IDENTIFYING CLOUD FIELD REGIMES FROM WORLD WIDE MERGED CLOUD ANALYSIS VIA K-MEANS CLUSTERING

THESIS

Stewart G. Almeida, Captain, USAF

AFIT-ENP-MS-21-M-098

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.

AFIT-ENP-MS-21-M-098

IDENTIFYING CLOUD FIELD REGIMES FROM WORLD WIDE MERGED CLOUD ANALYSIS VIA K-MEANS CLUSTERING

THESIS

Presented to the Faculty

Department of Engineering Physics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Atmospheric Science

Stewart G. Almeida, BS

Captain, USAF

March 2021

AFIT-ENP-MS-21-M-098

IDENTIFYING CLOUD FIELD REGIMES FROM WORLD WIDE MERGED CLOUD ANALYSIS VIA K-MEANS CLUSTERING

Stewart G. Almeida, BS

Captain, USAF

Committee Membership:

Maj P. A. Saunders, PhD
Chair

Lt Col R. C. Tournay, PhD
Member

Maj K. E. Fitch, PhD
Member

AFIT-ENP-MS-21-M-098

Abstract

Joint histograms of cloud top height (CTH) and optical depth (OD) are created using the World-Wide Merged Cloud Analysis (WWMCA) dataset over a four year period (2014-2017) to identify average cloud field regimes and assess the application of utilizing the WWMCA dataset with the AFIT Sensor and Scene Emulation Tool (ASSET). Two selected regions encompassing the Florida peninsula and a portion of the Pacific Ocean off the west-central coast of South America are examined over the months of January and July. Cloud field regimes are identified by running generated hourly OD-CTH histograms through k-means clustering, with optimal cluster number (K) evaluation performed by calculating and comparing silhouette scores and heuristic elbow method results. Varying cluster groupings are plotted out to distinguish discrepancies between these multi-cluster analysis. Initial results indicate $K = 3$ as the optimal number of clusters to use, generating three major cloud field regimes unique to each region with high relative frequency of occurrences (RFO). Notable departures from silhouette score calculations to cluster evaluation call into question the validity of silhouette score usage to determine optimal K values, which is discussed alongside future improvements and applications of the cloud field regimes identified.

Acknowledgments

I would like to express the utmost appreciation to my research advisor, Maj Saunders, for his thought provoking insight and invaluable guidance, both of which allowed this project to develop and grow in an otherwise tumultuous year. I would also like to thank the ASSET team, whose input and on-going efforts are what fueled and brought this project into reality.

And to my family, whose love and support continue to restlessly drive me forward.

Stewart G. Almeida

Table of Contents

	Page
Abstract	iv
Acknowledgments	v
Table of Contents	vi
List of Figures	vii
I. Introduction	8
II. Methodology	14
ASSET Background	14
WWMCA Background	16
Selected Regions of Interest	18
Development of Joint Histograms	20
K-Means Clustering Process and Implementation	22
Standardization of Data for Clustering Process	25
Cluster Evaluation - Silhouette Score	25
Remapping of Cluster Grouping Identification	28
III. Analysis and Results	30
Chapter Overview	30
January - FL Region	30
January - SA Region	36
July - FL Region	43
July - SA Region	48
IV. Conclusions and Recommendations	54
Appendix - Acronym List	58
Bibliography	59

List of Figures

	Page
Figure 1. Geographic Regions of Interest	19
Figure 2. Mean OD-CTH Histogram over FL Region - January	31
Figure 3. Silhouette Scores over FL Region - January	32
Figure 4. 6 Cluster Analysis for FL Region - January	33
Figure 5. 3 Cluster Analysis for FL Region - January	36
Figure 6. Mean OD-CTH Histogram over SA Region - January	37
Figure 7. Silhouette Scores over SA Region - January	39
Figure 8. 4 Cluster Analysis for SA Region - January	40
Figure 9. 3 Cluster Analysis for SA Region - January	42
Figure 10. Mean OD-CTH Histogram over FL Region - July	43
Figure 11. Silhouette Scores over FL Region - July	44
Figure 12. 4 Cluster Analysis for FL Region - July	45
Figure 13. 3 Cluster Analysis for FL Region - July	47
Figure 14. Mean OD-CTH Histogram over SA Region - July	49
Figure 15. Silhouette Scores over SA Region - July	50
Figure 16. 2 Cluster Analysis for SA Region - July	51
Figure 17. 3 Cluster Analysis for SA Region - July	53

IDENTIFYING CLOUD FIELD REGIMES FROM WORLD WIDE MERGED CLOUD ANALYSIS VIA K-MEANS CLUSTERING

I. Introduction

The impact of clouds and the range of effects they have on Earth's global climate system are extensive and a topic of interest and concern. Clouds on average cover about two-thirds of the Earth, with large spatial and temporal variability. Differences in total cloud cover over land and over the ocean can range by roughly 17% alone, and modeled representation of this variance is lacking (King and others, 2013:3826). Therefore, applications in remote sensing are all the more pertinent given the relationship between radiative effects and clouds with respect to satellite sensors; varying sensors will have different representations of cloud parameters, such as total cloud cover and optical depth, based simply on sensor spatial resolution (Wielicki and Parker, 1992:12799). Furthermore, coherent characterization and representation of clouds within numerical models and simulations continues to pose a challenge despite our current understanding of how this particular weather phenomenon changes both in the dynamic and microphysical sense. While these hurdles have long made the depiction of clouds within models an ongoing obstacle, continued efforts towards improving the foothold of accurate cloud representations in the real world will remain key towards overcoming this challenge.

Analytical approaches towards representing cloud distributions have long been varied and conducted. For example, Warren and others (1988) utilized ground-based observations as the main data source to describe the global cloud climatology in a series of four atlases accounting for total cloud cover, type, frequency of occurrence and more. Other studies have offered additional inquiry on regional and global cloud cover and cloud distributions. Such studies either took advantage of additional surface observations as the temporal resolution increased, increased satellite observations as the implementation of growing technology from an above point of view progressed and became more frequent, or from a combination of these advances (Warren and others, 2007:717; Söhne and others, 2008:4421; Hahn and others, 2001:11; Xi and others, 2010:1). In addition, variability within similar datasets has become a substantive point of interest. For example, satellite data offers varying levels of agreement based on cloud properties and which sensors and satellites are used for comparison (Marchand 2013:1941; Karlsson and Devasthale 2018:1; Stubenrauch and others, 2013:1031). Therefore, as data between these and other datasets continues to accumulate and evolve, additional analyses and inter-comparisons will be necessary to continue to elevate our knowledge of cloud distributions.

Atmospheric numerical modeling often involves the use of parameterizations as part of the modeling process, which is a means of replacing certain atmospheric phenomena, due to their complexity or size relative to the resolution of a model used, with simplified descriptions and approximations that allow for physical representation of these phenomena. Overcoming the task of cloud parameterization has been an issue that

has challenged the meteorological community for decades given the range and complexity of scales clouds fall under such as those pertaining to microphysical processes and radiation interactions. Many strides and techniques have been developed, however, to address this problem. One such technique involves the usage of cloud system resolving models embedded within a broader general model dubbed superparameterization (Randall and others, 2003:1547). Another technique involves a step-by-step process of parameterization improvement. By applying a composite of observations and model output to a specified criterion tied to cloud formation/maintenance and focusing on errors consistent between the composite average and model, adjustments to cloud parameterizations can be made to offer gradual improvements when looking at case studies (Jakob, 2003:1399). A key component of the latter approach relies on having available long-term datasets with the necessary information to properly evaluate how a model is performing. As such, one goal of this study is to determine the quality and viability of a particular dataset used towards cloud regime identification and classification.

A number of numerical models currently in use or being developed, rely on an increasingly accurate representation of clouds to be able to capture their effects both temporally and spatially. One example are general circulation models (GCMs), a type of climate model that employs a mathematical representation of Earth's global circulation. A secondary example are numerical weather prediction models (NWP), which incorporate current meteorological observations to forecast the future state of weather. While these numerical models serve to predict some future state of weather or climate,

other numerical models of interest seek to emulate an aspect of interest. One such tool, developed by the Air Force Institute of Technology (AFIT), is the AFIT Sensor and Scene Emulation Tool (ASSET), a physics-based image-chain model where synthetic electro-optical and infrared (EO/IR) sensor data are generated alongside realistic artifacts (Young and others, 2017;10178A-1). ASSET, a toolset which is openly available to the user community, allows the user to view or generate a scene image with simultaneous viewing from a number of heterogeneous EO/IR sensors, subsequently enabling multi-sensor fusion as well as the usage of a constellation design for said sensors (Steward, 2020). Clouds act as a major source of cluttering for EO/IR sensors, however, and are difficult to remove with background suppression given the range of motion and morphology they undergo, even at one particular location.

Assessments of cloud distributions have shown probable relationships between distinct cloud regimes/distributions and cloud parameters. Early studies, such as those conducted by Lau and Crane (1995) and Tselioudis and others (2000), would identify certain “dynamical” regimes by analyzing parameters such as sea level pressure anomalies (SLPA). Three distinct dynamic regimes for SPLA were created by identifying the 50th percentile points that equally divide the positive and negative frequency distribution of SPLA. Values above the positive 50th percentile were defined as “positive-SLPA”, below the negative 50th percentile point as “negative-SLPA”, and in between the two points as “zero-SLPA”. Then, available cloud observations from the International Satellite Cloud Climatology Project (ISCCP) were connected to these SLPA regimes to create a unique relationship of associated cloud regimes, noting how the

background cloud field utilized in this study is modulated depending on which SLPA regime is present. Jakob and Tselioudis (2003) expanded upon the work of creating distinct cloud regimes by forgoing their connection to dynamical parameters . They concluded that a given region's cloudiness can be attributed to a set number of distinct cloud regimes, of which having set dynamical regimes is not necessarily a requirement (Jakob and Tselioudis, 2003:4). They showed that, with just observed cloud information alone, distinct cloud regimes can be identified via cluster analysis. Cluster analysis is an unsupervised learning algorithm that seeks to create groups (clusters) within a dataset where the members of a particular group share more similar properties to each other than from those of another group. The cluster analysis performed by Jakob and Tselioudis (2003) was conducted on joint histograms of cloud top pressure and optical thickness values provided by the ISCCP, offering a simple yet effective method of binning certain cloud types and their frequency of occurrence over their areas of interest.

This study focuses on utilizing a global cloud dataset to identify distinct cloud regimes over a myriad of locations through all hemispheres as an extension of previous work conducted by Jakob and Tselioudis (2003) using similar techniques. The World Wide Meteorological Cloud Analysis (WWMCA) dataset will serve as the global cloud dataset to be used, offering a number of cloud properties at relatively high temporal resolution (1-hr) over a large, quasi-global area. Reliable results from previous studies that utilized a global satellite cloud dataset with the goal of creating cloud regime classifications (Tselioudis and others 2000:312; Jakob and Tselioudis 2003:2) offer a solid foundation for performing cloud data analysis with contemporary machine learning

capability. Previous comparisons between a number of varying satellite datasets have been conducted (Marchand and others, 2010:1), as well as the strengths and shortfalls of the WWMCA dataset on cloud detection (Pasillas, 2013:53). However, little work has been done with the WWMCA dataset regarding cloud regime identification and representation. Therefore, the primary purpose of this study is to develop global classifications and regimes of clouds, and to test if the regimes derived from this unique dataset can be successfully utilized by the high temporal and spatial resolution ASSET model.

Section 2 describes the data used for this study as well as the analysis techniques in further detail. In section 3, results pertaining to cloud regime identification are discussed. Finally, section 4 presents concluding remarks and possible future applications of this study.

II. Methodology

ASSET Background

One of the primary motivations for examining cloud regimes is to improve scene generation and analysis from ASSET products. ASSET serves as a physics-based image-chain model where, upon being given a source image to serve as the basis for a background scene, emulates a number of remote sensing processes and outputs a dynamic array of data frames. Originally intended for internal use at AFIT for student research pertaining to acquiring absolute knowledge on object position and radiometric signature (detection/tracking) or the usage of large datasets (machine learning), ASSET allows one to test various sensor configurations while remaining low on computational costs and model how a sensor responds to differing levels of irradiance at-aperture. Users of the model are able to specify their desired tunable parameters via an ASCII text file.

ASSET fills a unique role between various simulations in its field of view and capability; where high-fidelity simulators are focused on a small field of view without expending high computational costs and more simplistic simulators allow for basic analysis at the cost of explicit information on elements that can affect calculations done on data, ASSET falls in between these ends of the spectrum. Rather than simulate a scenario, ASSET allows one to quickly emulate data from a number of sensors spanning a wide field of view (WFOV) that is representative of real WFOV sensors without the high computational expense; the value gained from realistic but time consuming calculations of a particular situation (such as ray tracing) may not be worth the expense compared to simpler computations and emulated aspects, of which ASSET's value may

be shown. It should be noted that situations which require exact radiometric properties may still be better suited using higher fidelity models such as the Digital Imaging and Remote Sensing Image Generation (DIRSIG) or the Monte Carlo Scene (MCScene) model. For additional information concerning a more in-depth overview of ASSET, its parameters and how it operates, readers are encouraged to read Young and others (2017), who provide an extensive look at the capabilities and shortfalls of the ASSET model.

