Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Master of Science

TITLE Study of Cloud Microphysics using Data Aggregation & Prediction of Liquid Water Content using Machine Learning

PRESENTED BY Yash Shailendra Gokhale

ACCEPTED BY THE DEPARTMENT OF

Chemical Engineering

HAMISH GORDON, PROFESSOR AND ADVISOR

12/6/2020
DATE

ANNE S. ROBINSON, PROFESSOR AND DEPARTMENT HEAD

12/7/2020
DATE

Study of Cloud Microphysics using Data Aggregation & Prediction of Liquid Water Content using Machine Learning

MASTER OF SCIENCE PROJECT REPORT

2020

Yash Shailendra Gokhale

Department of Chemical Engineering

Advisor: Dr. Hamish Gordon

Table of Contents

Table of Figures

Table of Tables

Glossary

Acronym	Word
CCN	Cloud Condensation Nuclei
CDNC	Cloud Droplet Number Concentration
GCM	Global Climate Model
KAZR	Ka-Band ARM Zenith Radar
LWC	Liquid Water Content
LWP	Liquid Water Path
MC3E	Mid-latitude continental convective clouds
MPL	Micropulse Lidar
MWR	Microwave Radiometer
PF	Pre-factor
RL	Raman Lidar
SGP	Southern Great Plains

Background & Motivation

Clouds serve as a key regulator of the earth's average temperature. Minute changes in the extent or location of clouds leads to significant changes in the atmosphere, coupled with changes caused by external factors such as greenhouse gases and human-induced aerosols (National Aeronautics and Space Administration, 2005). It is the need of the hour that realistic and fitting simulations of clouds are possible, which is imperative for accurate representation of the earth's climate and the effect of clouds on the atmosphere. The challenges in modeling cloud systems arise from the dependence of the processes on both macrophysical (particle height, cloud fraction and thickness) and microphysical (water phase, droplet size and cloud droplet density) properties. Thus, vertical profiles of liquid and ice water contents, along with the number and size distribution of the cloud particles are required to quantify the role of clouds in climate change and atmospheric processes (Dunn, Johnson, & Jensen, 2011).

In addition to the complications arising due to the complex processes involved in clouds, uncertainties in the numerical weather models are further needed to be incorporated into the climate model for accurate estimations. The main reasons for uncertainties are put forth by Seifert (Seifert, 2011), as follows:

1. Potential gaps in the empirical and theoretical description of cloud processes (nucleation, aggregation, splintering of particles, etc.) exist
2. Inherent variability can be found in clouds, cloud and aerosol droplets such as spatial-temporal structures in clouds, particle size distributions, etc.
3. Computational models are unable to model the processes in complete due to strong non-linearity and high complexity of cloud processes

In order to simulate the effects of direct and indirect effects of aerosols on radiation balance, it is important to represent the number and mass of cloud particles and aerosol particles independently (Gordon, et al., 2020). Cloud droplet concentration is mainly determined by the value of maximum saturation and CCN (Cloud Condensation Nuclei) activity spectrum. Droplet concentration is one of the most important parameters in climate models, as they are instrumental in determining the major microphysical cloud properties (height of precipitation onset, type of precipitation and radiative cloud properties) and the effect of aerosols on clouds is translated through the cloud droplet number concentration (Pinsky, Khain, Mazin, & Korolev, 2012).

Clouds are formed by processes which are strongly dependent on liquid water content and ice water content. Mixed clouds generally refer to clouds where liquid and ice particles coexist, whereas those with an absence of ice particles are termed as liquid clouds (Ntwali, Mugisha, Vuguziga, & Kakpa, 2016). For this project, clouds are assumed to behave as liquid clouds and subsequent climate models are developed exclusively through these liquid particles.

Cloud droplet number concentration (CDNC) is an essential parameter to investigate the influence of cloud dynamics on radiation and atmosphere. CDNC is dependent on size distribution and chemical composition of the droplet, cloud droplet velocity and on other dynamical effects related to the water transportation, growing aerosol particles and cloud droplets (Kivekas, et al., 2008). As put forth by Kivekas et al. (Kivekas, et al., 2008), using sophisticated parametrizations for CDNC estimations in large-scale models is dependent on the availability of exact information on aerosol number size distribution and its exact chemical composition. In most cases, such reliable information is absent; hence, a practical approach to solve this issue is to rely on parametrizations which only depend on bulk aerosol properties and other easily available parameters.

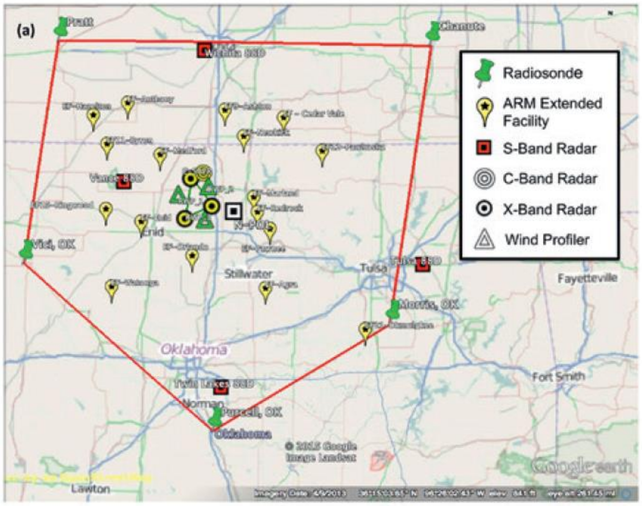


Figure 1: The sounding network that encompasses the central radar array (N-Pol, C-SAPR, triangular array of X-band radars in yellow, and 915-MHz profilers in green triangles) and SGP CF (Jensen, 2016)

The aim of this project is to test the efficacy of standard parametrizations put forth in literature, which are dependent, solely on bulk droplet properties and can be estimated purely through calculations. Data retrieved from the SGP site of the MC3 (Mid-latitude continental convective clouds) experiment. MC3E took place from 22nd April to 6th June 2011, mainly focused around the SGP (Southern Great Plains) site, with a combination of both airborne and ground-based instruments. The objectives of this campaign were motivated by the urge to understand the physical processes which drive the life-cycle of these convective clouds and the features of its precipitation (Jensen, 2016).

Parametrizations

Three standard parametrizations from literature were chosen to estimate the CDNC, the data of which was obtained from the SGP site and was further processed through a clustering methodology, explained in the later part of the report. Each of these parametrizations are dependent on atmospheric parameters at a particular time and altitude.

Yang et al. (Yang, et al., 2019) puts forth a new method to estimate in-cloud super-saturation fluctuations from ground-based remote sensing measurements. The analytical formula derived for the CDNC relies on the fundamental idea that the gradient of liquid water content (LWC) coupled with w can be used to estimate the mean super saturation between two layers of the cloud. The derived analytical parametrization is:

$$N_d = \frac{2e^{3\sigma_x^2}\rho^2}{9\pi}\frac{\sigma^3}{q^2}$$

Where N_d is the cloud droplet number concentration, $\sigma_x = \ln 1.4$ is the standard deviation obtained for a log-normal cloud droplet size distribution, ρ is the density, σ is the backscatter coefficient and q is the liquid water content.

The second parametrization used to estimate the CDNC is put forth by Pinksy, Khain, Mazin and Korolev (Pinsky, Khain, Mazin, & Korolev, 2012). The droplet concentration is mainly dependent on the temperature, cloud condensation nuclei concentration and particle updraft velocity.

Maximum saturation is estimated as: $S_{max} = C_3^{\frac{2}{2+k}} N_0^{\frac{-1}{2+k}} w^{\frac{3}{4+2k}}$

Based on this, the CDNC is estimated as: $N_d = N_o * S_{max}^k$, which simplifies to:

$$N_d = C_3^{\frac{2k}{2+k}} N_0^{\frac{2}{2+k}} w^{\frac{3k}{4+2k}}$$

k is a slope parameter which depends on geographical locations, meteorological conditions and time of the day, ranging from 0.3 to 1. N_o has typical values of 100 cm^{-3} for maritime clean conditions and ranges from 500-1000 cm^{-3}, depending upon the level of aerosol loading. w is the updraft speed of the particle and C_3 is a coefficient depending solely upon temperature.

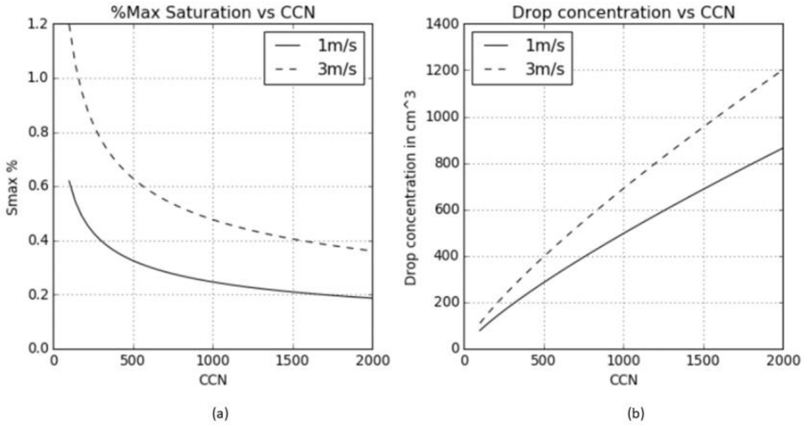


Figure 2: Recreated plots from Pinsky et al. (Pinsky, Khain, Mazin, & Korolev, 2012) to visualize the effect of CCN and constant updraft speed on the CDNC.

