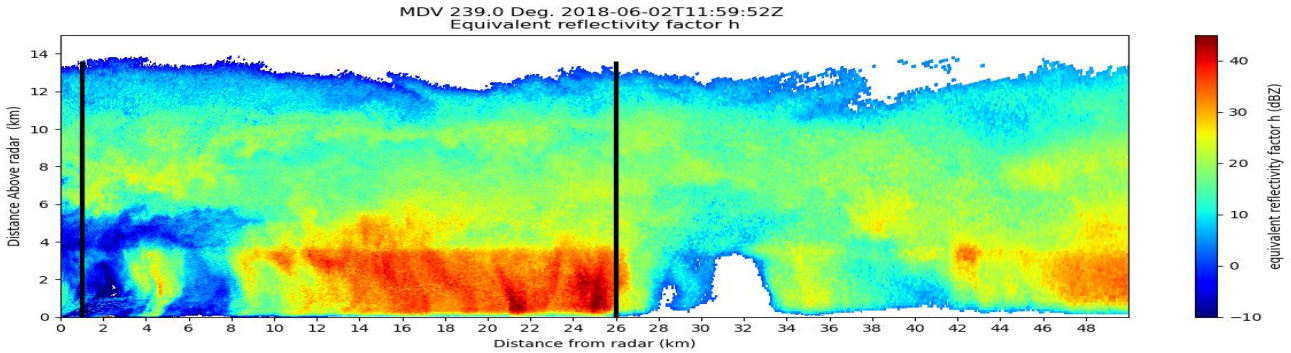


Figure 2. RHI plot of equivalent reflectivity factor. The black lines shows the area of study.

The Doppler velocity profile at radar site is obtained by operating the radar in plan-position indicator(PPI) mode at 85 degree elevation so that we are getting the terminal fall velocity of the hydrometeors. An example of a gridded Doppler velocity profile of a stratiform cloud is shown below.

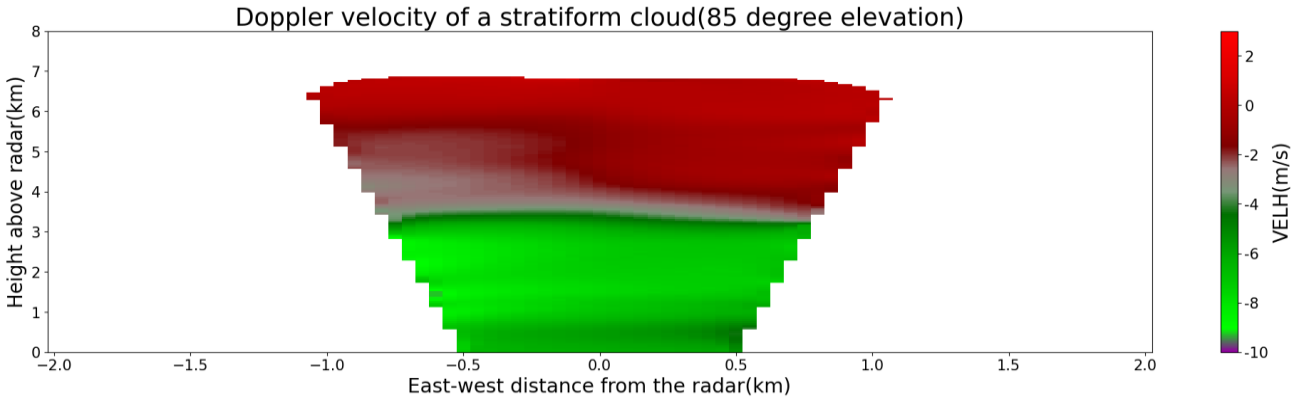


Figure 3.Gridded(100×100) vertical profile of Doppler velocity from a PPI scan(85° elevation)

We are using the equivalent reflectivity factor and Doppler velocity products in order to study the instensity of storm at every heights and the fall velocity profile of hydrometeors. The data taken for the study is in the period from June to August 2018.

The **Joss-Woldvogel disdrometer (JWD)** readings are taken from the MDV site and HACPL in order to calculate the rain intensity(RI) and mass weighted diameter(D_m). The JWD readings are at every 30 second interval which is converted to 1 minute interval for this study. Here we are taking the mean value of RI for every 12 minute intervals corresponding to the radar observations.

3.2METHODOLOGY

3.2.1 Prepration of Data

The horizontal equivalent reflectivity factor(dBZ) of RHI scan from June 1st to August 31st 2018 is taken for the purpose of this study. The data is given in polar coordinates(elevation and range) which is converted to cartesian coordinates(distance and height) and interpolated into 100X100 m grids in the horizontal and vertical. Then the mean vertical profile of reflectivity is taken over the HACPL and radar site by calculating mean reflectivity between 0.5km east and 0.5km west of HACPL site and upto 2km east of the radar site for each height. The **radar reflectivity factor is used as a proxy for the intensity of storms**. The equation for radar reflectivity factor is given by,

$$dBZ = 10 \times log_{10}(Z_e) \quad (1)$$

where Z_e is reflectivity given by,

$$Z_e = \int_0^{Dmax} |K|^2 N_0 e^{-\Lambda D} D^6 dD \quad (2)$$

Here K is the dielectric constant of the target and 'D' is the diameter of droplet and the remaining term is the Marshall Palmer drop size distribution.

For further analysis of the rain microphysics we take the Doppler velocity H(m/s) by taking the 85° elevation of the PPI scan. Since we are getting the relative velocity of hydrometeors with respect to radar using this variable , operating at 85° elevation will give the velocity of hydrometeors relative to radar **that resembles its actual fall velocity**. This can be done just over the radar site.

For each of the 1-min samples, the integral rainfall parameters like rain intensity (RI) and mass weighted mean diameter (D_m) are calculated over 12 min intervals. RI (mm h^{-1}) is given by

$$RI = \frac{6\pi}{10}\frac{1}{A\times t}\sum_{i=1}^{20}(n_i D_i^3) \quad (3)$$

where A (m2) is disdrometer measurement area(0.005 m^2), t (seconds)is data averaging time interval, n_i is number of drops measured in size class i, and D_i (mm) is the mean diameter of drops in size class i. Drop number concentration $N(D_i)$ (m^{-3} mm^{-1}) corresponding to size class i is given by

$$N(D_i) = \frac{n_i}{A\cdot t\cdot V(D_i)\cdot \Delta D_i} \quad (4)$$

where $N(D_i)$ is the number of drops of diameter D_i (mm) per unit volume corresponding to i^{th} diameter class,, $V(D_i)$ (m s−1) is the (theoretical) terminal velocity of drop corresponding to class *i* (**Gunn and Kinzer 1949**) and ΔDi (mm) is the diameter interval of size class *i*.

DSD moment of order n (M_n) is defined as,

$$M_n = \sum_{i=1}^{i=20} N(D_i)D_i^n \Delta D_i \quad (5)$$

and D_m (mm) is given by (**Bringi and Chandrasekar 2001**),

$$D_m = \frac{M_4}{M_3} \quad (6)$$

Using reflectivity, doppler velocity, RI and D_m we analyses the vertical strucuture and drop size distribution of the precipitating clouds over HACPL and MDV site. From the case studies done on June 1st 2018 over HACPL site and June 26th 2018 over the MDV site we arrives at certain conditions based on reflectivity profiles (which will be discussed in detail on the next section)that enables us to classify these clouds into shallow,stratiform,mixed stratiform-convective(transition) and convective clouds. Using these conditions we plots the CFAD's i.e Contoured Frequency by Altitude Diagrams (**Yuter and Houze 1995**) for each type of clouds. We have only included cloud profiles where the reflectivity values exceeds at least 5 dBZ at any level from 500 m to 2 km above the radar. CFADs have been constructed with 5 dBZ reflectivity bins at 0.1 km height interval between 0.5 and 10 km. The number of occurrences at each level is normalized by the maximum number considering all reflectivity bins among all vertical levels. A grid is considered only if the normalized frequency is at least 10%.

3.2.2 Pre Processing of data for Machine Learning

The reflectivity profiles of each cloud obtained over HACPL is taken one by one and **manually labelled as 0,1,2 and 3** for **Stratiform, Transition ,Convective and Shallow clouds** respectively in order to provide **as training and testing inputs and targets** for **supervised machine learning**. The inputs will be the individual cloud reflectivity profiles and the targets are the labels indicating it as one of these 4 types of clouds. Since we have more than 2 classes to classify this is a **Multiclass classification problem**. The labelling is done on the basis of understanding we have about the vertical reflectivity profile of individual clouds and the different microphysical processes associated with each type of cloud.

The input is fed to the models as tabular data with each row representing a cloud profile and each column representing the height levels(gridded at 100 m). Corresponding to each rows(clouds) we have its target classes(0,1,2,3).The data points is known as **features** to the model. In the next step, to reduce the complexity of data given to the models we select only certain features that are most important for the classification task. This is known as **feature engineering** in machine learning. By doing this, we select **40 features(here height levels)** from each cloud vertical profiles of reflectivity. The features selected and its explanation will be given in the next section. Finally the tabular data is converted to exponential space in order to handle the 'Nan' values in the data. This is in the sense that since the exponential of no number can be zero, after taking the exponential we put zero wherever there where 'Nan' values.

The data used for the study includes a total of 1658 samples consisting of Stratiform, Transition and Convective clouds with 40 features in each sample. The shallow clouds are not included because the condition for determining a shallow cloud is just based on the height of cloud that is these clouds does not extend above the melting layer. The height of melting layer is taken to be 5.3 km(±0.19 km) from the study of **M.C.R Kalapureddy et al 2022** over the Western India during the peak ISM(July and August) which is also in close agreement with **Das et al 2011.** There are totally **898 samples of Stratiform clouds, 493 convective clouds and 267 samples of Mixed Stratiform-Convective clouds** taken from June to August 2018 for the study. The number of samples also shows the frequency of occurrence of these clouds over HACPL since stratiform is seen the most number of times after shallow clouds. The frequency of occurrence of shallow clouds is in the order of thousands which makes about 90 percentage of total number of precipitating clouds. This makes it easy for identifying shallow clouds because even some outliers(i.e clouds that reaches just upto the

melting layer so that it is complex to call them plainly shallow) won't make much of a difference to the model since most of them doesn't reach the 0^{o} isotherm.