Within ASSET clouds serve as a source of necessary clutter given their impact on signal processing, where users are able to introduce artificial clouds into a scene by adjusting parameters within the mentioned text files. The importance of proper cloud representation within a scene is an ongoing issue; clouds change not just their position but shape over a given time, altering the performance of algorithms on other aspects of a scene given this source of potential error to radiometric values as a sensor operates. By introducing clouds into a scene, users are able to refine proper sensor usage and viewing angles that can be representative of real world situations without the potential high computational and time sink costs of running more complex algorithms and higher fidelity models. Herein lies the issue of proper cloud representation within ASSET. For most users, altering the numerous list of parameters within ASSET may be unfeasible and extremely tedious for all given situations. Users may opt to alter what parameters they deem necessary while simply providing more broad limitations on the majority of parameters, allowing ASSET to handle the grunt work of generating and placing an object such as clouds over a scene. Currently there is not an extensive dataset within ASSET that will allow it, given little or no user input, to place clouds representative of

their distribution based on geography and season at various locations around the globe. While the generation of a random cloud field is doable given the current ASSET iteration, that does not necessarily mean that the clouds generated would be characteristic of what type of clouds that would be typical over a given location.

WWMCA Background

WWMCA data is used in this study as the main source of satellite data to examine cloud regimes, of which is produced by the United States Air Force's Cloud Depiction and Forecast System II (CDFS II). The CDFS II, an upgrade to previous processing systems utilized by the Air Force Weather Agency (AFWA), continues a nearly 40 year long record of global cloud analysis for real time analyses and future forecasts of cloud cover and microphysical properties to be utilized by the Department of Defense (DoD) for its operations. While earlier models relied more solely on available data from polar orbiting satellites such as the Defense Meteorological Satellite Program (DMSP) for its cloud analysis, the operational implementation of the CDFS II marked a notable shift towards utilizing available data from both government and commercial sectors (Brown, 2002; Horsman II, 2007). The CDFS II produces its global cloud analysis by combining the imagery of five geosynchronous and four polar-orbiting satellites, merging the results into a single image. Cloud analysis is done via a threshold technique, where temperatures sensed that are colder than the expected background temperature or brightness values sensed that are brighter than the background warrant the placement of a cloud at that particular grid point. Background temperatures and visual brightness values are initially provided the Surface Temperature (SFCTMP) analysis model and Snow Depth & Sea Ice

(SNODEP) model respectively, though the former model offers manual adjustments to values fed to account for differing characteristics between the incorporated satellites used and for the geography of the background being examined over. Values such as cloud top height are assigned by comparing brightness temperatures of cloud tops from derived satellite values to global NWP derived vertical profiles that match the same temperature value.

WWMCA products are produced hourly, projected onto a global domain that consists of two hemispheric polar-stereographic grids, allowing for a global view of all regions on Earth. Horizontal resolution for the gridded data is at 24km, with 4 floating layers of vertical resolution available for select variables. Current available provided data from WWMCA extends far back to 2002, though prior to 2014 the number of available variables pertaining to clouds were limited. From 2014 and onward, additional variables useful for the probable examination of cloud regimes such as cloud layer optical depth and cloud layer water path became available, prompting a slight restriction to what years would be available for use.

Four years of cloud data between 2014-2017 from the available WWMCA dataset were examined by extracting a number of cloud parameters from all existing files and regrouping the data back into one singular file that corresponds to a particular month for all four years. The cloud parameters extracted are as follows:

1. Cloud Top Height (CTH)
2. Optical Depth (OD)
3. Cloud Layer Particle Size (CLPS)

4. Cloud Base (CBH)
5. Cloud Layer Water Path (CLWP)
6. Cloud Type (CT)

Note that while a total of 6 cloud parameters were extracted from each individual file available from the WWMCA dataset, CTH and OD remained the parameters of choice to further progress through this study with the remaining parameters stored aside for potential inquiry in examining various relationships between different cloud parameters and potential future studies. It should be noted that while WWMCA data has an hourly temporal resolution, the number of files available for each month and year analyzed is not consistent due to gaps of missing data. Part of the process for regrouping the individual data files also involved filling in hourly values of missing data with a blank numeric data value, subsequently known as "not a number" or NaN via Python. This method of assigning blank numeric data values into the overall dataset allows one to handle the vast majority of available data files without worry, following proper masking techniques for these data files, of any future calculations being affected by the varying number of data files such as determining the average of a cloud parameter for a particular month.

Selected Regions of Interest

Two target locations were selected as areas of interests for this study as shown in Figure. 1 below to examine notable and expected distributions of clouds, covering a wide range of different cloud structures. One location focuses on the Florida peninsula and surrounding waters, now addressed as the "FL region", that spans the area between 24.5° to 31°N and 79.75° to 87.75°W. Florida itself has long been studied concerning sea

breeze fronts and associated convection (Nicholls and others, 1990;1), where increased opportunities for convective systems to develop over an extended period of time, such as during the summer months, allows such an area to be a suitable choice for analyzing clouds associated with convection such as cumulonimbus and cirrus anvils, as well as low/mid level clouds associated with daytime convection or mentioned sea breeze fronts.

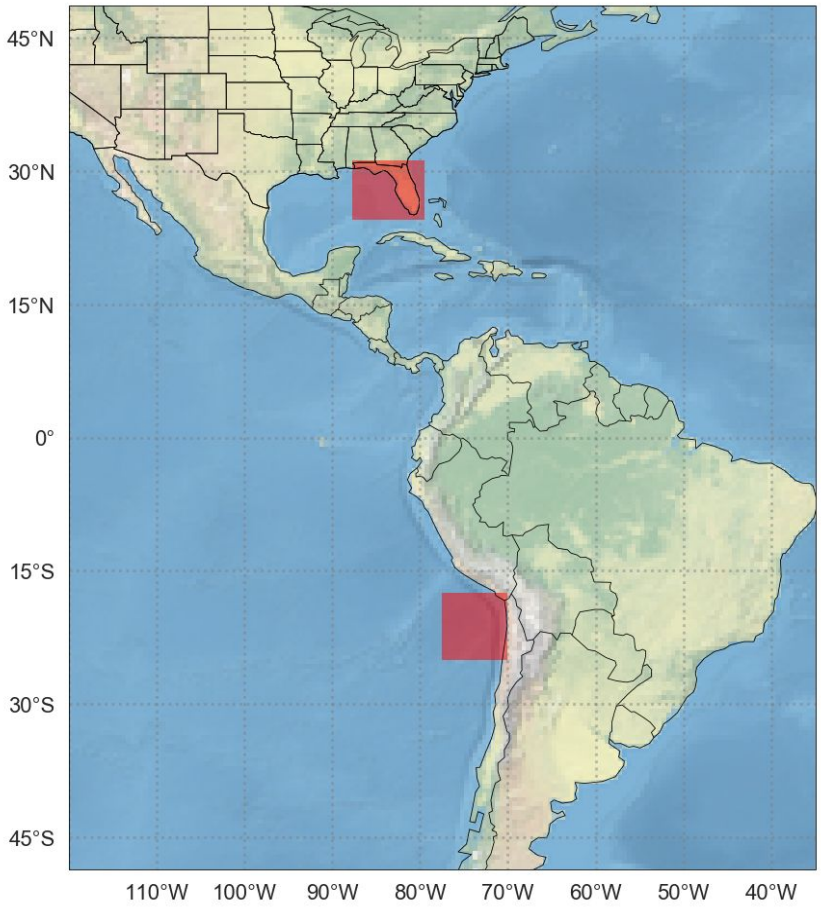

Figure 1. Geographic view of North, Central, and South America with selected regions examined for this study shown in red filled-in boxes. Northernmost box covers the state of Florida, termed the "FL region" (24.5° to 31°N, 79.75° to 87.75°W), while the southernmost box focuses off the western coast of central South America, termed the "SA region" (17.75° to 25°S, 70.25° to 77.5°W)

A second location focuses on the Pacific Ocean off the central west coast of South America, now addressed as the "SA region", that spans between 17.75° to 25°S and 70.25° to 77.5°W. This area is noted for having not only an extensive area of coverage, but also "the most persistent subtropical stratocumulus cloud deck in the world" due to upwelling and colder currents interacting with drier subsiding air aloft, of which both factors offer increased favor in the development of a low level temperature inversion and boundary layer clouds (Bretherton and others, 2004:967; Wang and others, 2004:274). Such a location would be very suitable as a choice for attempting to identify low level cloud regimes from the WWMCA dataset despite a weakness of under identifying those same clouds.

Development of Joint Histograms

To examine the relationship between cloud parameters and regime classification, joint histograms between these selected cloud parameters were created. Previous studies that delved into examining similar cloud properties and cloud regimes utilized available joint histograms provided by their dataset of choice, commonly products offered by the ISCCP (Jakob and Tselioudis, 2003; Tselioudis and others, 2000). WWMCA data does not offer products such as those from the ISCCP, requiring manual creation of these joint histograms.

To begin the process each individual netCDF4 file, which now comprises all hourly data points over a specific month, has the noted cloud parameters placed in one of two respective data arrays in Python. Initially, these arrays are two-dimensional (2D) in their own regard where the value of a data point is located at its noted latitude and

longitude, specified based on both of the noted regions being examined. These 2D arrays are converted into one-dimensional (1D) data arrays where each array is now simply the list of all data values across a grid space for one cloud parameter. These individual 1D data arrays are then combined together to create joint histograms between CTH and OD. Predefined bin edges are utilized for both parameters that comprise the joint histogram, and help serve two purposes. Maintaining a consistent number of bins for each respective axii allows for easier manipulation of multiple histograms when performing calculations that would normally be extremely dependent on the actual value of the data points within each histogram, which in turn would define bin ranges if created through more automatic methods. Additionally, a predetermined set of bin ranges allows for greater control on the restriction or representation of available data, which may serve as a hindrance or boon depending on range values selected.

As the focus of this study pertains to available clouds over an area of interest, zero values of CTH and OD are not included to rule out clear conditions over the two selected regions. CTH values start between 500m to 3000m, then subsequently increase in increments of 3000m up to 18000m. These ranges allow one to capture notable low-level clouds while also maintaining the ability to capture most clouds within the troposphere even at regions closer to the equator where the troposphere height is higher compared to regions closer to the mid-latitudes. OD bin value ranges are represented via a logarithmic scale that ranges from 0.1 to 60 (note that OD is a unitless value). The colorbar scale, also represented via a logarithmic scale, represents the frequency of occurrence (FO) for each bin range and is displayed as a percentage of the number of values that fall within a

selected bin range. FO values are calculated by first dividing the values generated from joint histogram calculations by the total sum of all values over that same histogram. This value is multiplied by 100 to offer a percentage value for each respective bin range, whereby the sum of all values will equate to 100. Percentage values below 0.1 are excluded in joint histogram plots to minimize displaying bins with significantly low values. The use of a logarithmic scale is incorporated due to the spread of data pertaining to FO and OD values; the visual representation of these two parameters within a joint histogram using a linear scale proved insufficient. Given the spread of values among OD data values, as well as keeping in line with bin ranges of similar joint histograms such as those part of the ISCCP dataset, the implementation of a logarithmic scale was deemed justified. This process is repeated for all hourly data values for the months of January and July for each year between 2014-2017, leading to four distinct arrays of data for each individual year per selected region.

These data arrays are then averaged together to determine the overall mean joint histogram over a selected region across all four years of data per month. As noted earlier, not all of the data was available for all hourly increments due to missing gaps within the WWMCA dataset. Application of blank numeric data values and masking techniques on these data arrays allows for the overall average to be computed while minimizing the difference in mean calculations across all hourly increments as best as possible as some hourly values may be averaged between four or fewer values due to these gaps. 24 total monthly average joint histograms were created between each respective region, serving as the basis for subsequent joint histograms created via cluster analysis.

K-Means Clustering Process and Implementation

To further determine the distinct relationships between cloud parameters and cloud regimes, the statistical method of cluster analysis is used. Cluster analysis is a task process that looks to examine a dataset and determine probable clusters within that dataset. These clusters consist of a cluster center, or centroid, and individual data points. These individual data points are assigned to a specific centroid based on some form of distance measurement between the data point and centroid, creating the unique clusters for a dataset. While there are a variety of clustering algorithms available for usage in partitioning data, K-means was the chosen method given its widespread applications and simplicity in working with multivariate observations and datasets (MacQueen 1963:289; Morissette and Chartier, 2013:15)

K-means clustering follows much of the basis laid out for general cluster analysis (Pedregosa and others, 2011:2825). Given an N-dimensional population dataset, the number of centroids incorporated into a dataset is determined by the number of k groups that is chosen by the user. The individual data points are then assigned to a specific centroid based on their euclidean distance from all of the centroids, where the smallest distance calculated determines which group a particular data point will belong to. The euclidean distance calculation follows the formula below:

$$d(x,y) = \sqrt{(x_c - x_p)^2 + (y_c - y_p)^2} \quad (1)$$

Where x_c and y_c refers to the coordinates of a cluster centroid, x_p and y_p refers to the coordinates of a datapoint from the WWMCA dataset, and $d(x,y)$ refers to the euclidean

distance between the two selected points. Once all data points have been assigned to a particular group, a new mean value for each group is calculated using the assigned data points to determine subsequent new cluster centroid coordinates, which is as follows:

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, \; y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

Where n denotes the number of data points that belong in a particular cluster, and x_i and y_i denote the respective coordinate values of a datapoint. Successfully completing and updating the centroid coordinates marks one full iteration through the clustering algorithm. This process of determining and adjusting centroid locations is repeated until the coordinates of the centroids cease to be updated, indicating a convergence within the clustering process.