The third parametrization is obtained by combining the lidar equation in Frisch et al (Frisch, Fairall, & Snider, 1994) and radar reflectivity equation in Donovan et al (Donovan, et al., 2000). The parametrization, which is obtained through the combination of the two equations, is dependent on an assumed prior distribution of the cloud droplet radius. In this case, two standard distributions were considered, Gamma distribution and modified exponential distribution.

The distribution is denoted as: $< r^k >$

Based on the lidar equation, extinction coefficient, α can be modeled as:

$\alpha = 2\pi N_d < r^2 >$.

As per Donovan et al. (Donovan, et al., 2000) which follows a lognormal distribution, reflectivity, Z can be modeled as a function of the radius distribution as: $Z = 64N_d(\frac{K}{K_w})^2 < r^6 >$.

Combining the two equations: $N_d = \sqrt{\frac{8\alpha^3 <r^6>}{Z*\pi^3 <r^2>^3}(\frac{K}{K_w})^2}$

Irrespective of the assumed droplet radius distribution, numerator has a factor of r^6, whereas denominator has the factor of $(r^2)^3 = r^6$, and thus, the resulting droplet number concentration is independent of the droplet radius. The chosen droplet radius distribution influences the value of the pre-factor. The parametrization for the droplet concentration can be restated as:

$N_d = PF\sqrt{\frac{\alpha^3}{Z}}$, where PF is the constant term, derived from the assumed droplet radius distribution and setting $\frac{K}{K_w} = 1$.

Gamma function is defined as: $\Gamma(k) = (k+1)!$. The value of k was varied from 0 to 7 and the subsequent pre-factor was recorded. Similarly, for the exponential distribution, it is dependent on the standard deviation of the data and the power to which it is raised. The pre-factor for the modified exponential distribution is thus evaluated as: $e(i, s) = e^{\frac{i^2 s^2}{2}}$ where i, s are the power of the distribution and standard deviation of the data respectively.

The value of PF was thus, tabulated as follows:

Table 1: Value of the constant parametrization for assumed droplet radius distributions for parametrization 3

Type of distribution	Distribution parameter	Pre-factor
Gamma	0	4.8188
Gamma	1	2.4536
Gamma	2	0.8675
Gamma	3	0.2328
Gamma	4	0.0501
Gamma	5	0.009
Gamma	6	0.0014
Gamma	7	0.0002
Exponential	10^{-5}	0.5079
Exponential	0.001	0.508
Exponential	0.05	0.5156
Exponential	0.1	0.5394
Exponential	0.35	1.0593
Exponential	0.5	2.2765
Exponential	1	204.92

From Table 1, it can be observed that there is a strong dependence of the type of droplet radius distribution chosen on the CDNC. An assumed Gamma distribution (k=0) has a 10^4 higher pre-factor than the Gamma distribution (k=7). Thus, choosing the optimum distribution for the droplet radius is an important factor. Based on the fitting of experimental data collected for the SGP site, a Gamma distribution (k=2) was found to be the best. Alternatively, an exponential distribution with standard deviation, $\sigma = 0.35$ was found to have similar values as the Gamma distribution (k=2). The fitting methodology has been explained in the latter sections. Measuring the droplet radius directly is a difficult task and thus, in order to approximate the CDNC, parametrizations which rely on radar reflectivity are used directly, eliminating the use of droplet radius.

Instrumentation & Data Source

Based on the three parametrizations chosen, multiple atmospheric parameters such as updraft velocity, backscatter coefficient, temperature, etc. are required for estimation of the CDNC. For the scope of this project, data was retrieved from various instruments, recorded at the SGP site for the month of May, 2011. Certain days, where satisfactory liquid clouds were observed, were chosen to generate the basic data for estimations of the CDNC.

Recordings from 8 instruments/modes were combined together, through a spatial-clustering methodology established (explained in the later sections of the report) and corresponding analysis was conducted on the clustered data. The stored data was in the form of .cdf files, having a combination of both descriptive and numerical data. Table 2 provides a summary of the type of instruments which are being placed at the SGP site and the nomenclature which is followed for every date of recording and recording site. The variables which are required for the CDNC calculations are extracted from these instruments, the summary of which can also be seen in Table 2.

Table 2: Data files used in the project, i's nomenclature, type of instrument used and extracted variables

Data File Nomenclature (For standard ARM files at SGP)	Instrument/ Mode	Extracted Variables
sgparsclkazr1kolliasC1.[date, site].nc	KAZRARSCL	Updraft Velocity Reflectivity
sgpmicrobasekaplusC1.[date, site].nc	MICROBASE	Liquid Water Content, Ice Water Content
sgp10rlprofbe1newsC1.[date, site].cdf	Raman Lidar	Extinction coefficient
sgpaosccn1colC1.[date, site].nc	CCN Particle Counter Aboard Aircraft	CCN, Maximum Saturation

sgp1rlprofext1ferrC1.[date, site].cdf	Raman Lidar	Temperature
sgp30smplcmask1zwangC1.[date, site].nc	Micro pulse Lidar	Backscatter coefficient
sgpmwrret1liljclouC1.[date, site].nc	Microwave Radiometer	Liquid Water Path, Cloud Base
sgpceil10mC1.[date, site].nc	Ceilometer	Cloud base

KAZR-ARSCL VAP is a value added product that merges collected KA-band radar moments with cloud base and cloud mask observations from micropulse radar, cloud base from ceilometer and information from soundings, rain gauge and microwave radiometer. This dataset provides cloud boundaries and best-estimate radar moments, which can be used to understand cloud processes and lead to improvement in cloud and earth system models (ARM, 2020).

MICROBASE product is a combination from multiple instruments: 35-GHz millimeter wavelength cloud radar (MMCR), the ceilometer, the micropulse radar (MPL), the microwave radiometer (MWR) and baloon-borne radiosonde surroundings, which are located at fixed ARM sites, in close proximity to each other. The MICROBASE VAP is obtained after the data from these instruments is processed into usable, value-added products, making it a combination of Value Added Products (Dunn, Johnson, & Jensen, 2011).

Raman Lidar is an active, ground-based, laser remote sensing instrument that provides height and time resolved measurements of water-vapor mixing ratio, temperature, aerosol and cloud optical properties. Number of detection channels are tuned to measure the elastically backscattered light from atmospheric aerosol, yielding measurements of extinction and depolarization ratio amongst others (ARM, 2020).

The micropulse lidar is a ground-based, optical, remote-sensing system which is primarily used to determine the altitude of clouds and detection of atmospheric aerosols. Pulses of energy are transmitted into the atmosphere, the energy of which is scattered back to the transceiver as a time-resolved signal, thereby aiding the detection of clouds and aerosols in real time (ARM, 2020). Due to the time delay between outgoing pulse and backscattered signal, backscatter coefficient can be recorded, which can then be used to detect the cloud base.

Microwave radiometer (MWR) provides time-series measurements of a height-column integrated amounts of water vapor and liquid water, using which integrated water vapor and liquid water path is derived from a statistical retrieval algorithm that uses monthly derived and location-dependent coefficients (ARM Research Facility, 2020).

Ceilometer, a remote-sensing instrument measures cloud height, vertical visibility and potential backscatter signals by aerosols. The advantage of using measurements from ceilometers is its ability to detect three cloud layers simultaneously, roughly upto maximum vertical range of 7700 meters (ARM Research Facility, 2020).

Detection of Cloud-base for liquid clouds

Cloud base is the lowest altitude at which a visible portion of cloud can be observed. For liquid clouds, it can be defined as the lowest altitude in a height column at which visible portion is observed. Cloud base can be directly detected by certain instruments such as a microwave radiometer, ceilometer, etc. However, during the absence of such instruments, it is imperative to be able to approximate the cloud base, in order to proceed with further analysis of the recorded data, to thereby determine the cloud microstructure. In this section, a searching methodology to detect cloud base was put forth, using the backscatter coefficient as the deciding variable. Various alternate methods to arrive at a closer approximation to the cloud base were tested to arrive at the final searching method.