3.2.3 Steps in Modelling

a) Training and Test data

At first the whole tabular data is taken(1658 rows and 40 columns) and is split into training and testing datasets. The test data is 10% and training data is 90% of the total dataset. The test data is split in such a way that it keeps the same percentages of classes in the split. This splitting is done so that the training data can be used **to fit the model** which is the data given to the model to learn and then the **fitted model can be tested** using the test dataset which is a new set of data the model is not familiar with. Training a model involves using an algorithm to determine model parameters (e.g., weights) or other logic to map inputs (the reflectivity profile) to a target (type of cloud).

b) SMOTE Oversampling

Now, we have 1492 samples for training the model and 166 samples for testing(0-90,1-27,2-49).The training data includes 808 stratiform clouds,240 transition clouds and 444 convective clouds. Here we can see that one class has more number of samples compared to the other two classes. This makes this an Imbalanced data classification problem where one class is the majority. An imbalanced data given as input causes a model to be biased towards the majority class by making it learn more about the features of those classes thereby ignoring the minority classes. This makes the model wrongly predict the minority class as the majority. A machine learning model expects inputs to have the same number of samples. Often real-world data sets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples.

In order to tackle this challenge, we uses an oversampling technique called Synthetic Minority Oversampling Technique **(Chawla et al 2002)** or simply **SMOTE.** This is a data augmentation technique. SMOTE first selects a minority class instance 'a' at random and finds its k nearest minority class neighbours. The synthetic instance is then created by choosing one of the k nearest neighbours 'b' at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b(**Imbalanced Learning: Foundations, Algorithms, and Applications,2013).** By this way we will generate more number of samples for transition and convective clouds and makes it equal to the number of stratiform clouds thereby making the data balanced.

c) Validation set and Loss function

After oversampling, we have 808 samples of each stratiform,transition and convective clouds. From this we split 10% of data as validation set. Validation data are used with each model developed in training, and the prediction errors are calculated. The validation data may be used several times to build the final model. The error in predictions are calculated using a loss function that can vary from model to model. In a multi-classification problem, we define the logarithmic loss function F as, in terms of the logarithmic loss function per label F_i as:

$$\mathrm{F} = \frac{-1}{N}\sum_i^N \sum_j^M y_{ij} \cdot \mathrm{Ln}(p_{ij})) = \sum_j^M (\frac{-1}{N}\sum_i^N y_{ij} \cdot \mathrm{Ln}(p_{ij})) = \sum_j^M F_i \qquad (7)$$

where N is the number of instances, M is the number of different labels, y_{ij} is the binary variable with the expected labels and p_{ij} is the classification probability output by the classifier for the i-instance and the j-label. The ultimate aim of a machine learning model is to reduce this cost/loss function. If the training loss is zero, which means the model studies the training data perfectly then it will not be able to generalize to a new dataset which here is the testing set. This is known as **Overfitting** of the model. Validating and testing our supervised machine learning models is essential to ensure that they generalize well.

d)Feature scaling

The next step is feature scaling which is also one of the feature engineering techniques. Many machine learning algorithms prefer or perform better when numerical input variables and even output variables in the case of regression have a standard probability distribution, such as a Gaussian (normal) or a uniform distribution. For this we make use of a **Quantile transform Scaler** that will map a variable's probability distribution to another probability distribution. The quantile function ranks or smooths out the relationship between observations and can be mapped onto other distributions, such as the uniform or normal distribution .The transformation can be applied to each numeric input variable in the training dataset and then provided as input to a machine learning model to learn a predictive modelling task

e)Hyper-parameter Tuning

We use hyperparameters to calculate the model parameters such as weights and bias. Different hyperparameter values produce different model parameter values for a given data set. Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning

algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

Here we are using a total of 7 ML models .Three of them are baseline models which includes **Logistic regression, Naïve Bayes and Decision tree** . Remaining 4 are complex ML models which requires hyperparameter tuning. They are **xgboost, light gbm, support vector machine and randomforest**. All the models are tuned using Optuna which is an automatic hyperparameter optimization software framework, particularly designed for machine learning. The lower and upper bounds of each hyperparameter we tuned for the models using optuna is shown in the table below.

Model	Hyperparameters	Lower	Upper	Values
xgboost	booster	-	-	dart
	Lambda	1e-8	1.0	7.856932170615428e-06
	Alpha	1e-8	1.0	8.618699602038801e-07
	max_depth	7	15	8
	Eta	1e-8	1.0	0.17417073737851846
	Gamma	1e-8	1.0	5.97391529967321e-08
	grow_policy	-	-	lossguide
	sample_type	-	-	uniform
	normalize_type	-	-	forest
	rate_drop	1e-8	1.0	0.28041784694385835
	skip_drop	1e-8	1.0	0.0003899171063467437
Light gbm	lambda_l1	1e-8	10	2.680331760378277e-07
	lambda_l2	1e-8	10	2.468977537167554e-07
	num_leaves	2	256	99
	feature_fraction	0.4	1	0.7859744578744218
	bagging_freq	1	7	3
	min_child_samples	5	100	15
	bagging_fraction	0.4	1	0.8944405940302624
svm	C	100	1000	499
	Kernel	-	-	rbf

	degree	1	6	3
Random Forest	n_estimators	50	1000	636
	max_depth	4	50	50
	min_samples_split	1	150	6
	min_samples_leaf	1	60	1

Table 3. Hyperparameter Tuning of Machine learning models

f) Evaluating a model

After training the model using the tuned hyperparameters we need to evaluate the performance of the model. For this we make use of an evaluation metric. To understand the significance of an evaluation metric we must discuss about the outputs of a machine learning algorithm. The ouputs of a machine learning model can be mapped into one of the following categories:

1. True positives(**TP**): A true positive is an outcome where the model correctly predicts the positive class.

   Eg: A cloud which is stratiform(positive) in nature and is classified as stratiform(positive) by the model.

2. True negative(**TN**): a true negative is an outcome where the model correctly predicts the negative class.

   Eg: A cloud which is not stratiform(negative) in nature and is classified as not stratiform(negative) i.e convective or transition by the model.

3. False positive(**FP**): A false positive is an outcome where the model incorrectly predicts the positive class.

   Eg: A cloud which is not stratiform(negative) in nature and is classified as stratiform(positive) by the model.

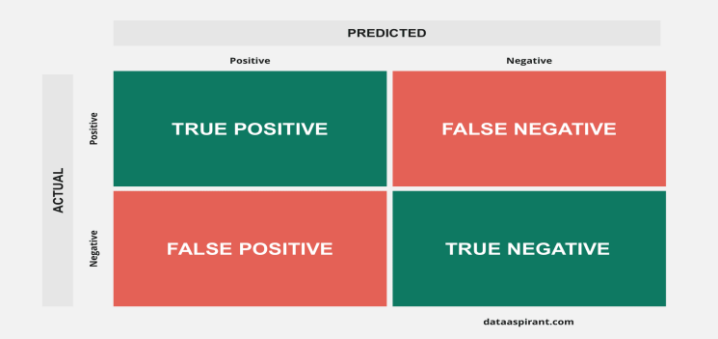
4. False negative(**FN**): A false negative is an outcome where the model incorrectly predicts the negative class.

   Eg: A cloud which is stratiform(positive) in nature and is classified as not stratiform(negative) by the model.

More number of True positives and True negatives makes a classification model better while False positives and False negatives which is caused due to misclassification reduces the performance of the model. The evaluation metrics we use are:

I.Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



II. Accuracy

Accuracy simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions. Accuracy may not be a good measure if the dataset is not balanced. Due to this we use Balanced Accuracy score. This metric is particularly useful when the two classes are imbalanced.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} = \frac{TPR + TNR}{2}$$

Where

$$True\ positive\ rate(TPR) = \frac{TP}{TP + FN} \quad True\ negative\ rate(TNR) = \frac{TN}{TN + FP}$$

III.Precision

It is a measure of correctness that is achieved in true prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives.

$$Precision = \frac{TP}{TP + FP}$$

IV.Recall

It is a measure of actual observations which are predicted correctly, i.e. how many observations of positive class are actually predicted as positive. Recall is a valid choice of evaluation metric when we want to capture as many positives as possible. Recall is a useful metric in cases where False Negative is higher concern than False Positive. This is same as sensitivity.

$$Recall = \frac{TP}{TP + FN}$$

V. F1-score

The F1 score is the harmonic mean of precision and recall. F1-Score is used when the False Negatives and False Positives are important. F1-Score is a better metric for Imbalanced Data.

$$F1\_Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)}$$

The test data is passed to the trained model and the model predicts the output. The output is compared with the true output(the ones which we have manually labelled) using these evaluation metrics. The higher the values of these metrics the better our model.

4.RESULTS AND DISCUSSIONS

4.1 The Evolution

In order to classify the precipitating clouds in WGs into shallow, stratiform, mixed stratiform-convective(transition) and convective clouds we first need to understand how to distinguish between these clouds based on their physical properties. We know that in the tropics stratiform precipitation occurs within a mesoscale convective system due to older convective cells i.e weakening of convection(**Houze 1997**). Analysing such a case will help us to identify both convective and stratiform clouds along with an intermediary stage where the convection is weakening and turning into stratiform which is basically a mixed stratiform-convective precipitation.

We used the X-band radar on RHI mode and observed such a case of evolution on **June 26th** 2018 from **11:25 UTC to 12:54 UTC** near the radar site. The vertical profile of reflectivity or VPR(from

RHI) and doppler velocity(from PPI 85^{o} elevation) profile during that time is analysed and the results are shown below.