A number of tunable parameters are available prior to running the k-means algorithm that can affect the output of the algorithm. To maintain consistency, most tunable parameters were kept at default values where the maximum number of iterations for a single k-means algorithm run was set to 300 and the maximum number of times the k-means algorithm would be run with differing centroids was set to 10. One important tunable parameter set was the random state of the algorithm, whereby keeping this value undefined would create a scenario where every total run of the k-means algorithm, even with the same cluster value set for each run, would result in different centroid initializations and subsequently different end cluster groupings. As the optimal value for clusters in the k-means algorithm is susceptible to a number of aspects (quantity of data, spacing of data points, subjectivity, etc.), multiple k-means runs would be performed over

each month for each region. To ensure that the aspect of randomness between identical runs for one location was minimized, the random state value for all runs of k-means was set to a value of 10. Note that potential values for the random state parameter can be any integer value, where 10 was arbitrarily chosen as the set value. Different random state values were run to test whether the end results would change significantly between each value set. In short, silhouette score values differed significantly between each different random state value, though the end cluster results remained nearly the same save for slightly varying cluster frequency values. This defined random state variable also allows others to perform the exact same calculations when using the WWMCA dataset as described above, allowing easy reproducibility of obtained results.

Standardization of Data for Clustering Process

One issue with a direct cluster analysis between varying cloud parameters falls under the method of determining which data points fall under which group. As the euclidean distance between data points and initial centroids is calculated to determine data point groupings, the weight of the parameters being examined calls for some form of normalization given up to a four order magnitude difference between OD and CTH. One method of normalization is the basic statistical standard score, or z-score, which is a measurement of the number of standard deviations a data point is above or below the population mean. Z-score was chosen as the method of standardization given its simple reduction of scale and variable components and for its noted strength when used with K-means clustering (Mohamad and Usman, 2013:3302). To address this issue reshaped data, modified during the joint histogram creation process, was subjected to Z-score

standardization, whereby z-scores were calculated for a given array of data to be used for clustering purposes by utilizing the following equation:

$$z = \frac{x-\mu}{\sigma} \tag{3}$$

Where x is the raw value for a datapoint, μ is the sample mean value from all data points, and σ is the sample standard deviation. These Z-score values were then subsequently run through the K-means algorithm to identify cluster groupings.

Cluster Evaluation - Silhouette Score

Data arrays between CTH and OD serve as the dataset being fed through the K-means clustering process following necessary standardization. The number of K clusters available to choose from will determine possible groupings between these cloud parameters, of which choosing an ideal value for K has some form of subjectivity to it depending on the type of data being examined and whether one has priori knowledge on possible labels within the dataset. Labels in this case refer to some type of augmentation to the dataset where an informative tag is placed upon said datapoints, which can serve to categorize or highlight an aspect of the data deemed noteworthy. Given that the dataset used does not fall under the "labeled" data category, the use of intrinsic methods to evaluate our clustering algorithm becomes feasible.

One method of determining the validation of a cluster analysis and subsequently the ideal number of clusters to use is by calculating the silhouette value (hereby silhouette score) for these clusters of data. The silhouette score ranges from [-1,1], and is a measure of how well an object belongs within a particular cluster by comparing the tightness and

separation of objects within and between clusters (Rousseeuw, 1987;56). The silhouette score for a single data point is given as:

$$s(i) = \frac{b(i)-a(i)}{max\{a(i), b(i)\}} \tag{4}$$

Where $s(i)$ is the silhouette score for a datapoint i, $b(i)$ is the minimum average distance from i to all other clusters that it does not belong to, and $a(i)$ is the average distance between i and all other data points within the same cluster. Silhouette score values near positive 1 indicate that a datapoint is compact and well matched to its assigned cluster, while values near negative 1 indicate that a datapoint is mismatched and not placed within an appropriate cluster. Low and negative values also indicate that the number of clusters used within the clustering algorithm may either be too few or too many in count.

To aid in determining the optimal number of clusters to proceed with, the average silhouette score for our k-means clustering model was calculated by training said model using K clusters that ranged from [2,9] clusters in value. These silhouette scores for all clusters were then plotted out in a simple line graph to illustrate which defined value set for K may represent the most optimal number of clusters. Silhouette scores were plotted using a heuristic technique known as the "Elbow Method", an additional method of cluster verification whereby the silhouette score point that leads to a more linear change of silhouette score values following this point would be marked with a dashed vertical line and be initially set as the optimal value for K. This initial marking is not always guaranteed for each plot as it heavily is dependent on the data being examined, any preprocessing techniques utilized, and the value of the data points themselves. Excessive overlapping of data points or worse, data points that are assigned to incorrect clusters,

would affect silhouette score values, and may easily or be able to distinguish a value of K that could be detected as the optimal one. Depending on the month and region being examined, multiple peaks of silhouette score values may be seen, calling for additional examination on the number of optimal clusters by re-running the k-means algorithm multiple times using different values for K.

Silhouette scores were calculated and plotted for both the FL and SA regions for the months of January and July. If automatically detectable via elbow method application on silhouette score plots, this value of K would serve as the number of clusters analyzed for one run through the k-means algorithm. Any notable peaks of silhouette score values would also be examined as well and subsequently run through the k-means algorithm.

Remapping of Cluster Grouping Identification

The end result of successful convergence from k-means clustering provides a new array of data whose Z-score values were each divided into their unique groupings based on the number of clusters set within the clustering algorithm; each Z-score value is assigned an integer value that denotes which group it pertains to. To illustrate these groupings for our unstandardized datasets, Z-score integer values are combined with our previous unstandardized joint histogram arrays that contain a four year average of hourly data, such that each hourly joint histogram is now assigned to its respective cluster grouping based on Z-score cluster grouping assignment. These clusters are broken up into their own individual data arrays by sorting for the respective cluster integer value, where the relative frequency of occurrence (RFO) for each cluster is calculated by dividing the number of hourly joint histograms that fall within a particular cluster by the total number

of hourly joint histograms available for the month. This total value depends on the number of days that belong to each month, being either 672, 720 or 744 for February (discounting leap years), 30 day and 31 day months respectively. Following cluster grouping separation, the overall mean for each cluster is calculated and plotted out again as a new joint histogram for comparison. To aid in comparison, bin count values were also extracted from each joint histogram to allow for focus on more centralized data points that offer some distinction between these clusters.

III. Analysis and Results

Chapter Overview

WWMCA data over four years from 2014-2017 are analyzed over two target regions (FL and SA regions) for each month to examine optimal cluster number groupings for the identification of prevalent cloud regimes over their respective regions. Results gathered are broken up first by month and then by region, and are presented below.

January - FL Region

The mean relationship between OD and CTH for the entire month of January over the FL region is shown in Figure 2 below, where all hourly joint histograms between OD and CTH from between 2014-2017 are averaged together. OD follows a logarithmic scale given the range and distribution of its values compared to values of CTH. Both CTH and OD do not account for non-zero values. The FO of data points that fall within each listed bin category is also displayed on a logarithmic scale again due to the distribution of values within each of the bins. For the month of January, the highest FO falls into three noticeable sections; relatively medium OD values (5.4-12.1) that extend from 500 to 12000m, mid-to-upper level clouds (9000-12000m) that vary in OD range, and optically thick (27-60) clouds that extend upwards from 500 to 9000m. Maximum cloud top heights between all four years was capped at roughly 15000m, with negligible FO values of OD at this layer below values of 2.4. The highest FO values are noted at occurring at the lowest CTH bin range (500-3000m), where OD bin ranges [5.4,12.1] and [27,60] hold roughly 18% and 12% of all plotted values respectively.

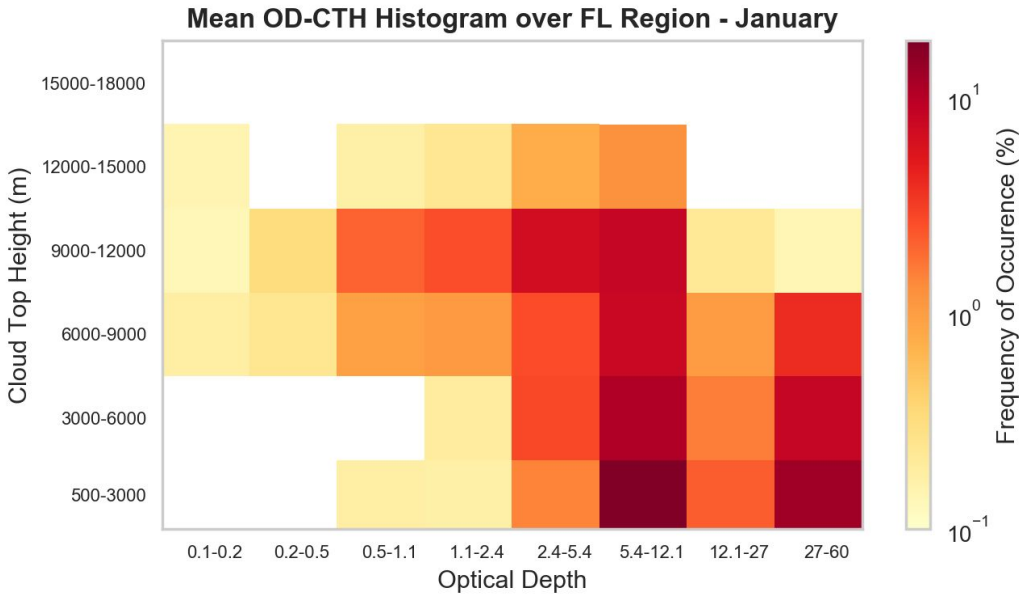

Figure 2. Mean joint histogram of CTH and OD averaged for all hourly histograms over the Florida (FL) region (24.5° to 31°N, 79.75° to 87.75°W) between 2014-2017 for the month of January. OD and FO color scale depicted on a logarithmic scale.

The overall background cloud field revolves around a wide range of optically thick clouds that retain CTH values of 9000m or less, where CTH values exceeding 12000m with low OD values may be attributed to cirrus clouds associated with convection given the large FO values tied to optically thick clouds that extend through much of the troposphere.

The silhouette scores for the FL region for the month of January can be seen on Figure 3 below where, if automatically detectable via the heuristic elbow method, the optimal value of K clusters to be used in the k-means clustering algorithm is marked with a dashed vertical line. Given a pre-defined random state of 10 for our k-means clustering algorithm, the optimal K value initially appears to be at a value of 6. While

this value is noted via the elbow method, $K = 2$ retains the highest silhouette score of 0.177, compared to a score of 0.175 at $K = 6$.

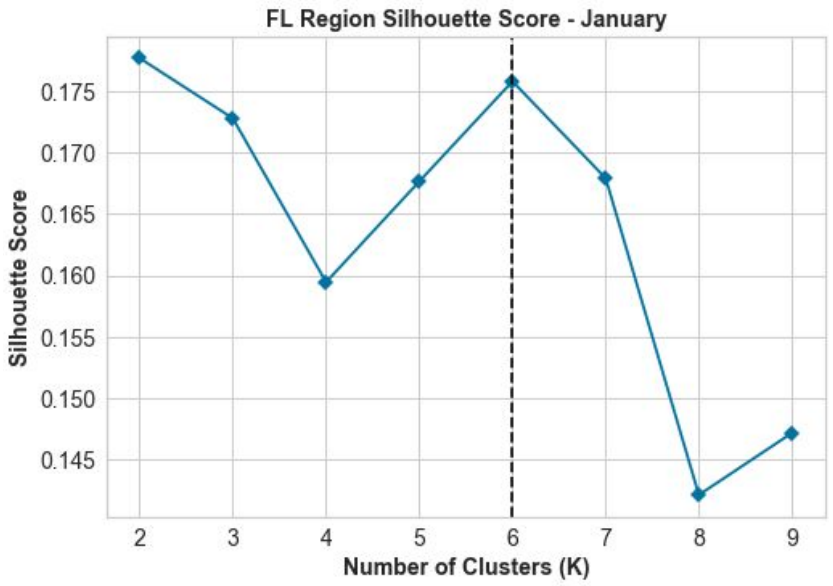

Figure 3. Silhouette score plot over the FL region for the month of January. Score values calculated for a cluster range of 2-9 where, using the heuristic elbow method, the optimal cluster number is marked with a dash line if detectable. Silhouette score values range from [-1,1], with values close to 1 indicative of well defined and well separated cluster groupings.