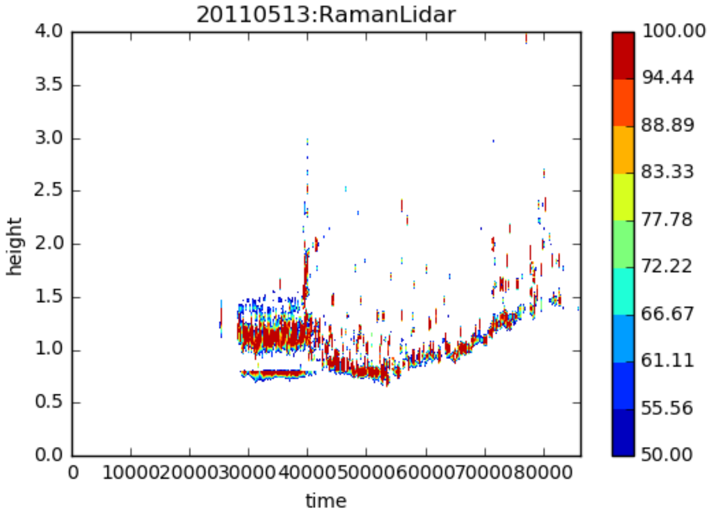
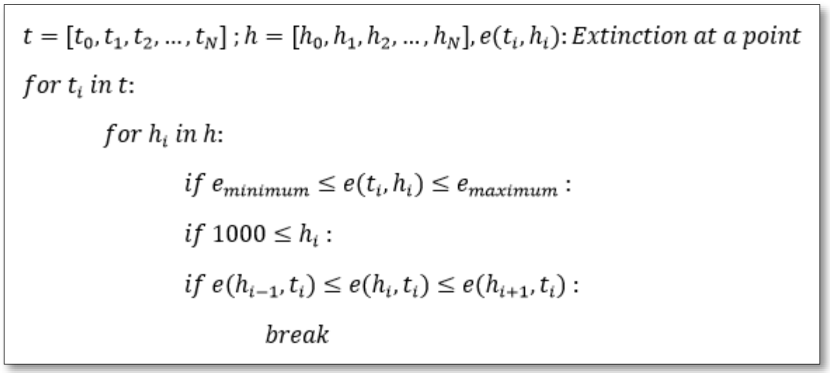


Figure 3: Height-time plot of extinction coefficient on '20110513' for the entire duration day. The cloud base for every height column was determined by applying the searching methodology to that column.

Pseudo-code for the searching algorithm was based on using a stopping criteria to detect the cloud base for a height column. The output of this searching algorithm would yield a time-series of the cloud base values.



```
$t = [t_0, t_1, t_2, \dots, t_N]$ ; $h = [h_0, h_1, h_2, \dots, h_N]$, $e(t_i, h_i)$: $Extinction\ at\ a\ point$
for $t_i$ in $t$:
    for $h_i$ in $h$:
        if $e_{minimum} \leq e(t_i, h_i) \leq e_{maximum}$ :
        if $1000 \leq h_i$ :
        if $e(h_{i-1}, t_i) \leq e(h_i, t_i) \leq e(h_{i+1}, t_i)$ :
            break
```

Figure 4: Pseudocode for the searching algorithm to detect cloud base using time and height resolved backscatter coefficient

The searching algorithm in Figure 4 takes into consideration three primary constraints to detect the first instance of the cloud base. The algorithm is run through every time instance and stopped when all the three conditions are fulfilled at the first instance. The first constraint limits the values of the extinction at a potential cloud base, in order to avoid any outliers caused by the error in recording of the data. Also, it is assumed that the cloud base should always be above 1 km, put forth by the second constraint. Lastly, in order to avoid faulty detection of cloud base, the cloud base has to have a higher value than a point below it, and a lower value of extinction above it. This constraint ensures that the presence of cloud above the cloud base is also taken into account and detection of any inherent fluctuations of the data is minimized.

The time series generated using this searching algorithm is a prerequisite to implement the spatial-temporal clustering methodology, in order to combine data from multiple instruments with different time and height resolution.

Sequential Clustering of Data files

In order to test the efficacy of the parametrizations for cloud droplet number concentration (CDNC), all the incoming atmospheric measurements should have the same resolution in time and altitude. However, as features are extracted from multiple radar files, each of the data files have a different time and altitude resolution. A sequential clustering method was developed to transform the data from an instrument, into the required time and height resolution.

Methodology

In order to superimpose values from one instrument onto another instrument having a different resolution, a three step clustering process has been implemented. The process of converting a parameter from one resolution to another resolution, two data files, from two corresponding instruments are used simultaneously.

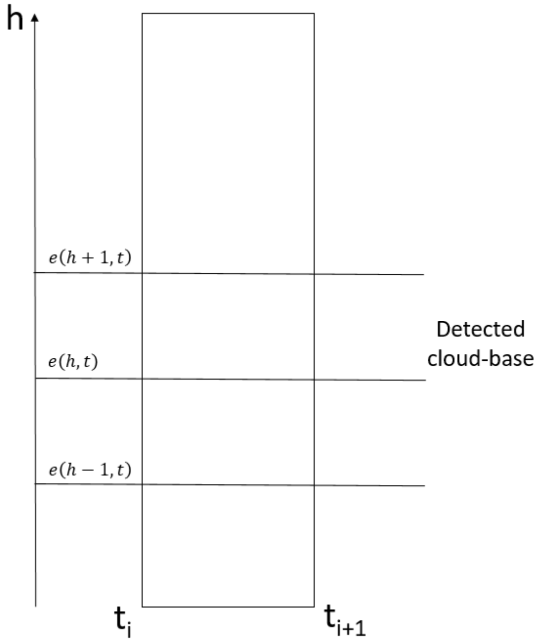


Figure 5: Graphic of the implementation of the searching algorithm for a single height column

Consider the case wherein two radars with time and height resolution as (t_1, h_1) and (t_2, h_2) are chosen for the clustering process. A parameter, 'Temperature', stored in the data file of Radar 1 with a shape of (t_1, h_1) has to be transformed to a shape of (t_2, h_2), so as to align all the variables in the same dimensions. As Radar 1 is to be transformed to Radar 2, we term (t_2, h_2) as the principal resolution, as it forms the basis of the clustering methodology. On the contrary, (t_1, h_1) is termed as the auxiliary resolution. Hence, mathematically:

$$t_2 = t_{pri}\ ;\ h_2 = h_{pri}\ ; t_1 = t_{aux}\ ; h_1 = h_{aux}$$

If $t_{aux} < t_{pri}$, the time resolution of the auxiliary radar has been increased; if $t_{aux} > t_{pri}$, the time resolution of the auxiliary radar has been decreased. Similarly, if $h_{aux} < h_{pri}$, the time resolution of the auxiliary radar has been increased; if $h_{aux} > h_{pri}$, the time resolution of the auxiliary radar has been decreased.

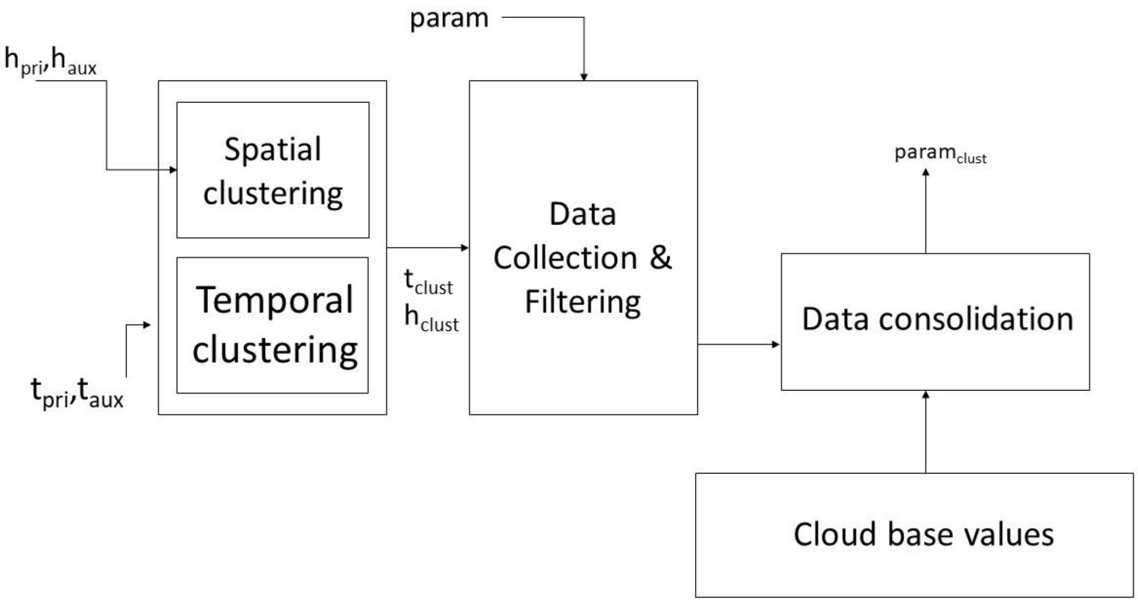


Figure 6: Process Diagram of the spatial-temporal clustering methodology for data superimposition of two instruments

A three step methodology, as showcased in Figure 6, is followed in a sequential manner to give out a dataset, $param_{clust}$, having the dimensions of (t_{pri}, h_{pri}), a change from the initial (t_{aux}, h_{aux}).