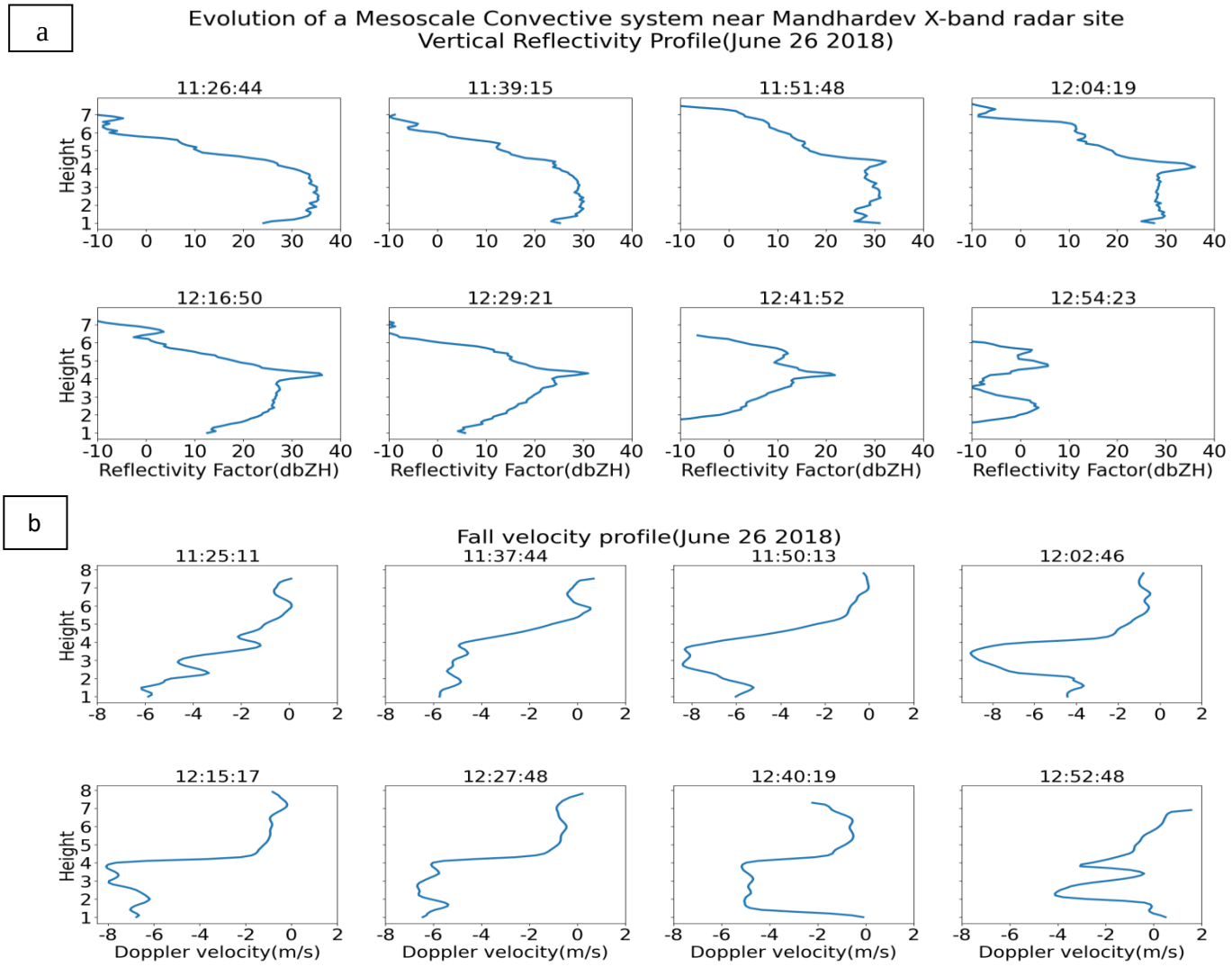


Figure 4(a)Vertical profiles of reflectivity from RHI scans (b)Doppler velocity of hydrometeors from PPI scans. All times are in UTC

From Figure 4 (a) we can see that the system is evolving as its vertical profile of reflectivity is having significant changes in its structure with time. At 11:26 we can see the reflectivity of cloud increasing from cloud top upto ~4km above radar(which is 1.3 km above MSL) and then not varying that much upto 0.5 km. We are not considering below 0.5 km since there are many orographic influences due to WGs. At 11:51 we can see a change in the structure as there is a sheer increase in reflectivity from 5 km to 4km. Below 4km the reflectivity values are decreasing slightly then again increasing. From 12:04 to 12:41 it is clear that there is a local maxima of reflectivity between 3-5 km and the peak value of reflectivity is decreasing with time. The reflectivity increase is in the order of 7-10 dBZ. We can also see that the reflectivity values below 4km is reducing on every height levels. At 12:54 the values are very low which indicates the cloud should be moving away.

Since the PPI scans extends for 12 minutes and the next would be an RHI scan, there must be a 12 minute gap between the outputs of both scans. But since we are taking 85^{th} elevation which is one of the final sweeps of the PPI scan the time difference between both these scans are so close that they can be used to define the same system of cloud. Figure 4 (b) shows the change in doppler velocity during these intervals. Since the radar is almost vertically pointing(85^{o}) this can be inferred as the vertical velocity(fall velocity) of hydrometeors. The **negative velocity indicates** that the hydrometeors are **coming towards the radar(**or simply falling**)** and positive indicates it is moving away from the radar(rising upwards). At 11:25 we can see that the vertical velocity is undulating between 3 and 5 km. All particles are still falling but the fall velocity is low and there is a tendency to rise upwards too. From 11:37 to 11:50 we can see the pattern changing as there is large increase in the magnitude of velocity from 5-4 km. Below 3 km the magnitude is slightly decreasing(trying to rise upwards). 12:02 to 12:40 depicts a steep and flat increase in the magnitude of downward fall velocity in the order of 5-6 m/s. Here we can see the tendency to rise decreases as water droplets are maintaining an almost constant fall velocity below 4km. The last plot shows water droplets trying to rise in lower levels which could be due to the influence of other clouds.

The reason for this phenomena can be attributed to the different cloud microphysical processes that happens in convective and stratiform clouds. **Chapter 6 and 9 of Cloud Dynamics** by **Robert.A. Houze Jr**(Department of Atmospheric Sciences, University of Washington, Seattle, Washington) clearly explains these processes. First we start by explaining stratiform precipitation.

Stratiform Precipitation

Stratiform precipitation is defined as a precipitation process in which the **vertical air motion is small compared to the fall velocity of the ice crystals and snow**. They generally occurs in nimbostratus clouds. The condition for the vertical velocity inside a stratiform cloud is,

$$|w| < V_{ice} \tag{8}$$

where |w| is the magnitude of vertical velocity and V_{ice} is the terminal fall velocity of ice/snow particles. The ice particles in the upper levels of nimbostratus must fall; they cannot be suspended or

carried aloft by the air motions as they grow. This can be explained in detail using the VPR of a stratiform cloud shown below.

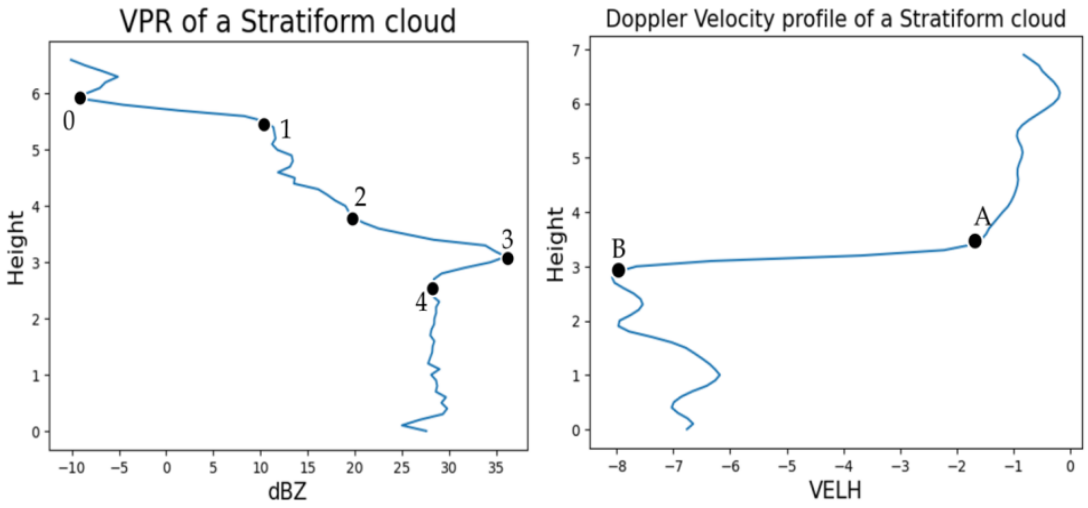


Figure 5.(a) VPR of a stratiform cloud. The 0,1,2,3,4 points are used to indicate important microphysical processes.(b) DV profile of a stratiform cloud

- The zone from 0 to 1 is associated primarily with ice particles *growing by deposition* which is the slowest microphysical growth mode(1-3 h). So the reflectivity which is proportional to D^6 (equation 2)does increase but slowly.

- Between 1 and 2 the particles continue to grow by *deposition , aggregation and some riming* and as it approaches the melting layer produces very large particles thereby increasing reflectivity.

- Between 2 and 4 **the radar bright band** is visible below the 0^o isotherm because snow/ice *melts to rain* showing a higher increase in reflectivity(7-10dBZ or more)becoming the **peak reflectivity** of that cloud at 3. Due to difference in dielectric constant of ice and water(0.197 for ice and 0.93 for water drops) the Z will increase by a factor 0f ~5(0.93/0.197) as it is proportional to $|K|^2$(equation 2).

- The sharp dropoff in Z from 3 to 4 is due to **2 reasons.** If the melting is completed at 3 the *water particles breaks up* into smaller rain drops thereby decreasing drop size and reflectivity($Z \sim D^6$).

- Second, the fall speed of particles suddenly increase from 1 to 3 m/s to 5-10 m/s for raindrops due to increase in drop size because of melting of ice to water. Since reflectivity is measured per volume of air, the mean concentration of rainwater must decrease sharply from 3 to 4 due

to the increase in downward flux of precipitation mass from 3 to 4(i.e *increase in terminal fall velocity* as seen in Figure 5(b)).