To begin examination of what these varying cluster groupings appear like visually, each cluster is individually plotted onto its own OD-CTH joint histogram based on the cluster number chosen for the k-means algorithm as shown in Figure 4 below. Much like Figure 2, each cluster is plotted as a mean of all data points that fell within that respective cluster with OD and the FO color scale (not labeled) being displayed in a logarithmic scale. Each cluster also has its RFO values labeled as well, indicating how often a respective cluster occurs

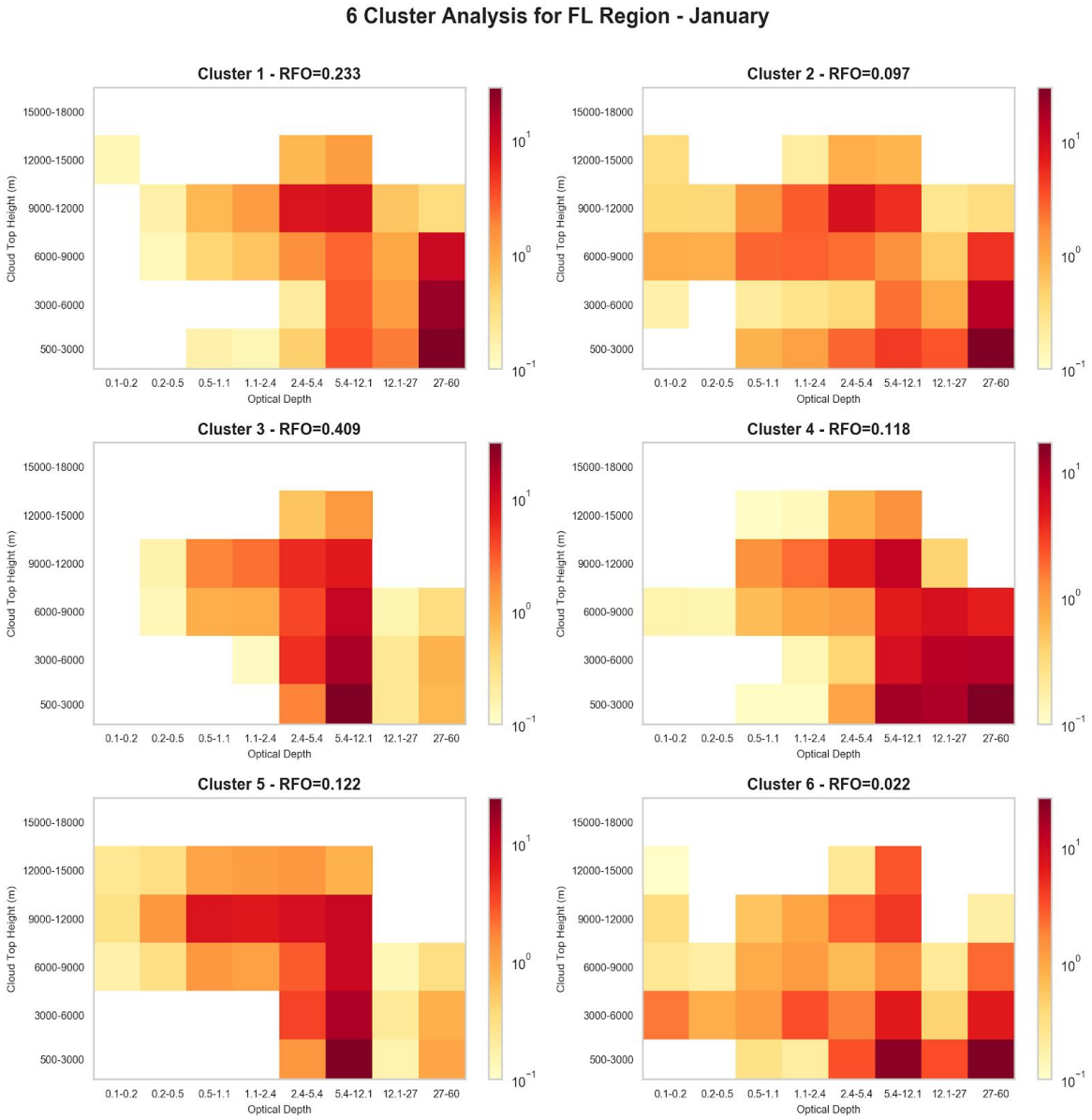

Figure 4. 6 cluster analysis over the FL region for the month of January. OD and FO colorbar are set to logarithmic scale, with RFO values for each respective cluster labeled. Note the varying colorbar scale between each cluster due to the four year average of hourly files difference from missing data.

Figure 4 showcases six distinct groupings of OD and CTH parameters across the FL region. Cluster 1 presents a concentrated spread of FO values across the OD range of

[27,60] with CTH values ranging from 500- 9000m. Similar to Figure 2, a band of relatively optically thin clouds also covers a marked number of FO values between the 9000-12000m bin set. Cluster 1 is the third most commonly occurring cluster with an RFO of 0.233. Cluster 2 attains a much lower RFO value at 0.097, making it the second least common cluster of the six analyzed. Cluster 2 retains a very similar cloud field pattern to cluster 1, with the only significant departures being slightly higher FO values for optically thin low/ level clouds, and an increase in FO values for mid/high relatively optically thin clouds within OD bin ranges [0.5,5.4] and CTH bin ranges [6000,12000]. Cluster 3, with an RFO of 0.409, is the most common cluster out of the six and focuses primarily on clouds than fall within OD bin ranges [2.4,12.1] with varying CTH values. Optically thin or very optically thick clouds have negligible to no recorded FO values for most CTH bin ranges.

Cluster 4, with an RFO of 0.118, retains similar upper level cloud FO distribution as seen in cluster 3, though the majority of FO values are focused on optically thick clouds that extend from 500m to 9000m. Similar to previous clusters, optically thin low level clouds remain absent. Cluster 5, with an RFO of 0.122, showcases two distinct regions of focus: varying CTH values for clouds in the OD bin range [5.4,12.1], and varying OD values for clouds in the CTH bin range [9000,12000]. This distribution of FO values falls more in line with the overall mean for the FL region, save for the focus on optically thick clouds primarily seen in clusters 1, 2, and 4. Cluster 6, with an RFO of 0.022, marks this particular cluster as the least commonly occurring cluster of the six.

Cluster 6 primarily focuses on optically thick clouds at the lowest CTH bin range, with a more spotty cloud field above this CTH bin range and for all other OD bin ranges.

While the elbow method depicted the optimal K value being 6 with a very similar silhouette score value to the maximum at $K = 2$, the number of similar cloud field patterns and notably low RFO values for some clusters called into question whether a lower K value would offer the same distinct cloud fields already observed with higher RFO values. To examine this possibility, a 2 cluster analysis (not pictured) was conducted and found that while the representation of optically thick clouds with varying CTH values and the notable band of clouds at the CTH bin range [9000,12000] was captured, some of the distinct cloud fields generated through higher K values may justify their existence with relatively sufficient RFO values. A three cluster run conducted can be seen from Figure 5 below, and showcases clusters 1 and 3 as the dominant clusters whose combined RFO values equate to 0.895. Cluster 2 serves as a smaller division between the two clusters, sharing high FO values for optically thick clouds (27-60) while also possessing a much more active cloud field for medium/low optically thick clouds that cover much of the mid/upper atmosphere. Compared to a six cluster analysis, a three cluster analysis retains the most distinct and dominant cloud fields which serve as the source for other splintered clusters via higher cluster analysis, while still offering a cloud field that captures a broader range of optically thinner clouds compared to just a two cluster analysis. A four cluster run (not shown) was shown to decrease the RFO of primarily clusters 2 and 3 from the three cluster run for its fourth cluster, of which shares a similar cloud field pattern to cluster 3 of the three cluster run.

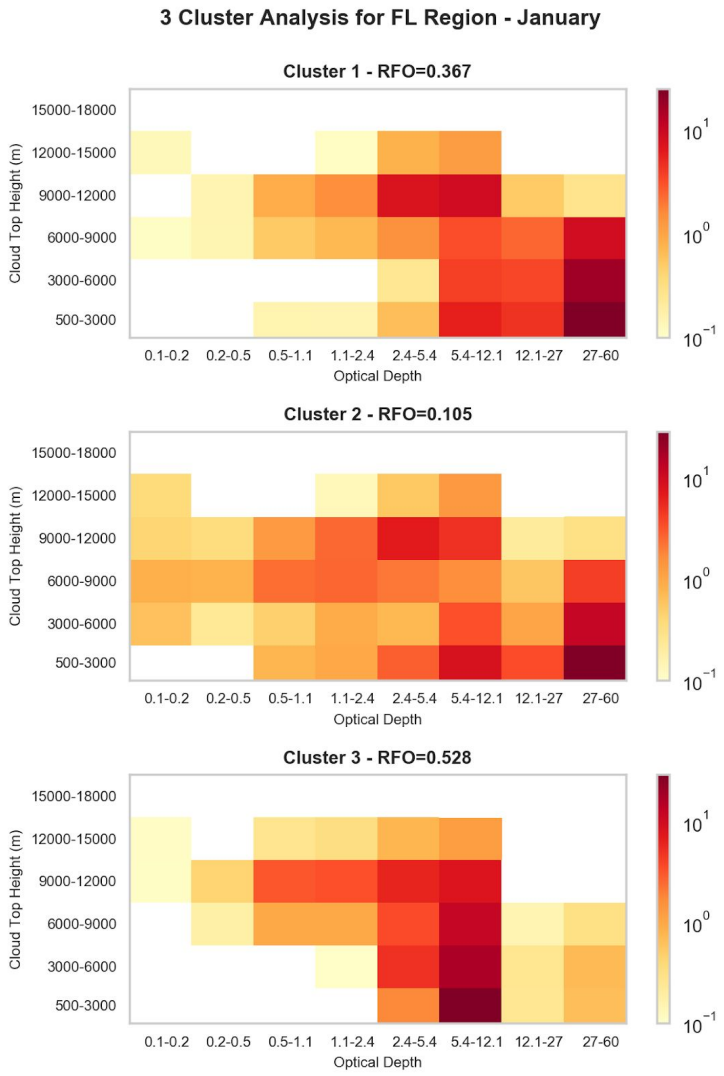

Figure 5. 3 cluster analysis over the FL region for the month of January. Compared to the 6 cluster run, the main dominant clusters remain present with the additional tertiary cluster (cluster 2) offering representation of a more active cloud field with more optically thin clouds in the low/mid levels of the atmosphere.

January - SA Region

The overall mean relationship between OD and CTH over the SA region for the month of January is shown in Figure 6 below. The mean joint histogram over the SA

region shows a marked difference compared to the mean joint histogram over the FL region, expected given the notable differences between the two regions.

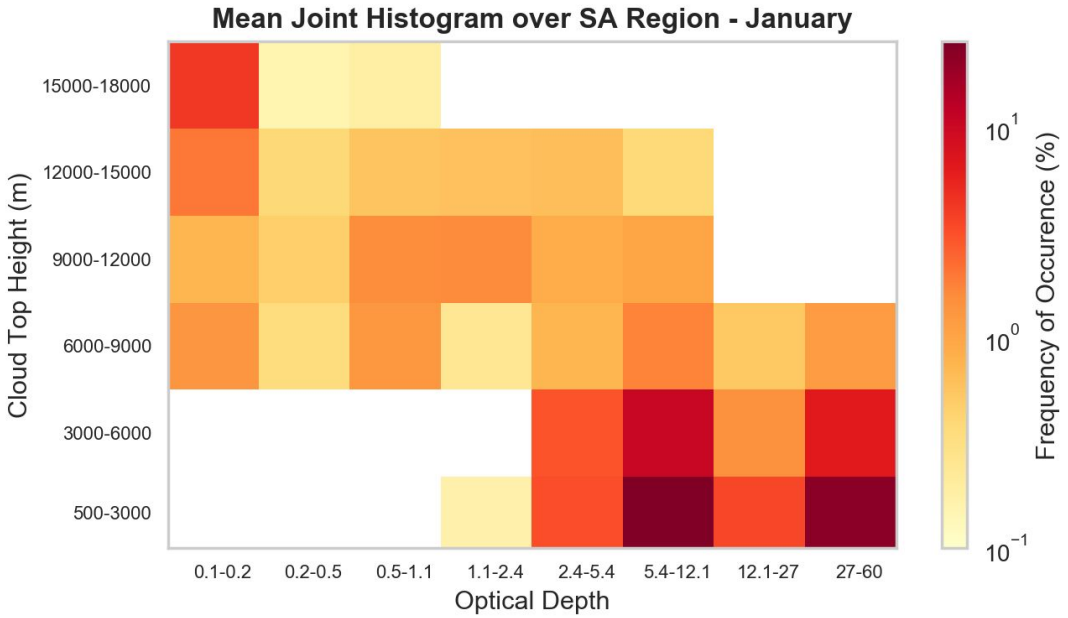

Figure 6. Mean joint histogram of CTH and OD averaged for all hourly histograms over the South America (SA) region (17.75° to 25°S, 70.25° to 77.5°W) between 2014-2017 for the month of January. OD and FO color scale depicted on a logarithmic scale.

While the FL region focuses mostly over land, the SA region focuses mostly over the ocean while being physically closer to the equator with an expected higher tropopause and subsequently higher CTH values. This latter point can easily be seen given the range of CTH values seen on Figure 5 with notable FO values within the maximum CTH bin range of 15000-18000m, values not recorded over the FL region for the same month. The focus of FO values is concentrated on three particular areas of the joint histogram. The first covers a wide range of CTH values that extend from 6000-18000m while also

possessing very low OD values (0.1-0.2), indicating the presence of optically thin clouds that extend through much of the mid and upper levels over the SA region as being prevalent. A second point of interest are the relative mid range OD values (5.4-12.1) that cover much of the low and mid levels of the troposphere (500-6000m), while the third point of interest is high FO values associated with very optically thick clouds mostly associated with the lower levels (500-3000m bin range) while also showcasing optically thick clouds that extend upwards in CTH value to 9000m. No optically thick clouds whose OD value exceeds 12.1 are seen at CTH values above 9000m, as well as optically thin clouds whose OD values are below 1.1 in the low/mid levels. Incidentally, the same OD bin ranges over the SA region hold the highest FO values similar to the FL region, with FO values of roughly 25% and 22% occurring in the OD bin ranges of [5.4,12.1] and [27,60] respectively.

Silhouette scores for the month of January over the SA region are shown in Figure 7 below. Similar to the previous silhouette score comparison for the FL region seen in Figure 3, the SA region initially exhibits an optimal cluster value at $K = 4$, noted for both its maximum silhouette score (0.156) and dashed vertical line that indicates elbow method support. Unlike silhouette scores for the FL region, score values for the SA region for the month of January take a notable decrease in value past cluster numbers of 5, with $K = 6$ offering the lowest silhouette score out of the tested cluster range. Overall though, silhouette score values are lower over the SA region as compared to the FL region during the same month.