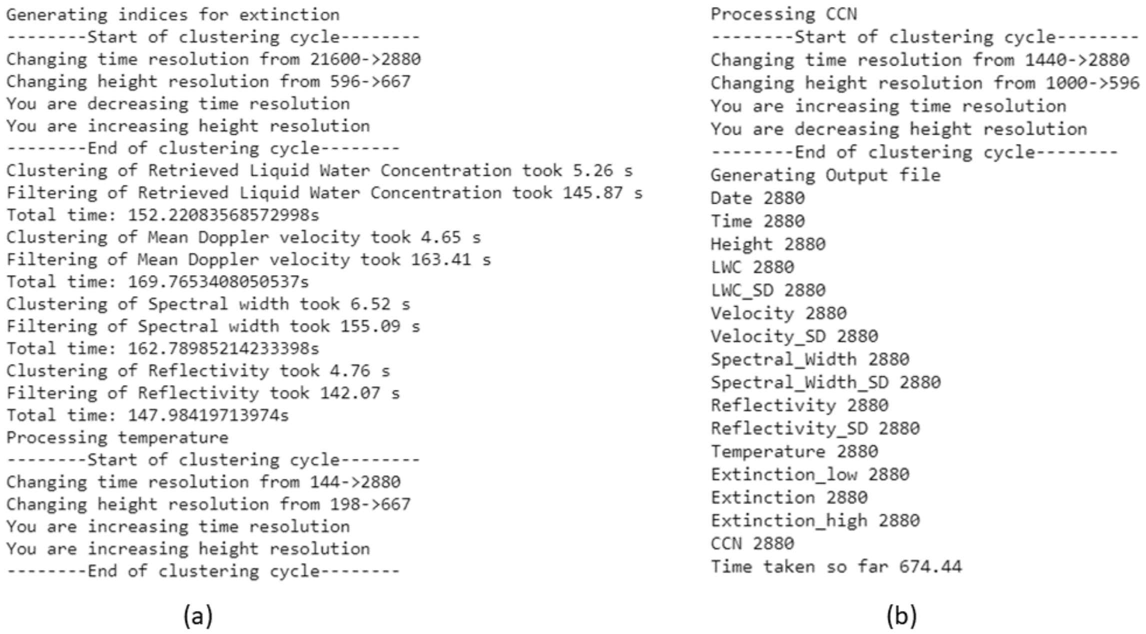
The first step involves generating two clustered lists of indices, generated simultaneously by combining the indices of t_{pri}, t_{aux} for time and h_{pri}, h_{aux} for height. All the indices of the t_{aux}/h_{aux} which lie between two adjacent indices in t_{pri}/h_{pri} are collected. Hence, the first clustering step yields two products, t_{clust} and h_{clust}, which is a collection of lists, which represent the clustered indices of the two input sets.

In the second step, t_{clust} and h_{clust} are run iteratively, to map every sub-list in t_{clust} to every sub-list in h_{clust} to yield an output which is of the same resolution as the principal resolution. In addition to collecting the data, the data is also filtered for unwanted values. For instance, if the parameter to be clustered is updraft velocity, all velocities which are lesser than zero are removed. Similarly, some values which are missing due to the inherent incapability of the instrument are also filtered.

In the final step, the filtered data is aligned with the time series of the cloud base, to ensure that only the values at the cloud base are rendered. The output $param_{clust}$ is the final output which is obtained through the three step clustering process.

Sequential Clustering using Python framework

An automated framework was developed using Python 3.6 to generate the clustered parameter, to be compliant with the resolution of the principal instrument. The framework outputs a single .csv file containing the required features at the cloud base. The only argument needed to be passed is the date on which the output file is required. Internally, the framework retrieves data from multiple files for multiple instruments, from the directed file path.



```
Generating indices for extinction
--------Start of clustering cycle--------
Changing time resolution from 21600->2880
Changing height resolution from 596->667
You are decreasing time resolution
You are increasing height resolution
--------End of clustering cycle--------
Clustering of Retrieved Liquid Water Concentration took 5.26 s
Filtering of Retrieved Liquid Water Concentration took 145.87 s
Total time: 152.22083568572998s
Clustering of Mean Doppler velocity took 4.65 s
Filtering of Mean Doppler velocity took 163.41 s
Total time: 169.7653408050537s
Clustering of Spectral width took 6.52 s
Filtering of Spectral width took 155.09 s
Total time: 162.78985214233398s
Clustering of Reflectivity took 4.76 s
Filtering of Reflectivity took 142.07 s
Total time: 147.98419713974s
Processing temperature
--------Start of clustering cycle--------
Changing time resolution from 144->2880
Changing height resolution from 198->667
You are increasing time resolution
You are increasing height resolution
--------End of clustering cycle--------
```

(a)

```
Processing CCN
--------Start of clustering cycle--------
Changing time resolution from 1440->2880
Changing height resolution from 1000->596
You are increasing time resolution
You are decreasing height resolution
--------End of clustering cycle--------
Generating Output file
Date 2880
Time 2880
Height 2880
LWC 2880
LWC_SD 2880
Velocity 2880
Velocity_SD 2880
Spectral_Width 2880
Spectral_Width_SD 2880
Reflectivity 2880
Reflectivity_SD 2880
Temperature 2880
Extinction_low 2880
Extinction 2880
Extinction_high 2880
CCN 2880
Time taken so far 674.44
```

(b)

Figure 7: Sample output for the Python framework, which follows a sequential order for the parameters. (a) is the first half of the output followed by (b).

The python based framework presented in Figure 7 is a convenient way of generating a .csv file, which is of the order of 1 Megabyte, from multiple .cdf/.nc files, each amounting to around 500 Megabyte of data. In addition to extracting information effectively from a large chunk of data, the Python framework also enables the user to keep track of the progress of the clustering algorithm. Depending upon the variance and complexity of the data, an average time of 10-15 minutes was observed for the days in consideration. The generated output file from this clustering serves as the starting point for further optimization and data analysis of the cloud systems.

Clustered Data Analysis

From the MC3E (Mid-latitude Continental Convective Clouds Experiment) which was conducted from 22nd April 2011 to 6th June 2011, six suitable dates were chosen, where satisfactory amount of liquid cloud was expected to be present. These six dates were:

1. 2011-05-05: Liquid cloud was observed visually for the latter half of the day
2. 2011-05-13: Liquid cloud was observed in patches throughout the day
3. 2011-05-14: Liquid cloud was observed visually from 03:00 hours to the end of the day
4. 2011-05-19: Liquid cloud was observed in patches throughout the day
5. 2011-05-27: Liquid cloud was observed visually for the latter part of the day
6. 2011-05-29: Liquid cloud was observed from 00:00 hours to 04:00 hours of the day

The clustering methodology was applied to each of these six days in question, and new data files of the format '.cdf' were generated, having information derived from multiple sources. Subsequent analysis was conducted either on the total data, or on individual dates.

Sensitivity Analysis of Variables

In order to assess the importance of variables and parameters in the parametrization put forth by Pinsky et al. (Pinsky, Khain, Mazin, & Korolev, 2012), effect of varying individual parameters was visualized for the entire dataset. As the parametrization depends on CCN, velocity and temperature, the effect of variance of velocity was analyzed, by comparing the droplet concentration based on the clustered values of CCN, temperature and a constant value of 1 m/s.

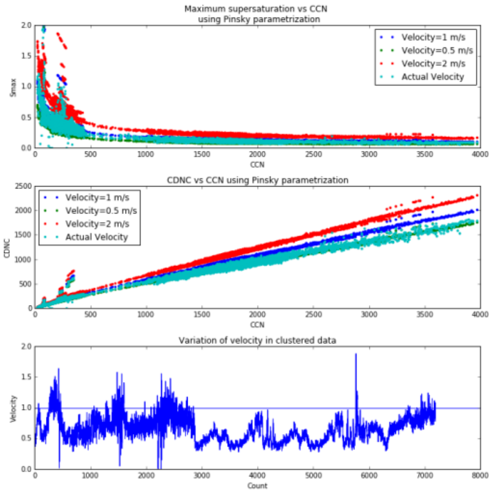


Figure 8: Effect of velocity on the maximum super-saturation and cloud droplet number concentration against the measured CCN (Cloud Condensation Nuclei). Effect of velocity is visualized, comparing the actual value of velocity against the plot with constant velocity.

From Figure 8, the variation of the clustered velocity can be observed. Although there is a slight variation in the clustered velocity, the effect of velocity can be observed in subplots 1, 2. Velocity does not play a major part in the maximum saturation and cloud droplet number concentration, which can be observed when the variables are calculated using a constant velocity, ranging from 0.5 to 2 m/s.

The CDNC put forth by Pinsky et al. (Pinsky, Khain, Mazin, & Korolev, 2012) depends on a parameter, ‘k’, which depends on the air quality and it varies from 0.3-1.

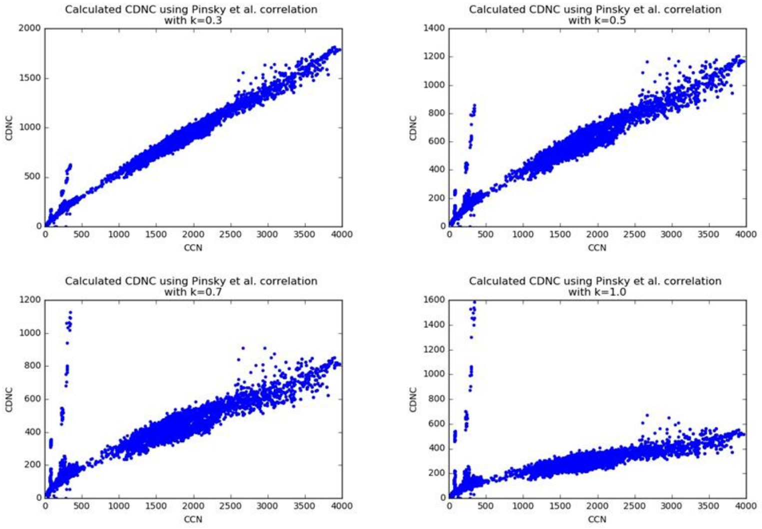


Figure 9: Calculated CDNC with multiple values of k, ranging from k=0.3 to k=1.