- Below 4, Z is not increasing and staying almost constant. This is due to the *evaporation* that happens below the melting layer because of insufficient updrafts which triggers *growth of droplets by collection*(*collision-coalescence*) which will increase the reflectivity.

What this implies about the dynamics of nimbostratus is that there must be a widespread gentle uplift of the air throughout the region of the cloud above the zero degree isotherm. This ascent must be strong enough to supply vigorous growth by vapor diffusion and some riming, but weak enough to allow sedimentation of ice particles.

So the microphysical processes such as *growth by deposition, aggregation , riming , melting , collision-coalescence , and increase in terminal fall velocity* are able to explain the vertical reflectivity profile of precipitating clouds which helps us to separate them based on their structure.

For a convective cloud(11:26 Figure 4(a)) where there are very strong updrafts below and above the melting layer, growth by deposition, aggregation, riming and collision coalescence below the melting layer are dominant processes while in order for melting process to be dominant it must satisfy equation(8) but the vertical velocity inside a convective cloud is much higher hence the particles are not allowed to sediment. Hence the phenomena of **radar bright band is absent in a convective cloud** most of the times.

For a mixed stratiform-convective cloud(11:51 Figure 4(a)) we can see that the updrafts are stronger below the melting layer so there is weak updrafts in the melting layer causing the particles to sediment but the breaking up of water droplets is not causing much of a decrease in Z as the droplets continue to grow by collection and Z increases. Hence we can say **that bright band signal is visible but not prominent and there is an embedded convection below the melting layer** for a transition cloud.

For a stratiform cloud(Figure 5 (a)) the updrafts are very weak below the melting layer but strong enough to allow ice/snow to melt. Below the melting layer since there updrafts are weaker growth by collection would be less so Z will drop due to evaporation or stay constant if there is no evaporation. Hence we get a **strong prominent radar bright band and lower reflectivity below the melting layer** for a stratiform cloud.

To further reinforce this we make use of the Rain Intensity(RI) reading from the Joss-Woldvogel disdrometer (JWD) in the same period of time. The mass weighted diameter(D_m) of droplets are calculated using the **equation(4,5 and 6).** The table below shows the RI and D_m values.

Time(UTC)	**11:26:44**	**11:39:15**	**11:51:48**	**12:04:19**	**12:16:50**	**12:29:21**	**12:41:52**	**12:54:23**
RI(mm h^{-1})	24.14	10.56	22.22	6.19	0.20	0.01	0.009	0.0004
D_m (mm)	1.57	1.40	1.59	1.57	1.12	0.47	0.10	0.07

Table 4 Disdrometer readings on June 26 over MDV

Table 4 shows that the rainfall intensity and mass weighted diameter is higher for the first three observations because these are convective and mixed stratiform-convective precipitation which has larger drop size distribution producing heavy rainfall. The remaining readings from 12:04 can be called as stratiform as RI and D_m are very low compared to the other causing widespread weaker precipitation.A similar case like this is observed on **June 1st 2018 over the HACPL** site and the RI and D_m is also calculated. The results are shown below.

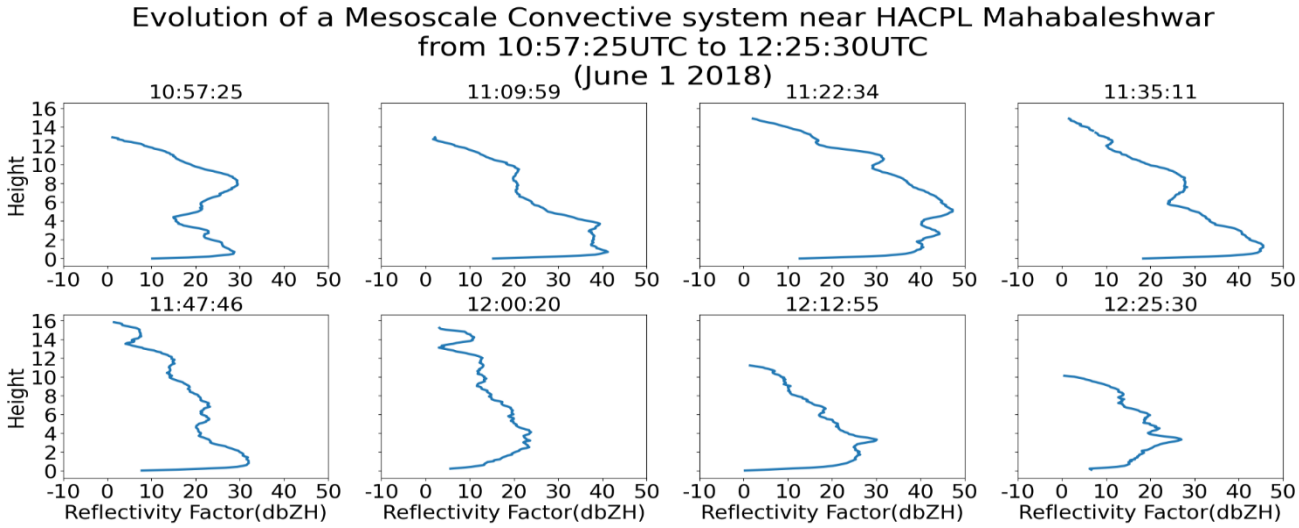


Figure 6. Reflectivity profile of clouds over HACPL

Time(UTC)	**10:57:25**	**11:09:59**	**11:22:34**	**11:35:11**	**11:47:46**	**12:00:20**	**12:12:55**	**12:25:30**
RI(mm h^{-1})	0.73	2.12	51.25	6.57	0.07	0.02	0.3	0.03
D_m (mm)	1.37	1.377	2.52	1.63	0.77	0.74	1.42	0.9

Table5 Disdrometer readings on June 1 over HACPL

This results are also in accordance with what we said above as we can see the high convective clouds weakening and developing a bright band forming stratiform precipitation. The RI (Table 3) shows

that at 11:22 we are getting heavy precipitation on the ground which also agrees with the reflectivity profile that exceeds 40 dBZ near the ground and having a very broader distribution. The D_m shows 2.52 mm which means very large droplets are falling to the ground. In the next steps we can see RI and D_m decreasing indicating stratiform precipitation.

4.2 Determining the reference structures for each cloud type and their frequency of occurrence during the period

The whole purpose of the evolution case studies was to distinguish between each cloud type over the WGs so that we get an ideal structure of these clouds in order to use as a reference for labelling the clouds to provide as target outputs for supervised machine learning models. Now we have arrived at the criteria for seperating these clouds. The type of cloud that was missing in the case study was shallow clouds. They are characterised using the presence of hydrometeors above and below the melting layer(**Das et al 2017**). A logical flow diagram depicting the classification criteria is shown below.

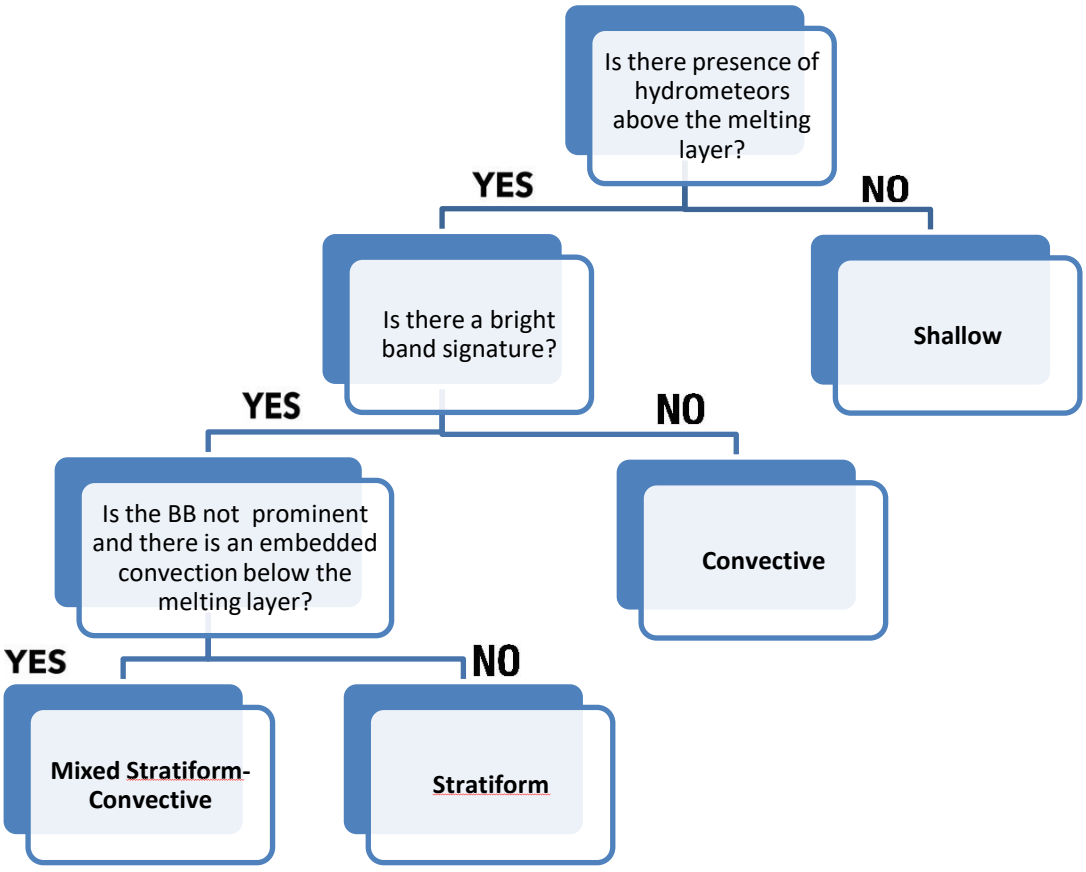


Figure 7. Logical flow diagram for classifying clouds

By using this criteria we determines an ideal structure for each type of cloud shown below.