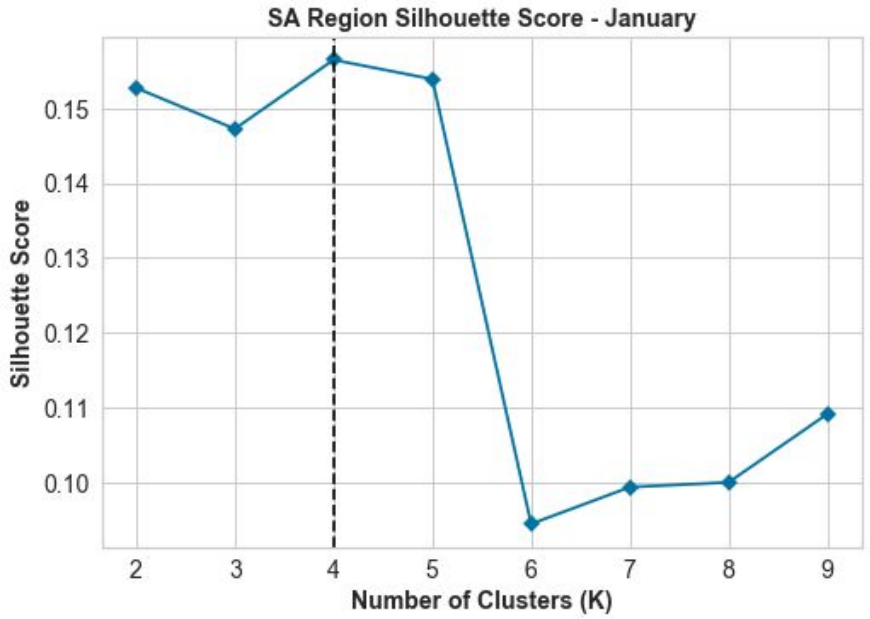

Figure 7. Silhouette score plot over the SA region for the month of January. The optimal cluster number, both in silhouette score maximum value and elbow method, are initially found at $K = 4$. Cluster values above 6 incur a notable relative sharp decrease in silhouette score values.

To further explore the potential optimal value for K clusters given the silhouette score results, multiple joint histogram cluster plots were created to examine RFO values for each unique cluster, and to examine notable differences between each cluster. Figure 8 below shows an initial 4 cluster analysis for the SA region given its maximum silhouette score. RFO values between each cluster indicate that clusters 1 and 3 account for over 85% of all available hourly data points for the SA region. Clusters 2 and 4 only account for roughly 15% of all available data points, with cluster 2 attaining the lowest RFO value out of all of the other clusters with a value of 0.071. Cluster 3, with the highest RFO value at 0.509, primarily focuses on high FO values coinciding with

optically thick clouds within the low and mid levels of the atmosphere as well as mid to upper level CTH values that coincide with very low OD values.

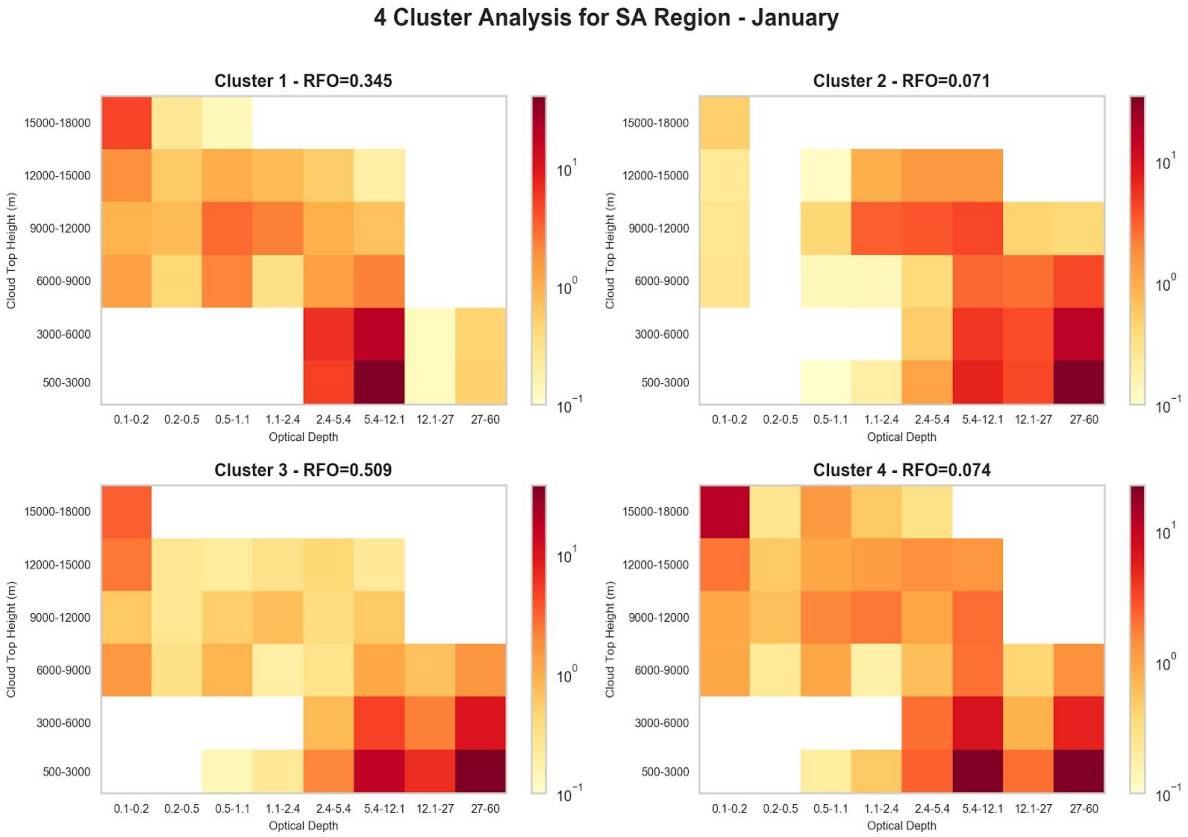

Figure 8. 4 cluster analysis over the SA region for the month of January. OD and colorbar are set to logarithmic scale, with RFO values for each respective cluster labeled. Note the slight variance in colorbar scale between each cluster, affected by the number of data points that fall within each cluster and subsequent four year hourly file difference due to missing data.

The OD-CTH bin with the highest FO value, being roughly 38%, is associated with high OD values (27-60) and low CTH values (500-3000m). Cluster 1 has the second highest RFO value (0.345), with a similar cloud field to cluster 3 save for a stronger focus on more optically thin clouds whose OD values are less than 5.4 in the low/mid levels. The

bin with the highest FO (~40%) is associated with relatively optically thin low level clouds, whose OD values range from [5.4, 12.1] and CTH values range from 500-3000m. Cluster 2 stems from clusters 1 and 3 with a focus on two regions: a section of optically thick clouds whose OD bins range from [5.4,60] with CTH bin values between [500m, 9000m], and a section of thinner clouds (1.1-12.1) in the upper atmosphere (9000m-15000m). Cluster 4 (RFO=0.074), is seen as a derivative from the three other clusters with similar low level optically thick clouds as seen in clusters 2 and 3 and a more active mid/upper level cloud field as seen in cluster 1. Of note is the representation of the highest and thinnest clouds in cluster 4 following the CTH and OD bin ranges of [15000m, 18000m] and [0.1,0.2] respectively, whose FO value for this particular bin range is the highest among the four clusters (~11%).

With two clusters possessing low RFO values in a four cluster analysis run, the reduction of cluster numbers became an area of interest. Looking at a three cluster analysis as shown in Figure 9 below, clusters 1 and 3 remain as the dominant clusters as seen during the four cluster analysis. Cluster 2 in the three cluster analysis shares a similar cloud field pattern to cluster 3 with the focus of FO values confined to the low/mid levels for optically thick clouds. One notable difference between clusters 2 and 3 is the distribution of FO values for medium OD bin ranges (1.1-12.1) with CTH bin ranges of [9000m, 15000m]. Cluster 2 in this case opts to place higher FO values within these bin ranges, where ~11.7% of FO values comprise this section of upper level clouds compared to only ~2.5% of FO values in cluster 3. Additionally, cluster 2 offers minute representation of varying optically thin clouds at the highest CTH bin range, similar to

cluster 1. Though the overall cluster number decreased in value, the cloud fields generated remain unique with relatively high RFO values compared to the higher cluster analysis performed. While four clusters remain a possible choice over the SA region for the month of January, three clusters still offer notable cloud field representation.

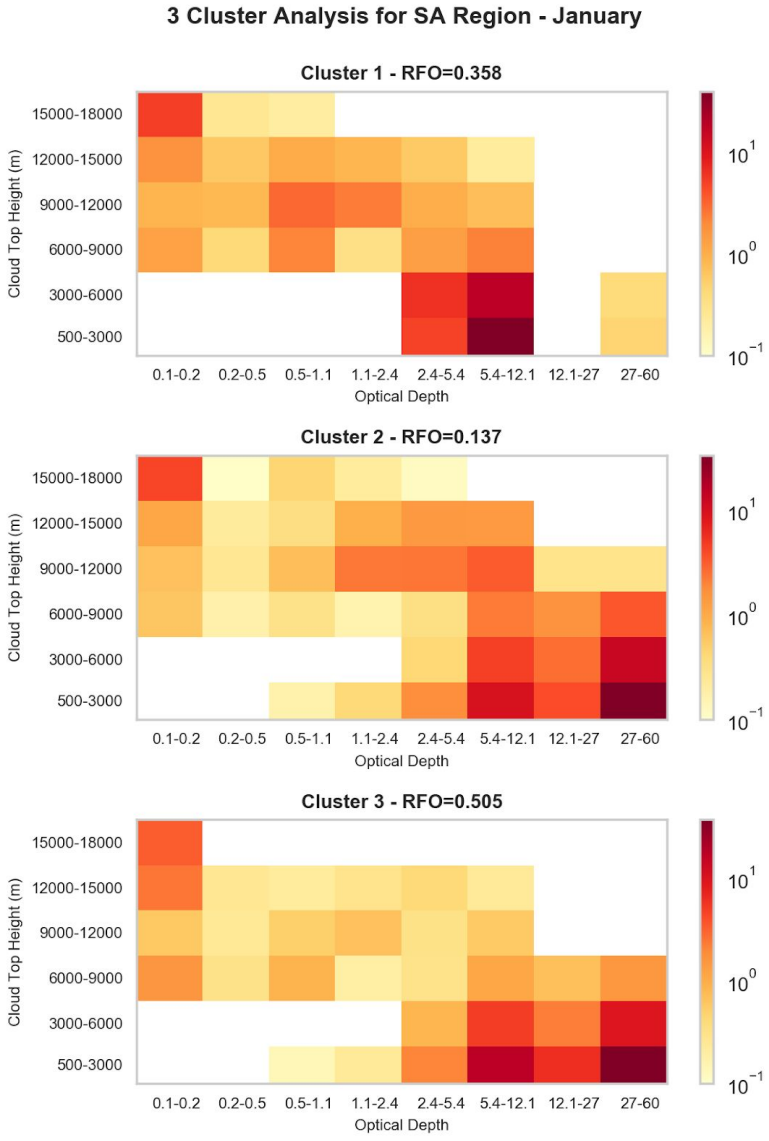

Figure 9. 3 cluster analysis over the SA region for the month of January. OD and FO colorbar are set to logarithmic scale, with RFO values for each respective cluster labeled. Clusters 1 and 3 remain as the predominant centroid groupings with cluster 2 having been splintered mainly from cluster 3.

July - FL Region

The mean OD-CTH joint histogram over the FL region for the month of July can be seen in Figure 10 below. The location of bins with relative maximum FO values is noticeably more varied than for those from the month of January with additional focus on upper level clouds of varying opacity. While bins of high OD values still retain large

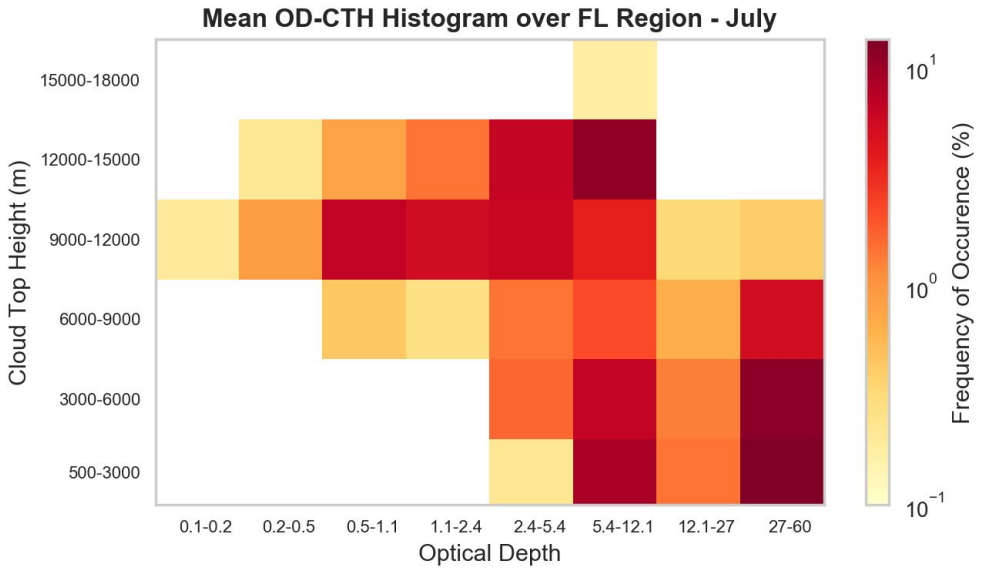

Figure 10. Mean joint histogram for all hourly histograms over the FL region for the month of July. OD and FO color scale depicted on a logarithmic scale, with a notable increase in FO values for upper level clouds whose OD value ranges from 0.5-12.1.