From Figure 9, it can be observed that, although the trend in the calculated CDNC is similar for all the values of k, the value of CDNC decreases with an increase in the value of k, as the general slope decreases with an increase in the value of k. Without taking into account the outliers, the value of CDNC increases with an increase in the value of CCN.

Parametrization Comparison

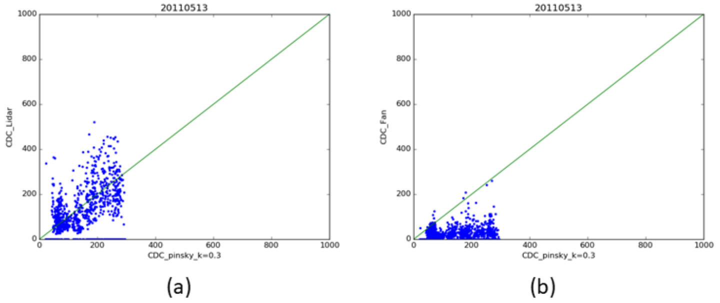


Figure 10: Correlation plots for CDNC derived from different parametrization for 20110513. (a): Lidar parametrization with gamma (2) distribution vs Pinsky parametrization ; (b) Yang parametrization with Pinsky parametrization

The correlation plots for multiple parametrizations in Figure 10 are indicative of the variance encountered in the recorded data from various instruments. For a sample day of 20110513, it can be observed that the Pinsky parametrization had a fair bit of accuracy with the Lidar parametrization. As the CDNC evaluated from Lidar parametrization is depedent on the basis of a fixed pre-determined distribution, a Gamma (2) distribution and a modified exponential (0.35) distribution was found to yield droplet concentrations within the specified range as the Pinsky parametrization. However, the Yang parametrization,as put forth by Yang et al. (Yang, et al., 2019) failed, as compared to the other two parametrizations. This can be attributed to the inaccuracies in measurement of the Liquid Water Content (LWC), as measured by the MICROBASE. As the clustered LWC at the cloud base has been determined through a searching algorithm implemented to find the cloud-base, further inaccuracies are inherited in the measurements of the extinction and liquid water content values at the cloud-base, which directly influence the calculation of the CDNC.

However, the Pinsky parametrization and Lidar parametrization do not always yield satisfactory results. As seen in Figure 11, the Lidar parametrization fails to match the CDNC values and yields a much lower value for the CDNC, as compared to the Pinsky parametrization. The possible reason for this deviation could be the dependence of the Lidar parametrization on the backscatter coefficient and droplet reflectivity. Inaccuracies in measurement of the radar reflectivity by the KAZR directly influence the CDNC calculation. On the other hand, the Pinsky parametrization is dependent on the updraft velocity, CCN and temperature, all three of which are comparatively easier to measure and would always yield a sensible value of the CDNC. Thus, even though the Pinsky parametrization has a high likelihood of providing sensible estimations for CDNC, they need not reflect the real values and could themselves be inaccurate and deviate from the actual measurement of the CDNC.

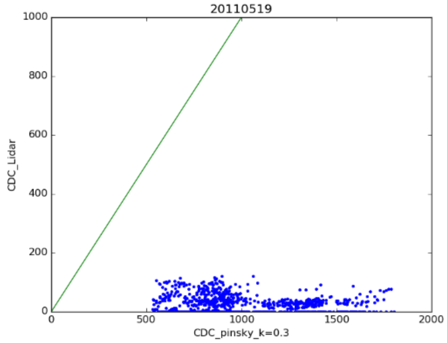


Figure 11: CDNC values calculated at the cloud base for 20110519, using the values for the parameters which are arrived through the clustering methodology established.

Predictive Models

Liquid Water Content (LWC) in liquid clouds influences the cloud dynamics and reflective properties of clouds (Gultepe & Isaac, 1996). Moreover, accurate estimations of liquid water content in cloud further helps in studying other parameters of GCMs (Global Climate Models) such as radiative absorption, optical thickness, etc. (Gultepe & Isaac, 1996) (Slingo & Schrecker, 1982). Several predictive models have been developed to establish relationships between liquid water content and other atmospheric parameters such as reflectivity, temperature and droplet velocity. In this section, effectiveness of several predictive models studied and analyzed to further improve the existing parametrizations of the atmospheric properties. For all the models mentioned in this section, the data points between the cloud base and cloud top were aggregated.

Linear Regression Models

Identifying simplistic trends, if any, need to be identified at the earliest instance before delving into developing complex and deeper prediction models. Linear regression models are tested to visualize any possible trend in the data. These regression models are fitted on individual dates, treated as a time-independent entity to assess whether linear models can be used to fit the complex data. The model follows the pattern:

$$LWC = a_0 + a_1 u + a_2 Z + a_3 CCN + a_4 N_d + a_5 T$$

where u: Updraft velocity; Z: Reflectivity; CCN: Cloud Condensation Nuclei; N_d: Droplet number concentration using Pinsky parametrization; T: Temperature; a_0: Bias

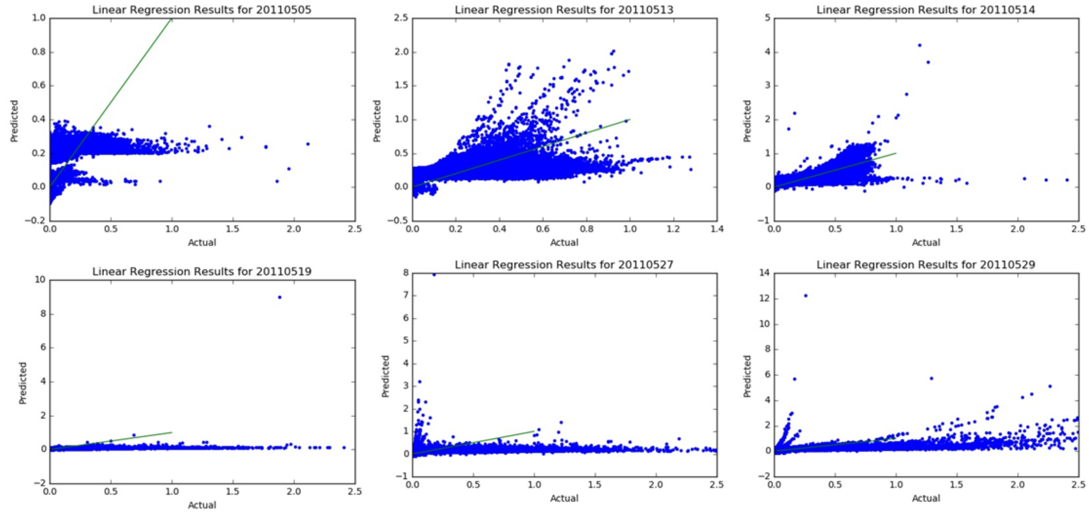


Figure 12: Simple linear regression models on a per day basis to visualize linear trend in the data

From Figure 12, it can be inferred that for the given dataset, no satisfactory linear trend can be observed. The linear regression model fails completely for the majority of the days, barring 2011-05-13. This implies a need for non-linear fitting models or higher order regression models to consider the inherent nonlinearities present in the data.

Decision Tree Models

Decision trees, which is a non-parametric supervised learning method, has been used for regression to predict the liquid water content. The benefit of using a decision tree model is its simplicity in comprehension and does not require any preprocessing of the data (scikitlearn, 2020). Although a decision tree might be slow to learn and unstable because of small variations in the data, hyper-parameter tuning is performed for the developed model in order to avoid overfitting and optimize the run-time.

Hyper-parameter tuning

In order to increase the accuracy of the Decision Tree algorithm developed and to avoid overfitting, optimal hyper-parameters such as depth of tree, splitting criterion, number of leafs per node, etc. were chosen through a grid search technique, implemented in Python using the GridSearchCV function of the sklearn library (scikitlearn, 2020). The exact implementation and detailed code of the grid search methodology can be referred to in the appendices.

The data under consideration was the combination of samples collected on each of the six individual dates, as listed in Clustered Data Analysis section. LWC was predicted using five dependent variables as: updraft velocity, droplet reflectivity, CCN, CDNC (calculated using Pinsky's parametrization) and temperature. Based on this input dataset, with an 80%-20% random train-test data split, the decision tree algorithm was used to predict LWC.