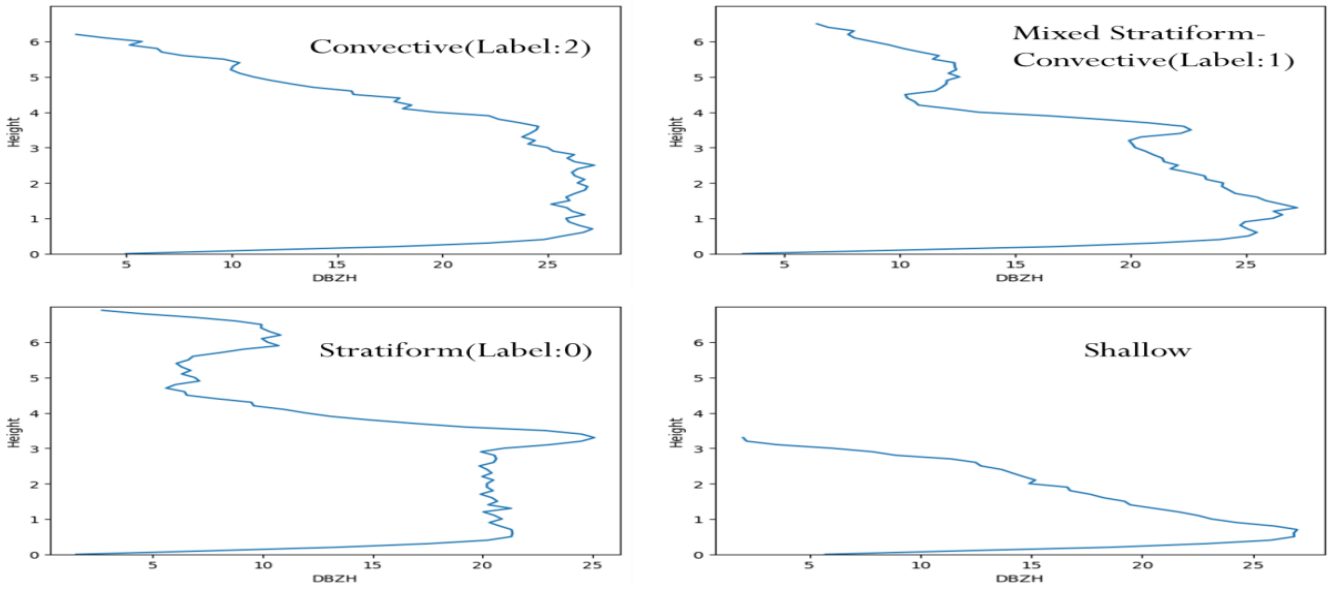
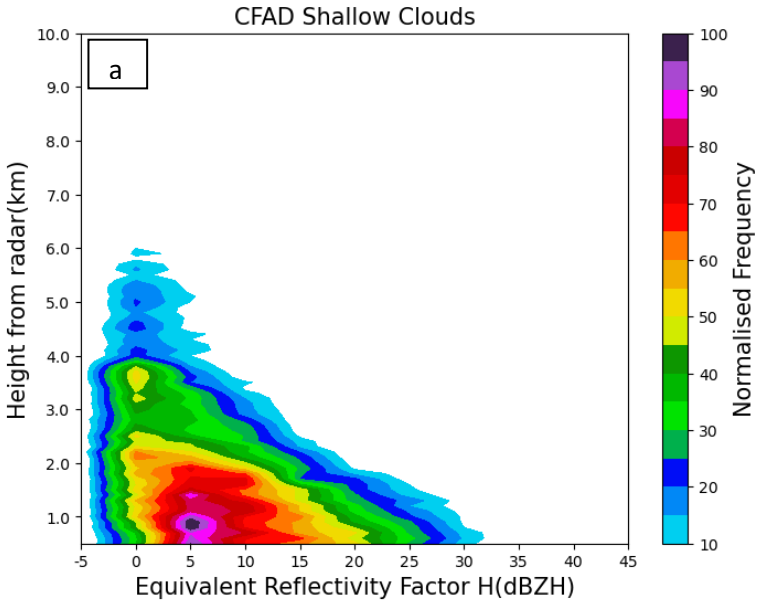


Figure 8. Ideal structures that are used as reference for labelling clouds

The CFADs(Contoured Frequency by Altitude Diagram)for each cloud type has been plotted to understand the relation between reflectivity and height for all the clouds from June to August 2018 and the occurrence of each type of cloud over the HACPL for the same period is found out.



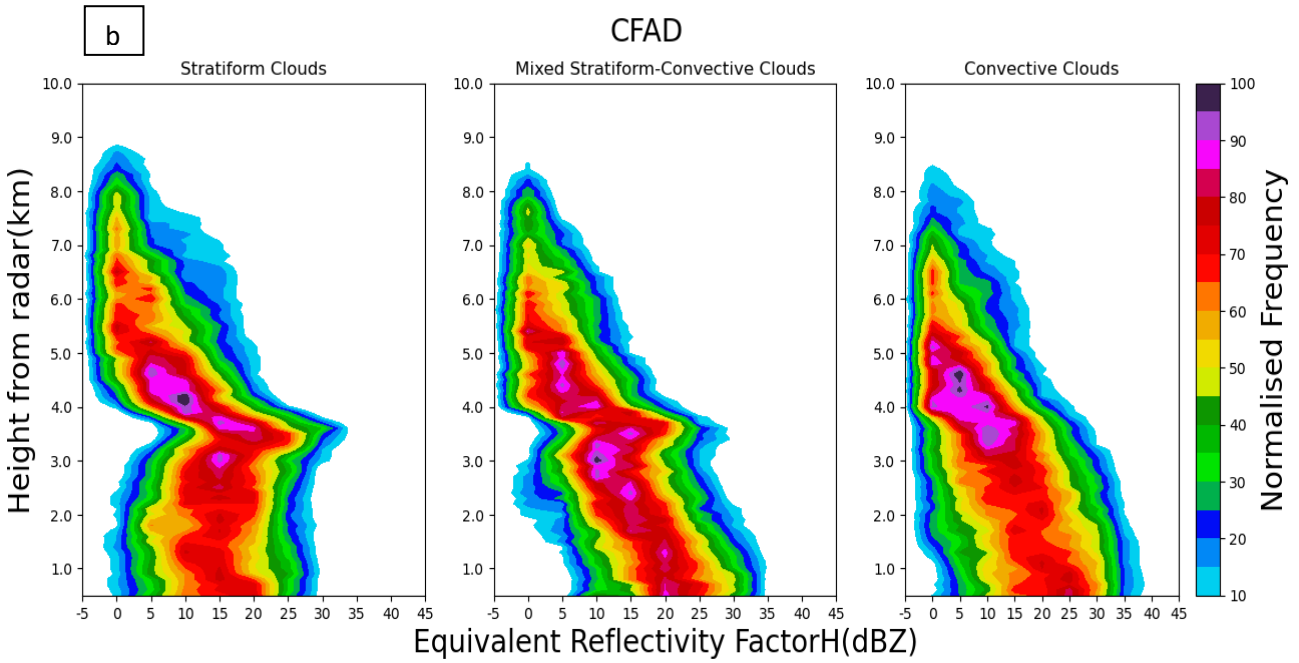


Figure 9. (a) CFAD plot of shallow clouds. The colormap shows the normalised number of occurrences of each reflectivity bins at a particular height. (b) CFADs of stratiform, transition and convective clouds.

Analysing Figure 9 (a), we can see a mode of 5dBZ at ~1km indicating the weaker precipitation of shallow clouds. Figure 9 (b) compares the other three type of clouds. For a stratiform we can see two modes one at 10 dBZ between 4-5 km which is almost the same height as the zero degree isotherm

and the other at 15-20 dBZ and ~3.5 km showing the peak of reflectivity due to melting. Then we can see the reflectivity decreasing showing another mode at 15 dBZ at nearly 3 km due to the breakup process. The reflectivity below the melting layer does not exceed 30 dBZ indicating the evaporation and absence of collision coalescence due to insufficient updrafts.

The mixed stratiform-convective clouds shows a similar structure but the increase in reflectivity is weaker between 4 and 3 km and increase more than the peak value below melting layer due to the embedded convection and updrafts and growth by collection process respectively. The convective clouds on the other hand does not show a local maxima and reflectivity values near the ground exceeds 35 dBZ and has a broader distribution due to vigorous updrafts and collision coalescence.

The occurrence of these cloud types during this period is shown in the figure below. It is clear that shallow clouds are dominating over HACPL with almost 84 % of the total number of clouds followed by stratiform with nearly 6 % , convective 5.3 % and mixed phase clouds with 4.7 % .

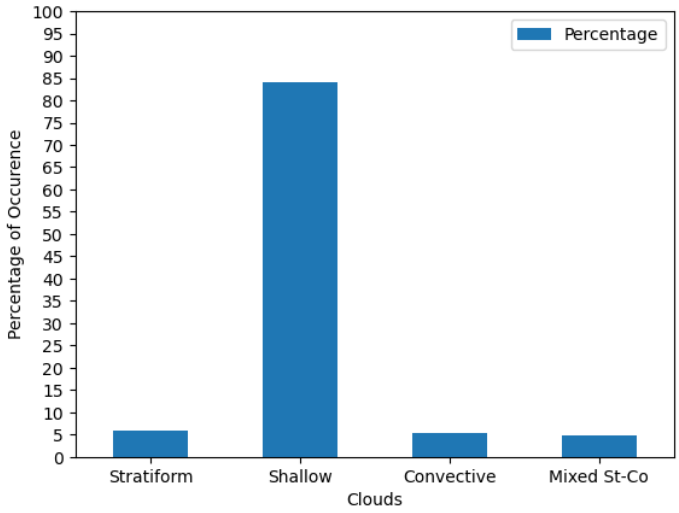


Figure 10. Histogram depiciting the percentage occurrence of each type of cloud in the period June to August 2018 over HACPL

4.3 Selection of features

The number of features given to a ML model as input plays an important role in the performance of the model. So instead of providing data at every heights we choose certain height levels that can define a specific type of cloud. The selected height levels are:

a) **1-2 km** – These levels **represents the ground or region below melting layer**. Since Convective and Transition clouds are defined to have higher reflectivity below the melting layer as can be seen on the CFADs.

b) **3-5 km** – These levels are chosen specifically to **detect the radar bright band** which will identify a stratiform cloud and also transition clouds.

c) 6-7 km – This represents **the region above the melting layer.** Convective clouds will have a broader distribution of reflectivity at these levels.