FO values within them, the month of July spreads higher FO values at varying OD bin ranges above 9000m. July highlights two distinct areas of focus: optically thick clouds that extend from CTH bin ranges [500m, 9000m] and a deck of clouds between [9000m, 15000m] whose optical thickness vary between OD bin ranges [0.5, 12.1]. Three bins are noted as having some of the highest FO values. Two of these bins are confined to the OD bin range [27, 60] with FO values of ~13% and ~12% at CTH bins of [500m, 3000m] and

[3000m, 6000m] respectively. The third bin is located at OD bin range [5.4, 12.1] and CTH bin range [12000m, 15000m] with a FO value of ~11%. Compared to the month of January over the same region, this same bin range only held a FO value of ~1%.

Silhouette scores over the FL region for this time period are shown via Figure 11 below, and show a more distinct pattern compared to silhouette scores for January alongside a different optimal K value set initially to 4. Also of note is the higher overall

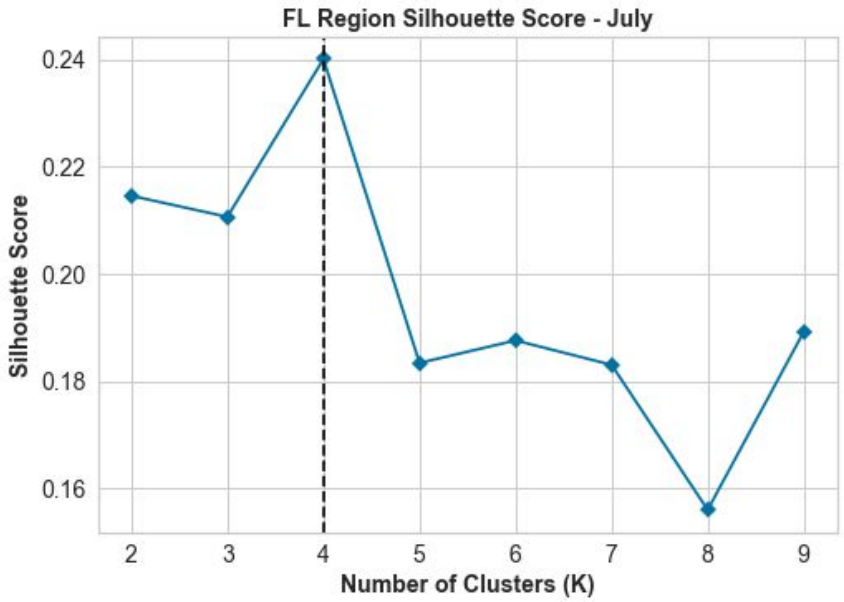

Figure 11. Silhouette score plot over the FL region for the month of July. The identified optimal K cluster number via the heuristic elbow method is found to be at 4, with a notable sharp decrease in score values with higher cluster numbers.

silhouette scores for the month of July with a K value of 4 attaining a silhouette score of 0.24. The identified optimal K number of clusters, together with higher silhouette scores, offer additional support in the number of clusters to choose from when running through subsequent k-means clustering.

Figure 12 below illustrates a 4 cluster analysis performed over the FL region for the month of July.

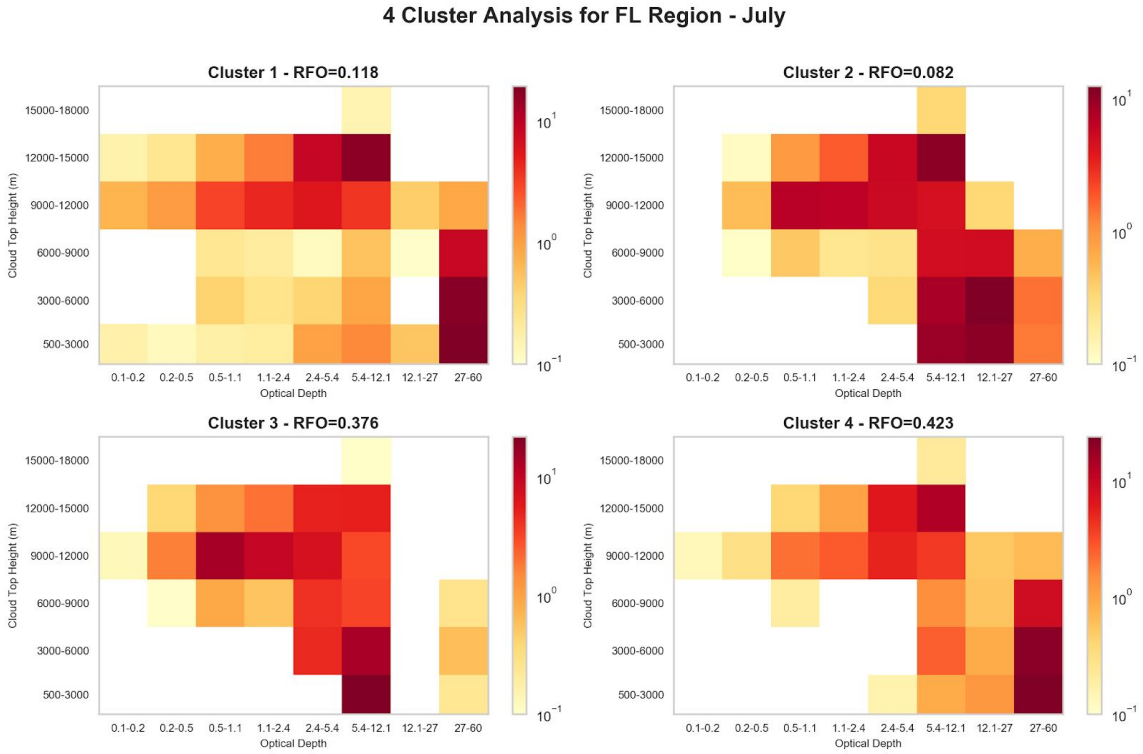

Figure 12. 4 cluster analysis over the FL region for the month of July. OD and FO colorbar scale depicted on a logarithmic scale. Clusters 3 and 4 are shown to be the most commonly occurring cloud fields, with cluster 4 sharing an overall common cloud field as cluster 1.

Cluster 1, with an RFO of 0.118, is the third most frequent cluster of the four with a notable representation of optically thinner clouds in CTH bin ranges below 9000m. FO values are concentrated over two main sections: optically thick clouds within the OD bin range [27, 60] with CTH values below 9000m, and more optically thin clouds whose OD values range from [0.5, 12.1] within CTH bin ranges [9000m, 15000m]. Below these ranges, FO values drop off considerably with optically thin clouds below 9000m

occurring far less often than the OD-CTH ranges mentioned. Cluster 2 has the lowest RFO value of the four clusters being set at 0.082, and is dominated by two cloud fields: optically thick clouds within OD bin ranges [5.4, 27] that extend from 500-9000m in CTH value ranges, and the same spread of varying optically thick clouds within the [9000m, 15000m] CTH bin ranges seen in other clusters.

Cluster 3 is the second most commonly occurring cluster with an RFO of 0.376 with a unique cloud field compared to the other three clusters. Most clouds within this cluster have OD values less than 12.1, with a strong focus on relatively thick low level clouds whose OD and CTH values are between [5.4, 12.1] and [500m, 3000m] respectively (FO~21%). FO value distribution is then concentrated on mid/upper level clouds of varying optical depths that primarily extend upwards to 15000m. Significantly fewer clouds with higher OD values are picked up; no clouds were identified, on average, between the OD range of [12.1, 27]. Cluster 4 has the highest RFO out of the four clusters at 0.423, and shares a very similar cloud field to cluster 1 save for the representation of optically thinner clouds below 9000m. The majority of FO values for cluster 4 are focused on the same bin ranges as noted for cluster 1 with the main difference between the two clusters, aside from FO percentage values across these same bin ranges, being little to no FO values plotted for clouds below OD value of 5.4 and below 9000m.

While 4 clusters was identified as a potential optimal value choice for K, the possibility of reducing the value of K while still retaining representative clusters prompted a 3 cluster analysis as shown in Figure 13 below. Again, this prompt was

obtained by comparing the differing clusters, particular cluster 1 and 4. While cluster 1 offers a more unique cloud field with respect to optically thinner clouds below 9000m compared to cluster 4, FO values below OD and CTH values of 5.4 and 9000m respectively only equated to ~3% difference in this range of cloud representation between the two clusters.

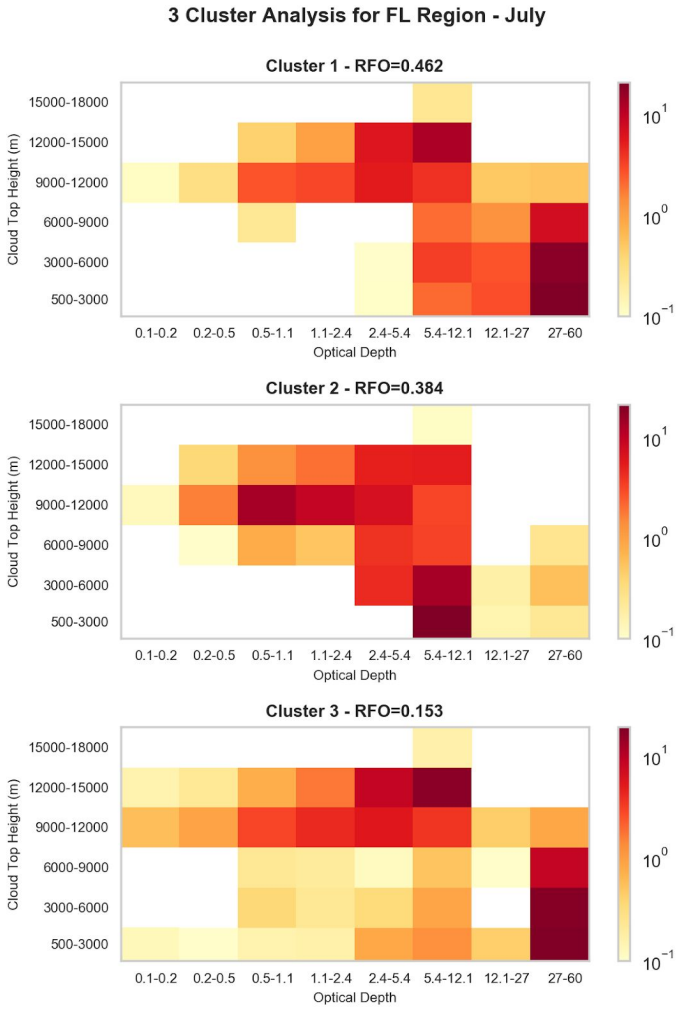

Figure 13. 3 cluster analysis for the month of July over the FL region. OD and FO color scale depicted on a logarithmic scale. The reduction to a 3 cluster analysis indicates how cluster 3, initially expected to only be found in higher cluster analysis, is one of the prevailing dominant clusters.

Reduction to a three cluster analysis, interestingly enough, showcases cluster 3, initially cluster 1 in the four cluster analysis, as a prevailing dominant cluster. In this instance, it can be inferred that the fourth cluster generated in the four cluster analysis (cluster 2) is a derivation from all prevailing clusters found in the three cluster analysis. FO distribution values for cluster 1 in the three cluster analysis highlight the focus on optically thick clouds whose OD values vary between [5.4, 60] below 9000m, compared to cluster 3's focus on the most optically thick clouds (OD>27) with greater distributions of optically thinner clouds below 9000m, as well as between [9000m, 15000m] for clouds whose OD values fall within [0.1, 0.5]. These differences, similar but slightly varied, indicate the potential to utilize a $K = 3$ value for cluster analysis rather than requiring this value to be four. Major distinct cloud fields are retained with relatively high RFO values amongst the clusters to justify their being.

July - SA Region

The mean joint OD-CTH histogram over the SA region for the month of July is shown in Figure 14 below. Major differences found in Figure 14 as compared to Figure 6 mostly pertain to the representation of optically thinner clouds (OD<12.1) with CTH values that extend from[9000m, 18000m], with the most noticeable decrease in FO value (~4%) occurring in the OD and CTH bin range of [0.1, 0.2] and [15000m, 18000m] respectively for the month of July. Overall, FO values remain highest over higher OD bin ranges (OD>5.4) below 6000m, indicating the focus on mid and low level cloud cover over the SA region.