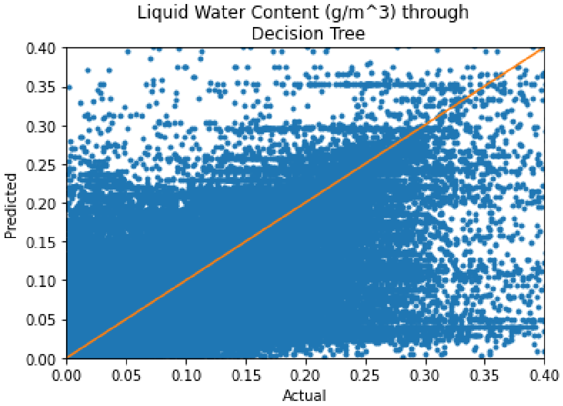


Figure 13: Grid search optimized decision tree regressor predictions on a randomly split dataset

Modeling on data sections

It can be observed from Figure 13, predictions on the entire dataset yields a scattered plot and no significant trend, either positive or negative can be observed from the fitted model. The possible source of inaccuracy could be arising from the variance due to multiple days being combined into a single dataset. The next step in improving the predictions is thus, to fit the decision tree on sections of data which would guide us to the possible sources of inaccuracies.

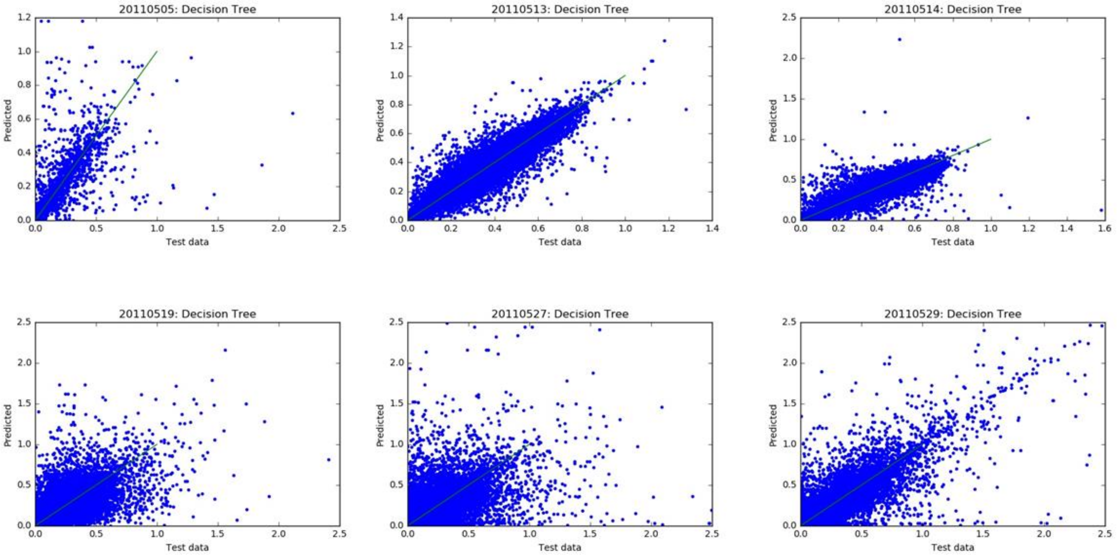


Figure 14: Decision Tree Regression Models with tuned hyper-parameters fitted on a per day basis

Developing fitting models on a per day basis leads to a better prediction of LWC, with most of the points lying on a close range to the ‘y=x’ line. From Figure 14, it can be observed that a better estimate on LWC can be obtained if the data is aggregated over a narrow time range and can then be treated as a grouped model to be developed (grouping on the basis of time but the model developed is time-independent).

Table 3: Number of samples in the dataset and accuracy of the decision tree fit on a per-day basis

Date	No. of samples	R^2
20110505	13185	0.473
20110513	278576	0.925
20110514	91436	0.847
20110519	570680	0.599
20110527	125601	0.314
20110529	88728	0.689

From Table 3, two features about the extracted dataset can be analyzed. The number of samples is indicative of the number of cloud particles above the cloud base can be observed. Satisfactory R^2 values, generally in the range of 0.6-0.9 are observed for most of the dates selected. However, the decision tree model fails on the total dataset, as seen in Figure 13. Moreover, as the decision tree model follows a closed form solution approach and is a non-parametric method, it is likely to fail on a larger dataset and could prove to be computationally expensive. This algorithm, in spite of performing better than the linear regression model, is not easily reproducible. Deeper neural neutral models would work better and faster, which is proposed in the latter sections.

Comparison of Measurement Techniques

Multiple instruments in use at a site can record the same atmospheric measurement. Liquid Water Path (LWP) is one such parameter which can be directly measured by an instrument or can be arrived at using standard equations. Microwave radiometer (MWR) can be used to approximately retrieve the liquid water path. For the MC3E, the microwave radiometer is able to directly retrieve the liquid water path, which is then stored in the corresponding data file. On the other hand, the value added product developed, MICROBASE is able to retrieve the liquid water content for the same site under scrutiny for MC3E. Based on the retrieval of the liquid water content, a time-series for LWP can be directly obtained as an integral of the liquid water content from the top to bottom of a height column (American Meteorological Society, 2012).

$$LWP_t = \int_{h_i=c_{base}}^{h_f=c_{top}} LWC_t(h) * dh$$

Thus, integrating LWC over a height column yields a time-series of the liquid water path for a fixed time period. In addition to comparing the LWP obtained through the microwave radiometer and MICROBASE, reflectivity recorded by the KAZR (Ka-Band ARM Zenith Radar) is used to estimate the LWC, which is then integrated for every height column to obtain the LWP estimates. The relationship between LWC and Radar Reflectivity (or reflectivity factor) is an empirical relation put forth by Liao and Sassen (Liao & Sassen, 1994), which assumes a cloud droplet number concentration of 100 cm^{-3}, as follows:

$$LWC = \left[\frac{N_d * Z_{liquid}}{3.6}\right]^{\frac{1}{1.8}}$$

KAZR is able to record the droplet reflectivity for both liquid and ice particles. However, as this study is only limited to liquid particles, an upper bound for the height column has been determined. For every time point in the KAZR recording for reflectivity, upper bound for liquid cloud has been defined as the last height index at which a liquid particle is observed (that is where LWC>0 and it is greater than the cloud base). For situations where the cloud base or cloud top could not be located, LWP was set to be zero.

The modified formula for the LWP calculated using KAZR is:

$$LWP_t = \int_{h_i=c_{base}}^{h_f=c_{top}} \left[\frac{N_d * Z_{liquid}(h)}{3.6}\right]^{\frac{1}{1.8}} * dh$$

For an ideal case, LWP retrieved from the microwave radiometer should coincide exactly with the calculated liquid water path from MICROBASE and KAZR. However, the large uncertainties present in the measurements, mainly due to the measurement techniques and the inherent variability found in clouds leads to a deviation in the values of the three LWPs, as seen in Figure 15 and Figure 16.

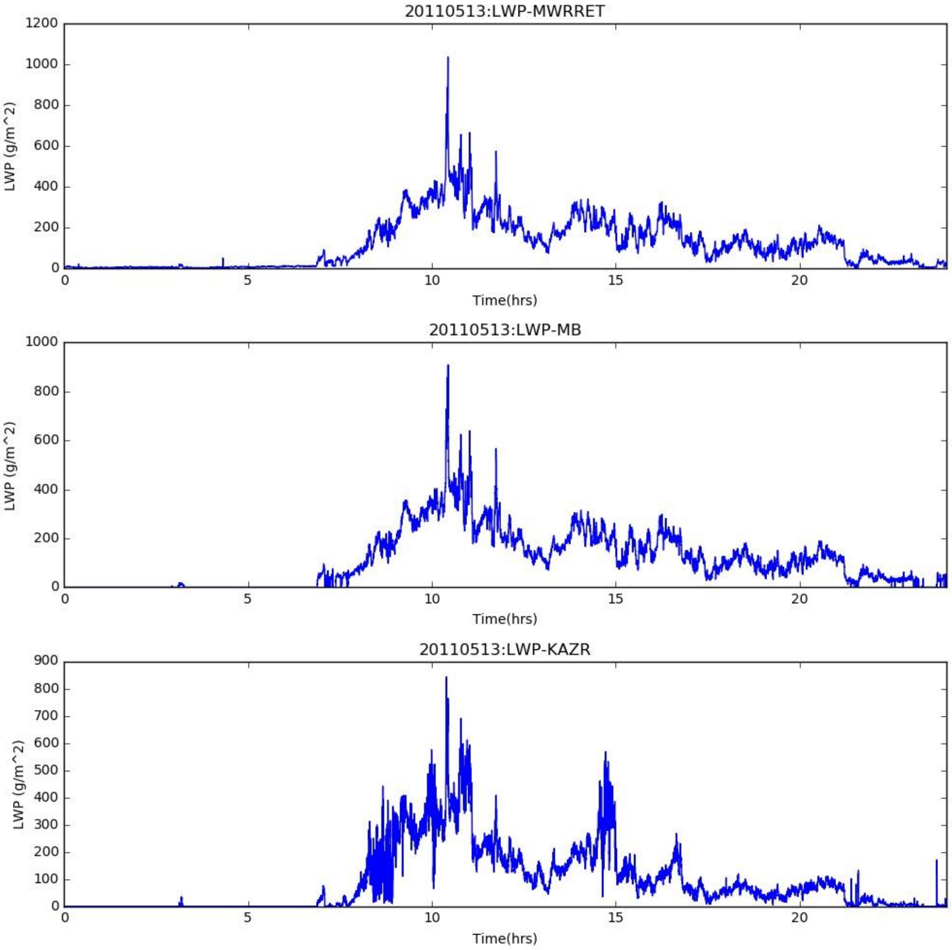


Figure 15: Time series of LWP for a particular day (20110513), which is indicative of the deviation in values of the LWP, retrieved directly or calculated from another retrieved variable.