Since the data is 100 m gridded in the vertical this makes it a total of **40 features.** Every input will have these 40 features to train the model.

4.4 Model Results

A total of 1685 samples consisting of 898 stratiform, 267 transition and 493 convective clouds over the HACPL site is used to train the model and 10% of the data is used for testing the model. The test set consists of 166 clouds with 90 stratiform, 27 transition and 49 convective clouds. The overall

results obtained on the test set from **7 machine learning models** that includes 3 baseline and 4 hyperparameter tuned models are shown in the table below.

	Model	**BAC**	**Precision**	**Recall**	**F1-Score**
Baseline	Decision tree	0.64	0.72	0.69	0.7
	Naïve Bayes	0.53	0.62	0.57	0.59
	Log Reg	0.66	0.72	0.7	0.7
Tuned	xgboost	0.78	0.83	0.83	0.82
	SVM	0.66	0.75	0.75	0.75
	RandomForest	0.73	0.79	0.78	0.78
	Light GBM	**0.81**	**0.84**	**0.84**	**0.84**

Table 6. Overall results of baseline and tuned models run on the test sets. BAC indicates Balanced accuracy score. The highlighted results are the best results obtained.All the values are weighted according to the number of test samples for each class.

From Table 6 we can see that **Tuned models performs better than baseline models** for this classification problem. **Light GBM**(Light Gradient Boosting Machine) gives the best results out of the 4 tuned models with a BAC of 81% followed by xgboost(Extreme Gradient Boosting). Therefore we take Light GBM model for further analysis.

Given below is the classification report of the Light GBM model.

	Precision	Recall	F1-Score	Support
Stratiform	0.88	0.9	0.89	90
Transition	0.68	0.78	0.72	27
Convective	0.86	0.76	0.8	49

Table 7. Classification report for each cloud. Support is the number of test samples of each cloud type. All the values are averaged for each class by weighing with the number of samples.

Table 7 shows F1- scores of 0.89 and 0.8 for Stratiform and Convective respectively while showing 0.72 for mixed stratiform-convective clouds. The reason for the better prediction of stratiform

clouds correctly can be attributed to **the distinctive bright band** between 3-5 km which makes its pattern more unique. The **higher values of reflectivity between 1-2 km** and the absence of bright band between 3-5 km is what makes the prediction of convective clouds. But the higher values of reflectivity below melting layer is also shared by transition clouds due to embedded convection.

The results of transition clouds are lower than the other two. This is mainly due to two reasons:

1. The original number of samples is very low compared to the other two(transition had only 267 samples). It is increased using SMOTE but this method has it's own limitations since it is generating only synthetic samples.
2. The physical properties of transition clouds is more complex compared to the other two. It has characteristics similar to stratiform clouds as well as convective clouds. Transition clouds shows a bright band phenomena between 3-5 km but the decrease in reflectivity is less due to updrafts hence making it not prominent. Similarly due to embedded convection these clouds have higher reflectivity values between 1-2 km. This makes their structure more complex to distinguish.

To get the correctly classified and misclassified number of clouds by the model we analyse the confusion matrix shown below. The number of samples for stratiform, transition and convective are 90 , 27 and 49 respectively.

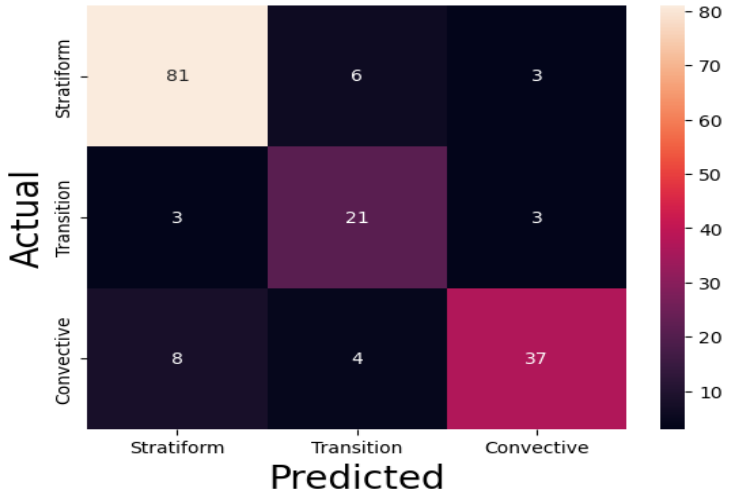


Figure 11. The Confusion Matrix for a Multiclass Classification model

The number of True positives(TP) for Stratiform, Transition and Convective are 81,21,37 respectively. So we can say that 27 clouds have been misclassified out of 166. There are 9 FN and 12

FP for Stratiform clouds. 6 FN and 10 FP for Transition clouds and 12 FN and 6 FP for Convective clouds.

The problem with this model is that **8 convective clouds have been wrongly predicted as stratiform clouds** and 3 stratiform clouds have been wrongly predicted as convective clouds. Even though we know that no model can be perfect, in our case misclassifying stratiform as convective and vice versa is the bigger mistake since both these clouds differ greatly in their precipitating intensities. Misclassifying convective or stratiform as transition is an affordable mistake compared to this one since transition clouds has properties lying between these two clouds.

So we have to improve on our model to minimize this problem as much as possible. For this, instead of doing one multiclass classification model we builds binary classification models for predicting stratiform and convective clouds separately. Since it is a binary classification, the complexity is reduced and one cloud can be predicted with more accuracy.

For building the first model which is designed for detecting Stratiform clouds, we label the stratiform clouds as 1 and every other clouds as 0. Similarly for the Convective cloud model, we label the convective clouds as 1 and everything else as 0. The Light GBM is used to construct both models. The results of the data run on both the models are shown below.

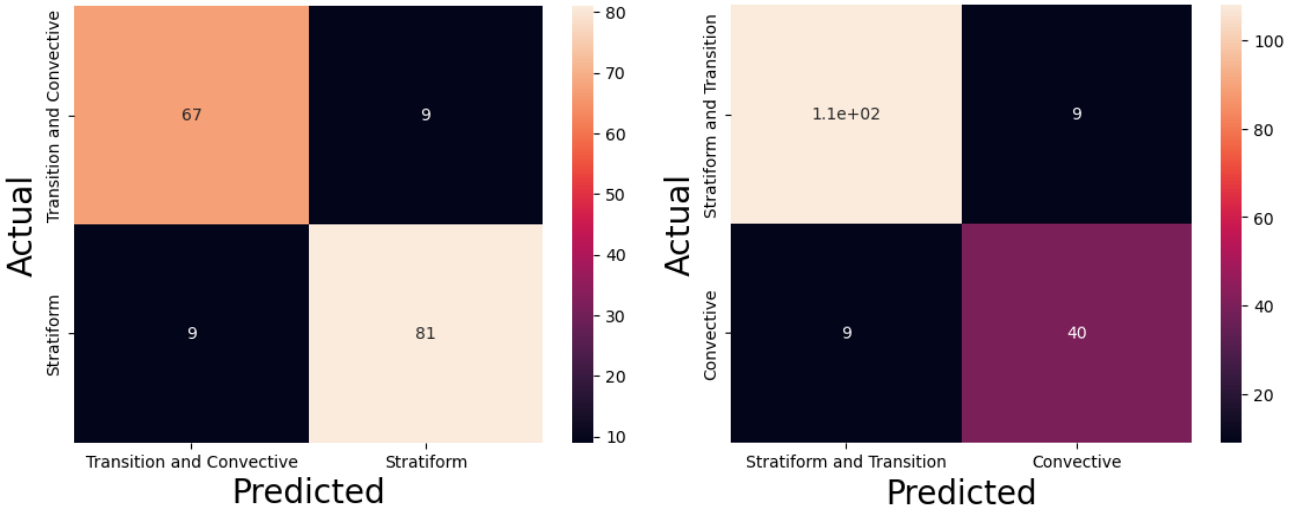


Figure 12.confusion matrix for the (a) Stratiform model (b) Convective model

Model (a) has a **0.9** f1-score for Stratiform clouds while Model(b) has a **0.82** for convective clouds. If the model (a) predicts 1 it is stratiform and 0 it is not stratiform(either transition or convective). Meanwhile if the model(b) predicts 1 it is convective and 0 it is not convective. **So if both the models predicts 0 we can assume that it should be a transition cloud.If both model predicts 1**

then it is a case where it could be either stratiform or convective.In order to tackle this we build another model which classifies into stratiform or convective entirely removing the transition clouds from the model. The flowchart of this process and the results obtained are shown below.