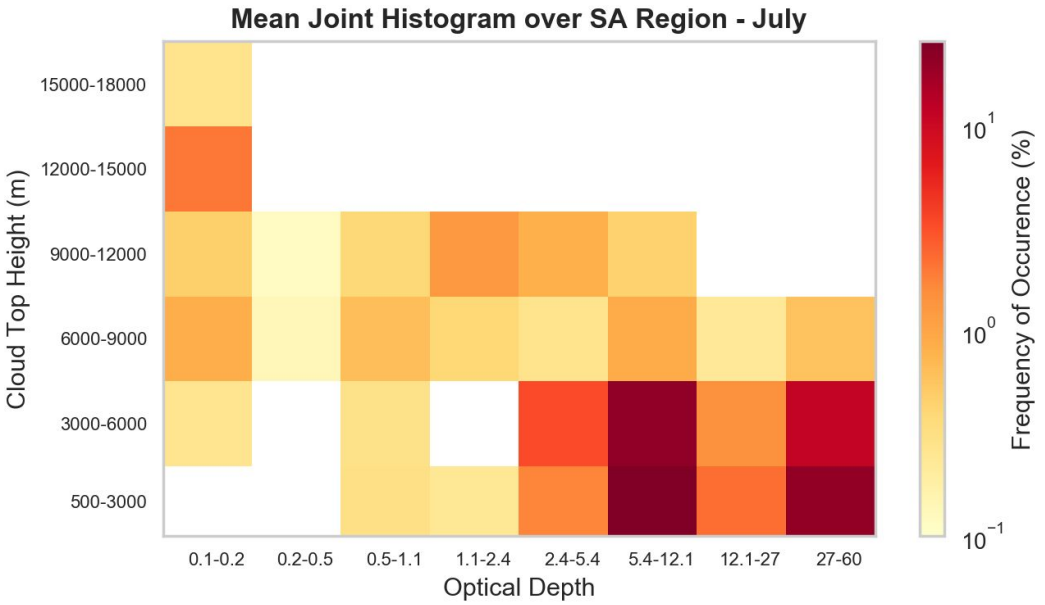

Figure 14. Mean joint histogram of OD and CTH for all hourly histograms over the SA region for the month of July, with a focus on low/mid level clouds with high optical depth values.

Silhouette scores for the month of July can be seen in Figure 15 below. Unlike previous silhouette score analysis, the SA region during the month of July exhibits an inconclusive set of score values. The maximum silhouette score, 0.158, is found where cluster number $K = 2$, and steadily decreases as K increases in value. The heuristic elbow method failed to detect (via Python's automatic application) an optimal cluster number based on silhouette score values. Choosing the optimal cluster number in this case is not as conclusive compared to other silhouette score plots, though given the relatively similar cloud field patterns between the two months over the SA region, support a rather low K value to be used

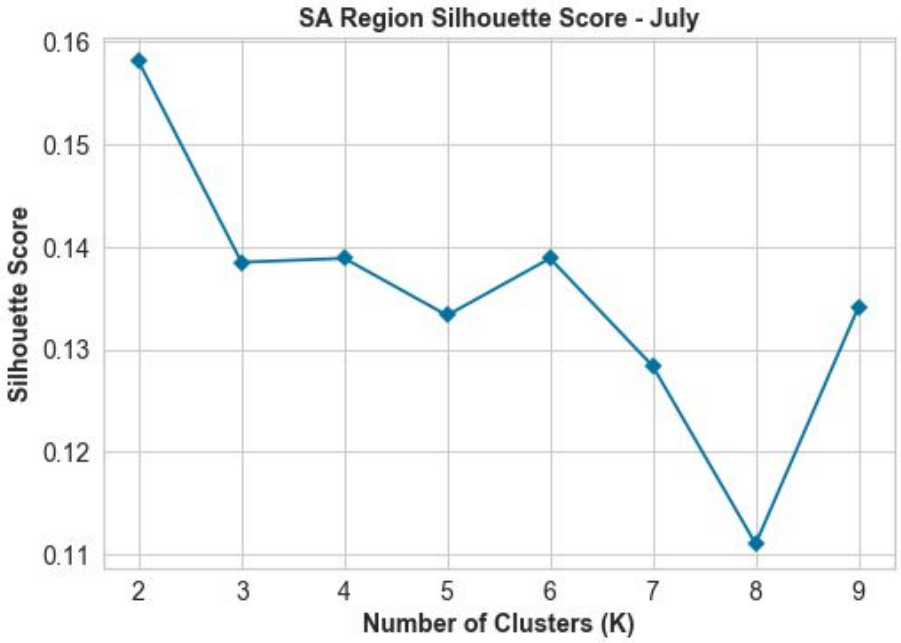

Figure 15. Silhouette score plot over the SA region for the month of July. The optimal cluster, based solely on silhouette scores, points to when $K = 2$. Note the failed application of the heuristic elbow method based on silhouette scores as no dashed line is plotted alongside score values.

Multi-cluster analysis is performed again over the SA region, first by examining the predetermined optimal cluster number at $K = 2$. The two cloud fields generated for a two cluster analysis offer simple derivations from the overall mean joint histogram as seen in Figure 14. Cluster 1, with an RFO of 0.423, focuses primarily on very optically thick mid/low level clouds with slight representation of very optically thin upper level clouds. The two bins with the highest FO values occur in the OD bin range [27, 60] and CTH bin ranges [500m, 6000m], where the combined FO values account for over 70% of values for cluster 1. Cluster 2 has the highest RFO value at 0.577, and shifts the focus of FO value distribution to thinner clouds within the same CTH bin ranges. For cluster 2, over 77% of all values fall within the OD bin range [5.4, 12.1] and CTH bin ranges

[500m, 6000m]. Both clusters, although with slightly varying quantities, represent some form of very optically thin clouds within OD bin range [0.1, 0.2] found within the mid/upper levels of the atmosphere. Indeed, the focus on low level clouds between the two clusters is primarily on the optical thickness of these clouds.

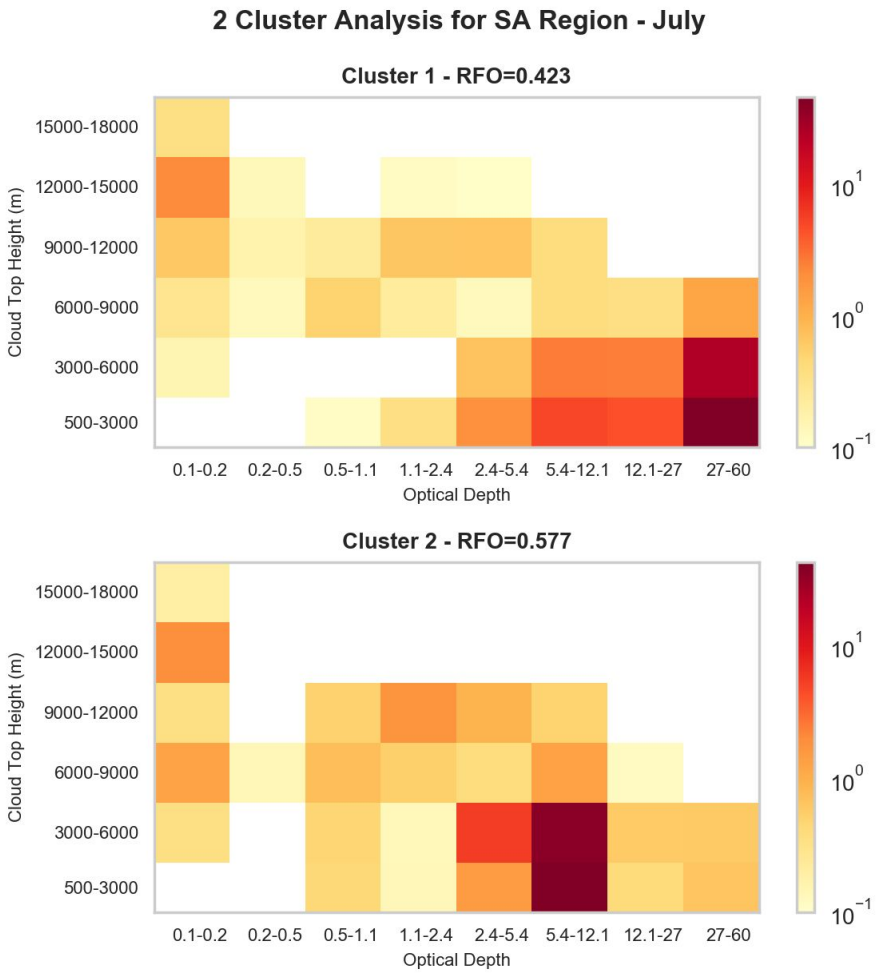

Figure 16. 2 cluster analysis over the SA region for the month of July. OD and FO color scale depicted on a logarithmic scale. The primary difference between the two clusters falls on the differing representation of mid/low level clouds with a distinct OD bin value range, where cluster 1 focuses on more optically thick clouds.

Of curious intentions is the possible application of utilizing a three cluster analysis over the SA region for the month of July. Previous analysis indicates that $K = 3$ offers a valid cluster number to select when undergoing the clustering process. The mean joint histogram for the SA regions shares similar cloud field patterns between the months of January and July, though overall the July mean is simpler in the spread of FO values; for the month of July, FO values are concentrated on two distinct OD bin value ranges compared to the marginal increase in representation of upper level clouds during the month of January. This three cluster analysis can be seen in Figure 17 below. Clusters 1 and 2 remain nearly the same in the three cluster analysis as compared to the two cluster analysis, with the most notable change in RFO value dropping from cluster 2. The tertiary cluster generated, cluster 3, can be seen as a derivative of cluster 2 given the similar distribution of FO value within the [5.4, 12.1] OD bin range. What separates cluster 2 and 3 is the shift of secondary focus on mid/upper level clouds within the OD bin ranges of [0.5, 12.1]. Cluster 3 focuses less on very optically thin upper level clouds and instead shifts this representation to clouds within the CTH bin ranges of [6000m, 12000m] in the noted OD bin ranges. For cluster 2, FO values for the noted bins only equates to ~3.7% while for cluster 3 FO values equate to roughly ~17.5%, a 13.8% difference in mid/upper level cloud representation between the two clusters.

A two cluster analysis captures the major defining cloud fields over the SA region, though a three cluster analysis retains these major cloud fields while also offering a unique variant of one of the major cloud fields. A four cluster analysis (not shown) was conducted to examine what further splintering of the clusters would offer. Results from

this brief analysis created a fourth cluster that, while possessing a relatively unique cloud field that focuses on very optically thick mid/low level clouds with moderate representation of optically thin upper level clouds, had too low of an RFO value (0.001) to justify its use.

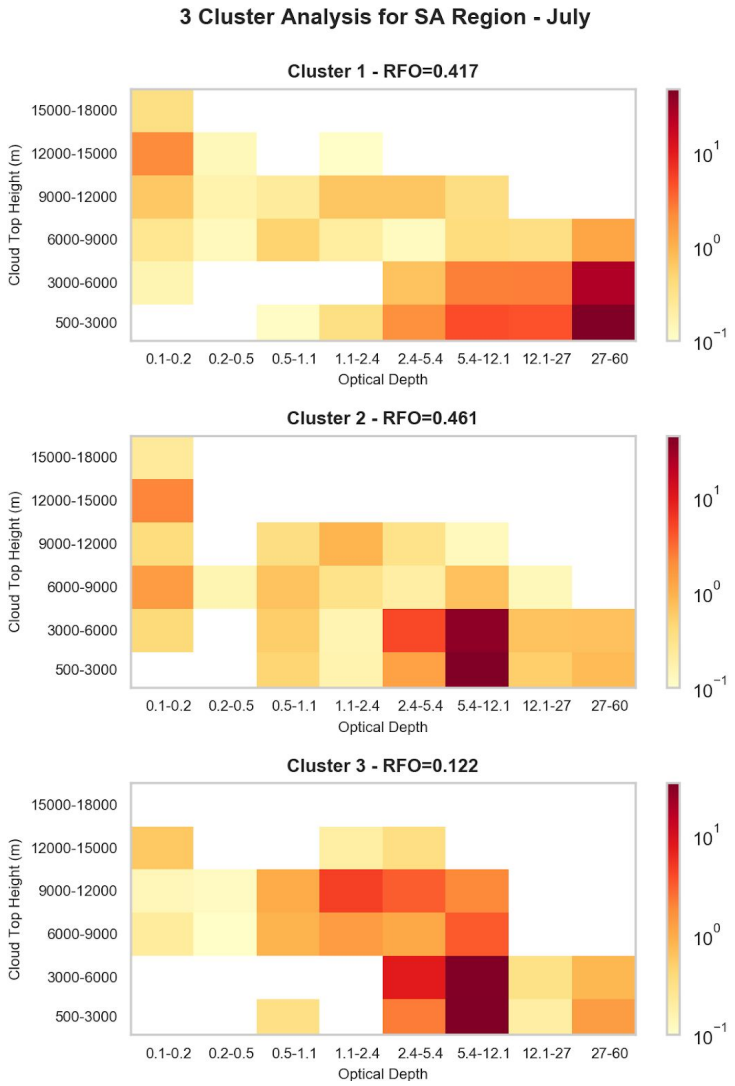

Figure 17. 3 cluster analysis over the SA region for the month of July. Clusters 2 and 3 highlight emphasis on low level clouds with a focus on significantly different OD values with cluster 1 serving as an intermediate group with a low/mid-level cloud field of variable OD values.

IV. Conclusions and Recommendations

A global satellite based cloud dataset is examined over two target regions across four years to determine probable cloud field regimes by way of k-means cluster analysis. To aid in the determination of the optimal cluster number to utilize in this cluster analysis, silhouette scores, line plots in combination with an additional heuristic form of evaluation via the "elbow method", are utilized. The optimal number of clusters derived for both the FL and SA regions varied in approach, though generally converged to when $K = 3$. Determination of this cluster number for subsequent cluster analysis was not consistent, and required further inquiry by examining and comparing a number of multi-cluster analysis. Incidentally, none of the silhouette score plots indicated that the optimal cluster number could be when $K = 3$, an interesting inquiry as both the SA and FL region had three instances of an optimal K value found both via silhouette score magnitude and the elbow method. Only the SA region for the month of July failed in its application of the elbow method, where its maximum silhouette score was found at $K = 2$. Both regions exhibited relatively similar cloud field patterns between their respective months; the FL region had a wider distribution of FO values between the low/mid/upper levels of the atmosphere, whereas the SA region had the majority of FO values concentrated on low/mid level clouds with varying OD bin value focus. While the range of OD values, being on a logarithmic scale, can vary significantly as one moves further upward in value, whether this change in cloud field representation is justified as its own unique cluster is another matter not fully explored within this study.