For 20110513, the trend of the retrieved LWP seems to be in accordance with the LWP calculated based on the recorded LWC from MICROBASE and in sync with the LWP calculated in a two-step process from the KAZR reflectivity. As the calculation of LWP from KAZR is a two-step process, there are higher levels of uncertainty and approximations involved due to the detection of cloud-base and upper bound, which lead to greater error in the estimated LWP. The estimates seem to be in good accordance for 20110513, as seen in Figure 15. However, the estimates can go wrong by a larger factor, if the cloud-base and upper bound approximations are not accurate, which are themselves approximated from the backscatter coefficient using a searching algorithm, as seen in Figure 16.

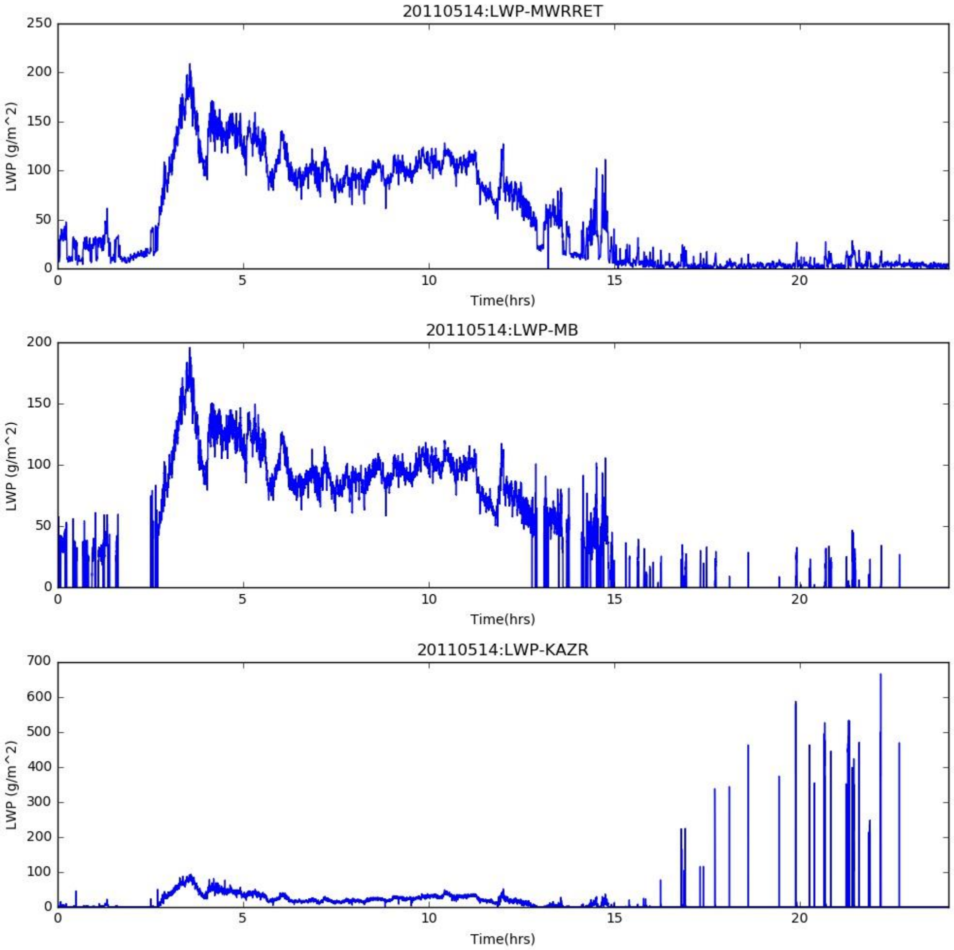


Figure 16: Time series plot for 20110514, wherein the estimations for LWP using KAZR reflectivity differ by a large factor due to error in detection of cloud-base and upper bound for a height column.

Addressing Uncertainties in Calculations

Arriving at the final estimate of the liquid water path from KAZR involves multiple steps, starting from clustering the data, followed by detection of cloud base and top, followed by correlating LWC and Z to finally integrating the LWC over a height column. Each of these step contribute to the uncertainty of the LWC estimated and thus leads to inaccuracies in the final calculations. Each of the uncertainties are addressed individually in this section of the report.

Spatial-temporal clustering of data

To arrive at the required time and height resolution for reflectivity, reflectivity derived from the KAZR data needs to undergo the clustering process, wherein the data is consolidated and filtered, to match the resolution of the principal instrument. During this process, there is a possibility that the inherent uncertainties in the measurement of the variable is amplified due to further transformations in the variable. This is the primary source of uncertainties in the data.

Detection of cloud-base and cloud-top

The presence of liquid cloud in a height column is detected using the searching algorithm, which estimates the cloud base using the extinction coefficient as its baseline parameter. As this searching algorithm is purely based on the extinction values and does not take into account, other parameters such as reflectivity, velocity and spectral width, which could influence the presence of liquid particles, the searching algorithm might yield inaccuracte results in the approximation of the cloud base. The cloud top is defined as the last point in a height column wherein the liquid water content has a positive value. On the basis of these two indices, the upper and lower portion of the liquid cloud is approximated and all the reflectivity values lying between these indices is chosen. From Figure 17, it can be seen that, for some parts of the day, cloud base and cloud top cannot be distinguished distinctly, and in such cases, cloud base and cloud top are assumed to have the same index. For cases whether the cloud top could not be detected by the searching algorithm, the cloud base and cloud top are set to zero. In both these situations, it ensures that the calculation of the liquid water path is unaffected, as the integration yields a zero value. Thus, this methodology of approximating the structure of the cloud contributes to large uncertainties in the reflectivity and ultimately to the values of LWP.

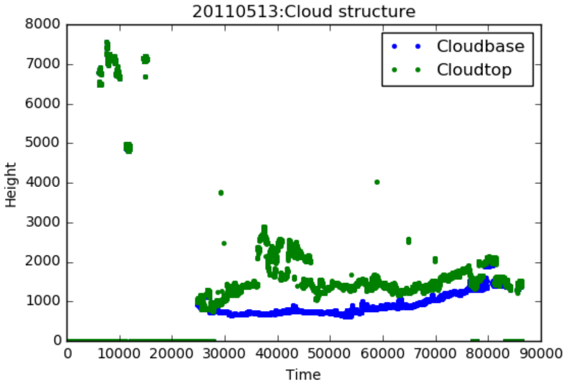


Figure 17: Liquid cloud microstructure determined using the searching algorithm

Empirical relation of reflectivity and LWC

The empirical relationship put forth by Liao and Sassen (Liao & Sassen, 1994) is used to estimate the LWC using reflectivity factor. This relationship assumes a constant cloud droplet concentration of 100 cm^{-3} to estimate the LWC. However, in reality, the CDNC is rarely constant, as seen in the calculations of CDNC using different parametrizations. Assuming a constant CDNC leads to inaccuracies in the calculated LWC, which is further carry-forwarded in the integration, to arrive at the LWP. Secondly, the power of $\left(\frac{1}{1.8}\right)$ is not fixed and cannot be generalized. Instead, the power is dependent on the type of conditions (continental or maritime) and the CCN concentration.

Integration over a height column

After arriving at the LWC values (either directly from MICROBASE or through the empirical relationship for reflectivity), these values are integrated over a height column. As the exact formula for LWC as a function of the column height is not known, numerical methods are to be used to arrive at the LWP value, for a liquid column. For this project, trapezoidal rule has been used to approximate the integration of LWC against a height differential.

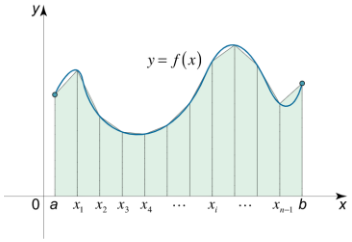


Figure 18: General schematic to implement trapezoidal rule for numerical integration (Math24, 2020)

Numerical integration using trapezoidal rule is arrived at, by the following method:

$$\int_a^b f(x)dx \sim \frac{\Delta x}{2}[f(x_o) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n)$$

Where $\Delta x = \frac{b-a}{n}$ and n is the number of steps

For this project, the approximation for the trapezoidal rule is done using the 'np.trapz' function in Python (NumPy, 2020). The numerical integration step, thus, further adds to the uncertainty in calculations.

Conclusion and Future Work

Representation of cloud microphysics is a key aspect of simulating clouds and understanding the atmospheric processes associated with the formation of clouds. In this project, major focus was directed towards developing a framework to combine data from multiple instruments and establish a methodology to correlate atmospheric parameters such as reflectivity, updraft velocity, temperature, LWC, LWP, etc. with each other. Moreover, analyzing the sources of uncertainty in these measurements and its effect on the predictions of variables was studied.