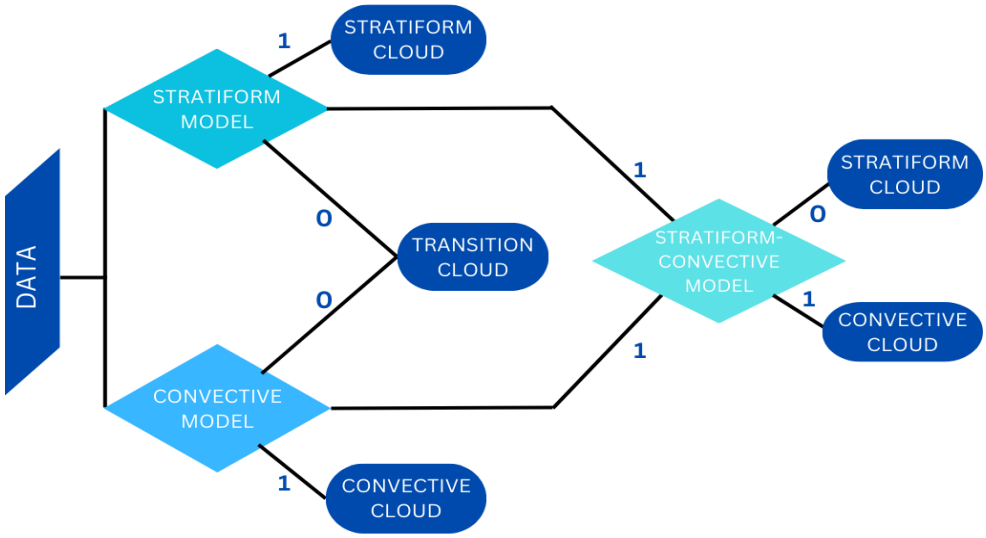


Figure 13. Flowchart for the ensemble binary classification models

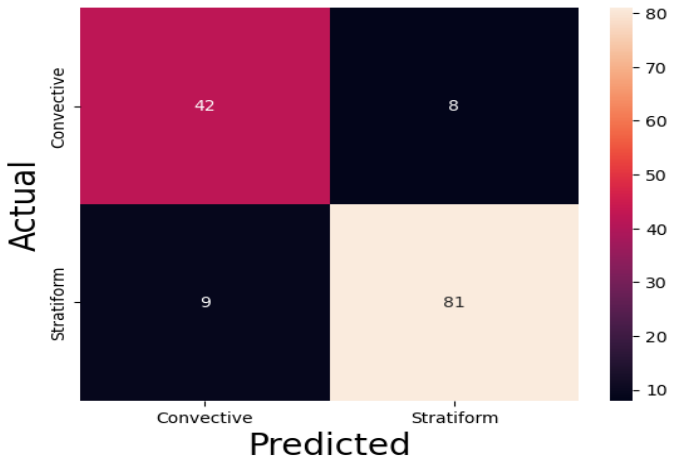


Figure 14. Stratiform-Convective Model

	Precision	Recall	F1-Score	Support
Stratiform	0.99	0.96	0.97	90
Transition	0.69	0.89	0.77	27
Convective	0.89	0.80	0.84	49

Table 8. Classification report from the results of the binary models.

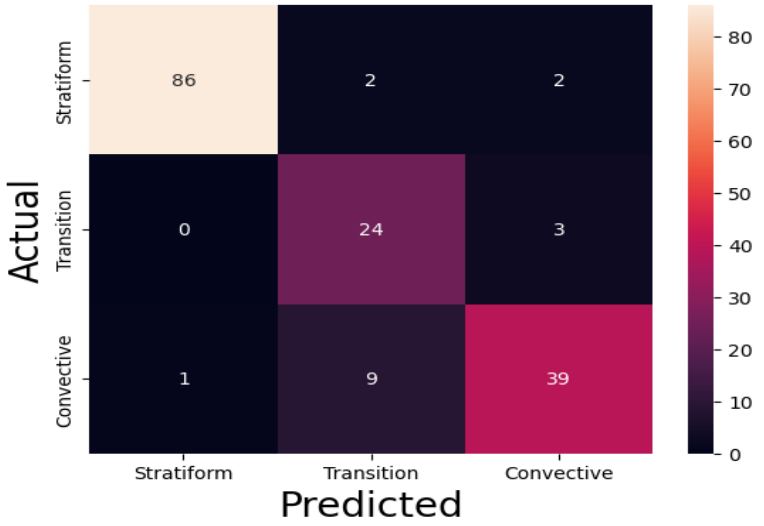


Figure 15. Final output from the 3 combined ensemble binary model

Figure 14 is a much improved result compared to Figure 11.This model has **0.88 BAC,0.91 precision,0.9 recall and 0.9 f1-score**. We can see that convective misclassified as stratiform and vice-versa has now reduced to 1 and 2 respectively. The number of TP for each cloud type has also increased and total number of FP and FN has been decreased.

We finally run this model on a data taken over the whole radar range not just the HACPL region. The clouds will be classified as Convective, Stratiform, Mixed Stratiform-Convective , Shallow clouds and Non precipitating clouds.

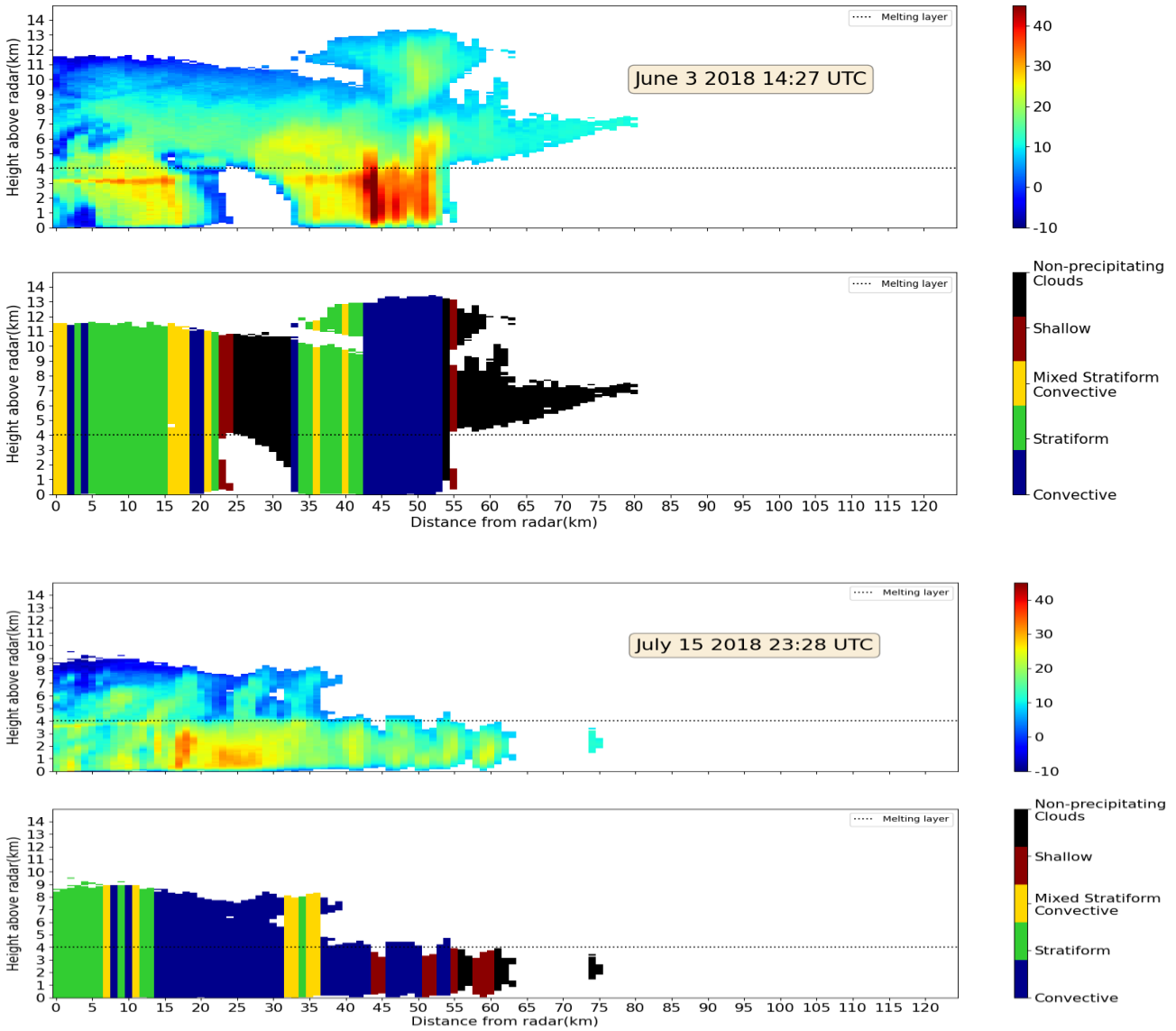


Figure 16. The model run on a data over the whole radar range taken on (a)June 3 2018 at 14:27 UTC. (b) July 15 2018 23:28 UTC.The VPR is taken averaging over 1 km in the horizontal

Analysing Figure 15(a), most of the convective clouds and stratiform clouds is predicted correctly by the model owing to its higher performance with both these clouds. Going deeper we can find that at 3 km from the radar a stratiform cloud is predicted by the model which when looking at the VPR can be seen that there is a clear bright band only for that cloud compared to its neighbours. The neighbouring clouds are predicted as convective because the bright band is not visible. Also at 36 km from radar there is a profile where there is a bright band along with a sign of convection below the melting layer. This has been correctly classified as transition cloud. Also clouds that does not show any reflectivity values up to 500m above the radar is considered as non precipitating clouds and clouds that does not reach the melting layer are shallow clouds. Figure 15 (b) also shows good results

with identifying more number of convective clouds at the same time not misclassifying stratfiorm as convective.

We can also see some misclassifications in both the figures. Clear bright band case at 16 km Figure(a) is classified as transition, but the presence of a high reflectivity just below the bright band makes this decision arguable. Also at 33 km a feeble bright band which should be classified as transition has been classified as convective.