As noted, both regions initially indicated a differing optimal cluster number to select when running through the clustering algorithm. While some score plots offered an initial optimal cluster number, such as seen in Figure 3, further inquiry into the cloud fields generated for each cluster and their respective RFO values called into question whether a different cluster number could be utilized to both retain the dominant and unique cloud fields identified while doing away with splintered clusters that had either too similar a cloud field pattern to other clusters or too low of an RFO value to justify their being.

These initial findings further called into question the very validity of relying on silhouette score analysis to determine the optimal cluster number for k-means clustering, though a number of points must be made regarding this issue. The use of silhouette scores offers a quick and valid snapshot on the quality of data being utilized for clustering purposes. As silhouette scores, mentioned earlier, are calculated by determining the distance a datapoint has to its own cluster compared to the distance to other clusters, this may point to an issue regarding how the data was handled or the very quality of the data itself. This notion is supported by low silhouette scores calculated over both regions for both months as only the FL region had one month of maximum silhouette scores above 0.2 but less than 0.3, while the actual range of silhouette scores range from [-1,1]. While still positive, these low values indicate overlapping between certain data points and clusters. This study, in its initial stages, first examined raw OD and CTH values before shifting to viewing joint histogram FO values for bin ranges. These raw values, averaged over similar time periods, were also clustered based on silhouette score values calculated.

Most of these silhouette score values (not presented here) exceeded those found in this study with maximum scores reaching 0.8. However, these scores were calculated off of raw averaged CTH and OD values, which did not accurately represent cloud fields depicted given the significant discrepancy between hourly data points across the four years when performing these averages. Nonetheless, the actual data points being utilized significantly altered calculated score values, an attribute that may arise itself from the bin ranges used. Silhouette score values, while not usable as an end all product, did offer a base to work off regarding optimal cluster usage and was particularly effective over the FL region compared to the SA region.

Such a discrepancy may partially be attributable to the geography of the target regions and a weakness of the WWMCA data used. As noted, the FL region encompasses much more physical terrain than the SA region where the latter is composed mostly of ocean; the additional land mass, given Florida's latitudinal position and shape, allows for sea breeze interactions and associated fronts to play a significant role in the distribution of cloud fields and development of convective systems (Nicholls and others, 1990). This difference in a more land based region view compared to a more ocean based region view is a topic that warrants further inquiry given previous studies that compared surface observations and a satellite dataset over land and ocean with notable differences (Hahn and others, 2001:13). Additionally, the WWMCA dataset is noted as having difficulty distinguishing low level clouds with similar temperatures, an interesting point of contention given the WWMCA dataset generally performs better over lower latitudes in

cloud detection yet may be encumbered by the extensive marine boundary layer cloud coverage (Horseman II, 2007; Wang and others, 2004:274).

Additional work can be done regarding the validity of the cluster analysis conducted over the two target regions. While beyond the initial scope of this study, the next step in assessing cloud regime identification via cluster analysis would rely on having some type of cloud type classification system in place dependent on values of OD and CTH such as those utilized by Jakob and Tselioudis (2003:1) with their work on the ISCCP dataset. The WWMCA dataset does not readily have such a system of cloud type identification by way of these two parameters, and would require the adaptation of similar regime classification tables already in use or the creation of a new empirical set of defining parameter values. A final measure would be to validate both the identified regimes and subsequent cloud type classification by comparing with real world data, be it surface observations or other available reanalysis datasets, to determine how the accuracy and validity of RFO values of identified clusters compares to other forms of observational data. Through these efforts, it may be shown that the development of representative cloud fields over an area of interest can be created using the relatively simple clustering method implemented together with equally simple cloud parameters. Improvements to how these cloud field regimes are generated serve as a useful base for future endeavors towards the advancement of scene emulation with respect to clouds.

Appendix - Acronym List

AFIT - Air Force Institute of Technology

AFWA - Air Force Weather Agency

ASCII - American Standard Code for Information Interchange

ASSET - AFIT Sensor and Scene Emulator Tool

CBH - Cloud Base

CDFS II - Cloud Depiction and Forecast System II

CLPS - Cloud Layer Particle Size

CLWP - Cloud Layer Water Path

CT - Cloud Type

CTH - Cloud Top Height

DIRSIG - Digital Imaging and Remote Sensing Image Generation

DMSP - Defense Meteorological Satellite Program

FO - Frequency of Occurrence

ISCCP - International Satellite Cloud Climatology Project

K - Cluster number

OD - Optical Depth

RFO - Relative Frequency of Occurrence

SFCTMP - Surface Temperature

SNODEP - Snow Depth & Sea Ice

WFOV - Wide Field of View

WWMCA - World-Wide Merged Cloud Analysis

Bibliography

Bretherton, Christopher S., Taneil Uttal, Christopher W. Fairall, Sandra E. Yuter, Robert A. Walker, Darrel Baumgardner, Kimberly Comstock, Robert Wood, and Graciela B. Raga. "The Epic 2001 Stratocumulus Study." *Bull. Amer. Meteor. Soc.*, vol. 85, no. 7, 2004, pp. 967-978.

Brown, Miles, " New Cloud Depiction System Goes Online." *Air Force Weather Agency*. 25 June 2002. 557Weatherwing.Af.Mil. 30 October 2006. Retrieved on 6 November 2020.

Hahn, Carole J., William B. Rossow, and Stephen G. Warren. "ISCCP Cloud Properties Associated with Standard Cloud Types Identified in Individual Surface Observations." *J. Climate*, vol. 14, no. 1, 2001, pp. 11-28.

Horsman II, Stephen J. *An Assessment of the World Wide Merged Cloud Analysis Using Interactive Graphics*. MS thesis. Naval PostGraduate School, Monterey CA, June 2007.

Jakob, Christian. "An Improved Strategy for the Evaluation of Cloud Parameterizations in GCMS." *Bull. Amer. Meteor. Soc.*, vol. 84, no. 10, 2003, pp. 1387-1402.

Jakob, Christian, and George Tselioudis. "Objective Identification of Cloud Regimes in the Tropical Western Pacific." *Geophysical Research Letters*, vol. 30, no. 21, 2003, pp. 1-4.

Karlsson, Karl-Göran and Abhay Devasthale. "Inter-Comparison and Evaluation of the Four-Longest Satellite-Derived Cloud Climate Data Records: CLARA-A2, ESA Cloud CCI V3, ISCCP-HGM, and PATMOS-x." *Remote Sensing*, vol. 10, 2018, pp. 1-27.

King, Michael D., Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. "Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites." *IEEE,* vol. 51, no. 7, 2013, pp. 3826-3852.

Lau, Ngar-Cheung and Mark W. Crane. "A Satellite View of the Synoptic-Scale Organization of Cloud Properties in Midlatitude and Tropical Circulation Systems." *Mon. Wea. Rev.*, vol. 123, no. 7, 1995, pp.1984-2006.

MacQueen, J. “Some Methods for Classification and Analysis of Multivariate Observations.” *5th Berkeley Symposium, University of California, Los Angeles*, 1967, pp. 281-296.

Marchand, Roger. “Trends in ISCCP, MISR, and MODIS Cloud-Top-Height and Optical-Depth Histograms.” *JGR Atmospheres*, vol. 118, no. 4, 2013, pp. 1941-1949.

Marchand, Roger, Thomas Ackerman, Mike Smyth, and William B. Rossow. “A Review of Cloud Top Height and Optical Depth Histograms from MISR, ISCCP, MODIS.” *Journal of Geophysical Research*, vol. 115, 2010, pp. 1-25.

Mohamad, Ismail Bin, Dauda Usman. “Standardization and its Effects on K-Means Clustering Algorithm.” *Research Journal of Applied Sciences, Engineering, and Technology*, vol. 6, no. 17, 2013, pp. 3299-3303.

Morissette, Laurence and Sylvain Chartier. “The K-Means Clustering Technique: General Considerations and Implementation in Mathematica.” *Tut. Quant. Met. Psy.,* vol. 9, no. 1, 2013, pp. 15-24.

Nicholls, M. E., R. A. Pielke, and W. R. Cotton. *Sea Breeze: Induced Mesoscale Systems and Severe Weather*. Grant No. NAG5-359. Fort Collins CO: Colorado State University, 21 June 1990.

Pasillas, Chandra M. An Evaluation of Northern Hemisphere Merged Cloud Analyses from the United States Air Force Cloud Depiction Forecasting System II. M.S. Thesis, Naval Postgraduate School, 2013.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, and others. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

Randall, David, Marat Khairoutdinov, Akio Arakawa, and Wojciech Grabowski. “Breaking the Cloud Parameterization Deadlock.” *Bull. Amer. Meteor. Soc.*, vol. 84, no. 11, 2003, pp. 1547-1564.

Rousseeuw, Peter J. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics*, vol. 20, 1987, pp. 53-65.

Söhne, Nathalie, Jean-Pierre Chaboureau, and Francoise Guichard. "Verification of Cloud Cover Forecast with Satellite Observation over West Africa." Monthly Weather Review, vol. 136, no. 11, 2008, pp. 4421-4434.

Steward, Bryan J., AFIT Sensor & Scene Simulation Tool (ASSET) Overview, Department of Engineering Physics, Air Force Institute of Technology, Wright-Patterson AFB OH, March 2020.

Stubenrauch, C. J., W. B. Rossow, S. Kinne, S. Ackerman, G. Cesana, and others. "Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel." *Bulletin of the American Meteorological Society*, vol. 94, no. 7, 2013, pp. 1031-1049.

Tselioudis, George, Yuanchong Zhang, and William B. Rossow. "Cloud and Radiation Variations Associated with Northern Midlatitude Low and High Sea Level Pressure Regimes." *J. Climate*, vol. 13, no. 2, 2000, pp. 312-327.

Wang, Yuqing, Shang-Ping Xie, Haiming Xu, and Bu Wang. "Regional Model Simulations of Marine Boundary Layer Clouds over the Southeast Pacific off South America. Part I: Control Experiment." *Mon. Wea. Rev.*, vol. 132, no. 1, 2004, pp. 274-296.

Warren, Stephen G., Carole J. Hahn, Julius London, Robert M. Chervin, and Roy L. Jenne. Global Distribution of Total Cloud Cover and Cloud Type Amounts Over Land. Colorado University. National Center for Atmospheric Research. 1988.

Warren, Stephen G., Ryan M. Eastman, and Carole J. Hahn. "A Survey of Changes in Cloud Cover and Cloud Types over Land from Surface Observations, 1971-96." *J. Climate*, vol. 20, no. 4, 2007, pp. 717-738.

Wielicki, Bruce A., and Lindsay Parker. "On the Determination of Cloud Cover from Satellite Sensors: The Effect of Sensor Spatial Resolution." *JGR Atmospheres*, vol. 97, no. D12, 1992, pp. 12799-12823.

Xi, Baike, Xiquan Dong, Patrick Minnis, Mandana M. Khaiyer. "A 10 year Climatology of Cloud Fraction and Vertical Distribution Derived from both Surface and GOES Observations over the DOE ARM SPG site." *JGR Atmospheres*, vol. 115, no. D12, 2010, pp. 1-12.

Young, Shannon R., Bryan J. Steward, and Kevin C. Gross. “Development and Validation of the AFIT Sensor and Scene Emulator for Testing (ASSET).” *Proc. SPIE*, vol. 10178, 2017.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
25-03-2020	Master's Thesis	March 2019 – March 2020

TITLE AND SUBTITLE
IDENTIFYING FOUR YEAR AVERAGE CLOUD FIELD REGIMES FROM WORLD WIDE MERGED CLOUD ANALYSIS DATASET BY WAY OF K-MEANS CLUSTERING

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Almeida, Stewart, G., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENP-MS-21-M-098

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AGENCY: Center for Technical Intelligence Studies and Research
ADDRESS
PHONE and EMAIL
ATTN:
(no sponsor enter: Intentionally left blank)

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RHIQ (example)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

Joint histograms of cloud top height (CTH) and optical depth (OD) are created using the World Wide Merged Cloud Analysis (WWMCA) dataset over a four year period (2014-2017) to identify average cloud field regimes and assess the application of utilizing the WWMCA dataset with the AFIT Sensor and Scene Emulation Tool (ASSET). Two selected regions encompassing the Florida peninsula and a portion of the Pacific Ocean off the west-central coast of South America are examined over the months of January and July. Cloud field regimes are identified by running generated hourly OD-CTH histograms through k-means clustering, with optimal cluster number (K) evaluation performed by calculating and comparing silhouette scores and heuristic elbow method results. Varying cluster groupings are plotted out to distinguish discrepancies between these multi-cluster analysis. Initial results indicate K=3 as the optimal number of clusters to use, generating three major cloud field regimes unique to each region with high relative frequency of occurrences (RFO). Notable departures from silhouette score calculations to cluster evaluation call into question the validity of silhouette score usage to determine optimal K values, which is discussed alongside future improvements and applications of the cloud field regimes identified.

15. SUBJECT TERMS
Clouds, Cloud Regimes, Clustering, Silhouette Score, WWMCA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Saunders, Peter A, Maj, USAF, AFIT/ENP
U	U	U	UU	64	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4743 (peter.saunders@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18