Through this project, a spatial-temporal clustering framework was ideated from scratch, which was followed by the development of an automated script, using Python 3.6, which was efficient in working with large files of atmospheric datasets and compressing into to a single compact file. Thus, this project was successful in aggregating data from multiple instruments and studying the cloud microstructure using the generated data file. Additionally, efficacy of three standard parametrizations was also studied through its utility for the MC3E. It was also inferred that the source of uncertainty in these parametrizations was arising from two primary reasons: inherent inaccuracies in the measurement of the data and formulation of the parametrization itself.

By comparing the efficacy of Figure 13 and Figure 14, it was concluded that using fitting algorithms such as regression models, decision tree models or higher order neural networks can be more effective if used on chunks of data, segregated based on time frame. Using the entire data for developing predictive algorithms can be computationally expensive without any success, since using a single modeling approach to capture such a wide range of uncertainties can lead to inefficiencies.

The inherent variance in atmospheric parameters poses multiple challenges to develop robust climate models. Moreover, it has been observed that a developed climate model for a region of interest is seldom effective and cannot be generalized to other regions. Location and its climatic conditions play a major role in defining the climate models. By applying the clustering methodology and the searching algorithm to detect the cloud base for MC3E, it can be inferred that even minute assumptions in the model could cause large uncertainties in the data, which ultimately lead to inaccuracies in predicting the atmospheric variables, critical to understanding the cloud microstructure. Establishing stronger assumptions in the clustering methodology and tuning the searching algorithm for cloud base of deep convective clouds is a possible future path of work, with an aim of improving the estimations and getting rid of certain inaccuracies in the methodology. It was observed that assumed cloud droplet radius distribution affects the estimations of CDNC to a large extent, since CDNC is dependent on reflectivity, which is inturn dependent on the sixth power of the droplet radius.

Modeling the behavior of LWC and LWP has been tested in this project, mainly using regression based models and non-parametric decision tree models. Although these models can be deemed effective in identify a trend between the dependent and independent variables, greater accuracy is essential in predicting the exact value of the variables such as LWC and LWP. A possible extension to this project is to implement higher order machine learning models such as a deep neural network model, as proposed in Figure 19.

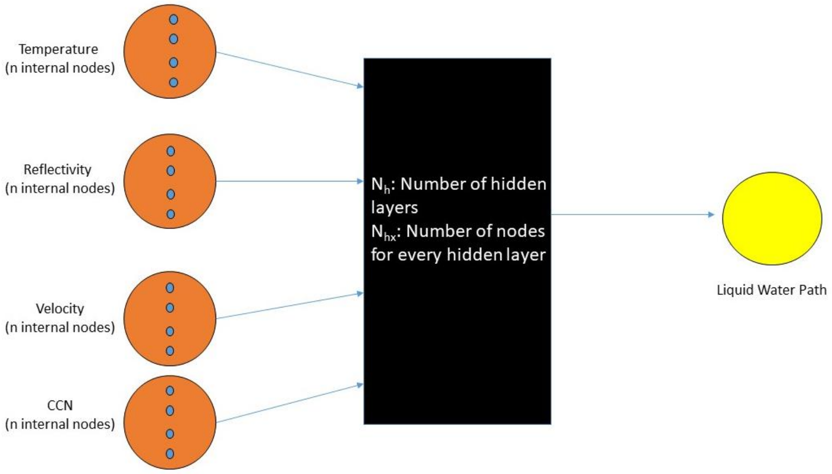


Figure 19: Neural network based prediction models to predict liquid water path in clouds

As the liquid water path is an integration of the LWC over a height column, using neural network based prediction models with a group of atmospheric variables (Temperature, Reflectivity, Velocity, CCN, etc.) at each height index as the independent variables and the liquid water path (measured by the microwave radiometer) as the dependent variable. The hyperparameters of this neural network model can be tuned by standard approaches.

The possible future work, thus, involves extending the clustering methodology and developing an automated framework to combine variables from multiple data sources, which would then aid in developing robust machine-learning based predictive models for critical variables such as LWP, LWC, etc.

Supplementary material

All the results were simulated using Jupyter Notebook software with the complete programming done using the Python language. Codes for the project are made available at:

Gokhale, Yash; Gordon, Hamish; Study of Cloud Microphysics using Data Aggregation & Prediction of Liquid Water Content using Machine Learning (2020), GitHub Repository, https://github.com/yashgokhale/CMU-MS-Research

Data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research Program.

Acknowledgements

This project report and the research conducted in developing these robust mathematical based climate model frameworks would not have been possible without the constant support of my guide, Dr. Hamish Gordon. I would also like to extend my sincere thanks to the Department of Chemical Engineering, Carnegie Mellon University for providing me with this golden opportunity to collaborate with experts in this field, which has been instrumental in aiding my growth as a researcher and a student of engineering.

References

American Meteorological Society. (2012, February 20). *liquid water path*. Retrieved from Glossary of Meteorology: https://glossary.ametsoc.org/wiki/Liquid_water_path

ARM. (2020). *KAZRARSCL*. Retrieved from ARM: https://www.arm.gov/capabilities/vaps/kazrarscl

ARM. (2020). *MPL*. Retrieved from ARM: https://www.arm.gov/capabilities/instruments/mpl

ARM. (2020). *RL*. Retrieved from ARM: https://www.arm.gov/capabilities/instruments/rl

ARM Research Facility. (2020). *CEIL*. Retrieved from ARM: https://www.arm.gov/capabilities/instruments/ceil

ARM Research Facility. (2020). *MWR*. Retrieved from ARM: https://www.arm.gov/capabilities/instruments/mwr

Donovan, D., Van Lammeren, A. C., Hogan, R., Russchenburg, H., Apituley, A., Francis, P., . . . Agnew, J. (2000). Combined Radar and Lidar Cloud Remote Sensing: Comparison with IR Radiometer and In-Situ Measurements. *Physics Chem Earth*.

Dunn, M., Johnson, K., & Jensen, M. (2011). *The Microbase Value-Added Product: A Baseline Retrieval of Cloud Microphysical Properties.* U.S. Department of Energy.

Frisch, A., Fairall, C., & Snider, J. (1994). Measurement of Stratus Cloud and Drizzle Parameters in ASTEX with Ka Band Doppler Radar and a Microwave Radiometer. *Journal of the Atmospheric Sciences*.

Gordon, H., Field, P., Abel, S., Barrett, P., Bower, K., Crawford, I., . . . Carlslaw, K. (2020). Improving aerosol activation in the double-moment Unified Model with CLARIFY measurements. *Atmospheric Chemistry and Physics*. doi:https://doi.org/10.5194/acp-2020-68

Gultepe, I., & Isaac, G. (1996). Liquid Water Content and Temperature Relationship from Aircraft Observations and Its Applicability to GCMs. *Journal of Climate*.

Jensen, M. P. (2016). The Midlatitude Continental Convective Clouds Experiment (MC3E). *Bull. Amer. Meteor. Soc., 97*, 1667–1686. doi:https://doi.org/10.1175/BAMS-D-14-00228.1.

Kivekas, N., Kerminen, V.-M., Anttila, T., Korhonen, H., Lihavainen, H., Komppula, M., & Kulmala, M. (2008). Parameterization of cloud droplet activation using a simplified treatment of the aerosol number size distribution. *Journal of Geophysical Research*. doi:10.1029/2007JD009485

Liao, L., & Sassen, K. (1994). Investigation of relationships between Ka-band radar reflectivity and ice and liquid water contents. *Atmospheric Research*, 231-248 .

Math24. (2020). *Trapezoidal Rule*. Retrieved from Calculus: https://www.math24.net/trapezoidal-rule/

National Aeronautics and Space Administration. (2005). *The Importance of Understanding Clouds.* NASA.

Ntwali, D., Mugisha, E., Vuguziga, F., & Kakpa, D. (2016). Liquid and ice water content in clouds and their variability with temperature in Africa based on ERA-Interim, JRA-55,MERRA and ISCCP. *Meteorology and Atmospheric Physics*. doi:10.1007/s00703-016-0447-z

NumPy. (2020, 06 29). *numpy.trapz*. Retrieved from NumPy: https://numpy.org/doc/stable/reference/generated/numpy.trapz.html

Pinsky, M., Khain, A., Mazin, I., & Korolev, A. (2012). Analytical estimation of droplet concentration at cloud base. *JOURNAL OF GEOPHYSICAL RESEARCH*. doi:10.1029/2012JD017753

scikitlearn. (2020, November 29). *scikitlearn*. Retrieved from Decision Trees: https://scikit-learn.org/stable/modules/tree.html

scikitlearn. (2020, November 29). *sklearn.model_selection.GridSearchCV*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Seifert, A. (2011). Uncertainty and complexity in cloud microphysics. *Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models.* Reading: ECMWF. Retrieved from https://www.ecmwf.int/node/12166

Slingo, A., & Schrecker, H. (1982). On the shortwave radiative properties of stratiform water clouds. *Quarterly Journal of the Royal Meteorological Society*.

Yang, F., Robert, M., Luke, E., Zhang, D., Kollias, P., & Vogelmann, A. (2019). A new approach to estimate supersaturation fluctuations in stratocumulus cloud using ground-based remote-sensing measurements. *Atmospheric Measurement Techniques*.