5. SUMMARY AND CONCLUSIONS

This study is motivated due to a lack of proper classification of the precipitating clouds that will result in better rainfall estimation over the WGs. The study focuses on clouds taken over the HACPL Mahabaleshwar(17.92°N, 73.6°E, 1.4 km above MSL). The types of precipitating clouds that exists over this region are Shallow, Stratiform, Mixed Stratiform-Convective(Transition) and Convective clouds in which shallow clouds are found to occur almost 84 % of the time followed by stratiform, convective and transition clouds. The shallow clouds are classified based on the criteria that it does not exceed above the melting layer which is taken to be 5.3km(±0.19km) over the WGs(Kalapureddy et al 2022).To understand and determine the criteria for classification for the remaining clouds, two case studies have been conducted on studying the evolution of a mesoscale convective system since older convective cells causes trailing stratiform precipitation(Houze 1997). Using reflectivity and doppler velocity from the X-band radar and rain intensity and mass weighted diameter from Joss-Woldvogel Disdrometer(JWD) characteristics of convective, stratiform and the intermediary stage between this transition which is a mixed stratiform-convective cloud has been discussed using the different microphysical processes that includes droplet growth by deposition, aggregation, riming, melting of ice/snow to water, breakup of water droplets, growth by collection and evaporation occurring on several height levels based on the type of cloud. Based on this study the criteria for classifying all the types of clouds and reference structures of vertical reflectivity profiles is finalised. CFADs has been constructed for each cloud type to understand the height dependency of reflectivity. Based on the CFADs we select reflectivity of certain height levels that are the most important in separating these cloud types(feature selection). The clouds are then labelled as stratiform, transition and convective based on their distinctive characteristics. 1658 samples consisting of 898 stratiform, 493 convective and 267 transition clouds is prepared for training and testing the machine learning models. 90 % of data is used for training the model while the remaining 10 % is used for testing.20% of data from the training dataset is used for validating the model during training. 7 machine learning models that includes 3 baseline and 4 complex models have been trained for multiclass classification

using the data set and the test results shows that tuned models performs better in which Light GBM gives the best results with a **0.84 averaged F1-score and 0.81 accuracy**. Analysing the classification report of this multiclass classification model it is found that stratiform and convective clouds performs better with **0.89 and 0.8 F1-score** respectively compared to transition clouds which shows **0.72 F1-score**. Due to this result we did another method by building 3 binary classification models for stratiform and convective clouds and predicting whatever that does not fall under these two as transition cloud. This approach gave a better result by increasing the **F-1 score of stratiform , transition and convective to 0.97, 0.77 and 0.84 respectively with an accuracy of 0.88 and averaged F1-score of 0.9 .** Finally the model is run on two days of data on June 3 and July 15 and we get a good classification with few misclassifications too. Future work will be focused on comparing the model results with established radar algorithms and classifying the clouds over the entire radar domain and several other radar domains throughout the WGs. The model can be improved by collecting more number of training samples from different years of data.

REFERENCES

Anagnostou, E. N. (2004). A convective/stratiform precipitation classification algorithm for volume scanning weather radar observations. *Meteorological Applications*, *11*(4), 291–300. https://doi.org/10.1017/S1350482704001409

Arnaud, P., Bouvier, C., Cisneros, L., & Domínguez, R. (2002). Influence of rainfall spatial variability on flood prediction. *Journal of Hydrology*, *260*, 216–230.

Biggerstaff, M. I., & Listemaa, S. A. (2000). An Improved Scheme for Convective/Stratiform Echo Classification Using Radar Reflectivity. *Journal of Applied Meteorology*, *39*(12), 2129–2150. https://doi.org/https://doi.org/10.1175/1520-0450(2001)040<2129:AISFCS>2.0.CO;2

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Churchill, D. D., & Houze, R. A. (1984). Development and Structure of Winter Monsoon Cloud Clusters On 10 December 1978. *Journal of Atmospheric Sciences*, *41*(6), 933–960. https://doi.org/https://doi.org/10.1175/1520-0469(1984)041<0933:DASOWM>2.0.CO;2

Das, S. K., Konwar, M., Chakravarty, K., & Deshpande, S. M. (2017). Raindrop size distribution of different cloud types over the Western Ghats using simultaneous measurements from Micro-Rain Radar and disdrometer. *Atmospheric Research*, *186*, 72–82. https://doi.org/https://doi.org/10.1016/j.atmosres.2016.11.003

Das, S., Maitra, A., & Shukla, A. K. (2011). Melting layer characteristics at different climatic conditions in the Indian region: Ground based measurements and satellite observations. *Atmospheric Research*, *101*(1), 78–83. https://doi.org/https://doi.org/10.1016/j.atmosres.2011.01.013

Fabry and Zawadski. (n.d.).

Fujiwara, M. (1965). Raindrop-size Distribution from Individual Storms. *Journal of Atmospheric Sciences*, *22*(5), 585–591. https://doi.org/https://doi.org/10.1175/1520-0469(1965)022<0585:RSDFIS>2.0.CO;2

Ghada, W., Casellas, E., Herbinger, J., Garcia-Benadí, A., Bothmann, L., Estrella, N., Bech, J., & Menzel, A. (2022). Stratiform and Convective Rain Classification Using Machine Learning Models and Micro Rain Radar. *Remote Sensing*, *14*(18). https://doi.org/10.3390/rs14184563

Gunn, R., & Kinzer, G. D. (1949). The Terminal Velocity of Fall for Water Droplets in Stagnant Air. *Journal of the Atmospheric Sciences*, *6*(4), 243–248. https://doi.org/10.1175/1520-0469(1949)006<0243:TTVOFF>2.0.CO;2

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st ed.). Wiley-IEEE Press.

Houze, R. A. (1973). A Climatological Study of Vertical Transports by Cumulus-Scale Convection. *Journal of Atmospheric Sciences*, *30*(6), 1112–1123. https://doi.org/https://doi.org/10.1175/1520-0469(1973)030<1112:ACSOVT>2.0.CO;2

Houze, R. A. (1997). Stratiform Precipitation in Regions of Convection: A Meteorological Paradox? *Bulletin of the American Meteorological Society*, *78*(10), 2179–2196. https://doi.org/https://doi.org/10.1175/1520-0477(1997)078<2179:SPIROC>2.0.CO;2

Kalapureddy, M. C. R., Patra, S., Dhavale, V., & Nair, M. R. (2023). CloudSat inferred contrasting monsoon intra-seasonal variation in the cloud vertical structure over Indian regions. *Climate Dynamics*. https://doi.org/10.1007/s00382-022-06643-0

Konwar, M., Das, S. K., Deshpande, S. M., Chakravarty, K., & Goswami, B. N. (2014). Microphysics of clouds and rain over the Western Ghat. *Journal of Geophysical Research: Atmospheres*, *119*(10), 6140–6159. https://doi.org/https://doi.org/10.1002/2014JD021606

Kumar, S., Hazra, A., & Goswami, B. N. (2014). Role of interaction between dynamics, thermodynamics and cloud microphysics on summer monsoon precipitating clouds over the Myanmar Coast and the Western Ghats. *Climate Dynamics*, *43*(3), 911–924. https://doi.org/10.1007/s00382-013-1909-3

Lavanya, S., & Kirankumar, N. V. P. (2021). Classification of tropical coastal precipitating cloud systems using disdrometer observations over Thumba, India. *Atmospheric Research*, *253*, 105477. https://doi.org/https://doi.org/10.1016/j.atmosres.2021.105477

Nandargi, S., & Mulye, S. S. (2012). Relationships between Rainy Days, Mean Daily Intensity, and Seasonal Rainfall over the Koyna Catchment during 1961–2005. *The Scientific World Journal*, *2012*, 894313. https://doi.org/10.1100/2012/894313

Ran, Y., Wang, H., Tian, L., Wu, J., & Li, X. (2021). Precipitation cloud identification based on faster-RCNN for Doppler weather radar. *Eurasip Journal on Wireless Communications and Networking*, *2021*(1). https://doi.org/10.1186/s13638-021-01896-5

Schuur, T. J., Ryzhkov, A. V, Zrnić, D. S., & Schönhuber, M. (2001). Drop Size Distributions Measured by a 2D Video Disdrometer: Comparison with Dual-Polarization Radar Data. *Journal of Applied Meteorology*, *40*(6), 1019–1034. https://doi.org/https://doi.org/10.1175/1520-0450(2001)040<1019:DSDMBA>2.0.CO;2

Smith, J. A., & Krajewski, W. F. (1993). A modeling study of rainfall rate-reflectivity relationships. *Water Resources Research*, *29*(8), 2505–2514. https://doi.org/https://doi.org/10.1029/93WR00962

Steiner, M., Houze, R. A., & Yuter, S. E. (1995). Climatological Characterization of Three-Dimensional Storm Structure from Operational Radar and Rain Gauge Data. *Journal of Applied Meteorology and Climatology*, *34*(9), 1978–2007. https://doi.org/https://doi.org/10.1175/1520-0450(1995)034<1978:CCOTDS>2.0.CO;2

Tawde, S. A., & Singh, C. (2015). Investigation of orographic features influencing spatial distribution of rainfall over the Western Ghats of India using satellite data. *International Journal of Climatology*, *35*(9), 2280–2293. https://doi.org/https://doi.org/10.1002/joc.4146

Tokay, A., & Short, D. A. (1996). Evidence from Tropical Raindrop Spectra of the Origin of Rain from Stratiform versus Convective Clouds. *Journal of Applied Meteorology and Climatology*, *35*(3), 355–371. https://doi.org/https://doi.org/10.1175/1520-0450(1996)035<0355:EFTRSO>2.0.CO;2

Utsav, B., Deshpande, S. M., Das, S. K., & Pandithurai, G. (2017). Statistical Characteristics of Convective Clouds over the Western Ghats Derived from Weather Radar Observations. *Journal of Geophysical Research: Atmospheres*, *122*(18), 10, 10–50, 76. https://doi.org/https://doi.org/10.1002/2016JD026183

Wang, H., Shao, N., & Ran, Y. (2018). Identification of Precipitation-Clouds Based on the Dual-Polarization Doppler Weather Radar Echoes Using Deep-Learning Method. *IEEE Access*, *7*, 12822–12831. https://doi.org/10.1109/ACCESS.2018.2867546

Wang, Y., Tang, L., Chang, P.-L., & Tang, Y.-S. (2021). Separation of convective and stratiform precipitation using polarimetric radar data with a support vector machine method. *Atmospheric Measurement Techniques*, *14*(1), 185–197. https://doi.org/10.5194/amt-14-185-2021

Williams, C. R., Ecklund, W. L., & Gage, K. S. (1995). Classification of Precipitating Clouds in the Tropics Using 915-MHz Wind Profilers. *Journal of Atmospheric and Oceanic Technology*, *12*(5), 996–1012. https://doi.org/https://doi.org/10.1175/1520-0426(1995)012<0996:COPCIT>2.0.CO;2

Yuter, S. E., & Houze, R. A. (1995). Three-Dimensional Kinematic and Microphysical Evolution of Florida Cumulonimbus. Part II: Frequency Distributions of Vertical Velocity, Reflectivity, and Differential Reflectivity. *Monthly Weather Review*, *123*(7), 1941–1963. https://doi.org/https://doi.org/10.1175/1520-0493(1995)123<1941:TDKAME>2.0.CO;2