Earth and Space Science

RESEARCH ARTICLE
10.1029/2022EA002373

Special Section:
Applications of Machine Learning Algorithms in Modeling Atmospheric Aerosols, Clouds, and Radiation

Key Points:

- Subclusters are created for an existing set of MODIS cloud clusters to explore the variability within the clusters
- The Davies-Bouldin index and subsom entropy were used as metrics to identify the least representative clusters
- Particular clusters are then directly analyzed using their subclusters, identifying overlooked behavior within these clusters

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:
A. J. Schuddeboom,
Alex.Schuddeboom@canterbury.ac.nz

Citation:
Schuddeboom, A. J., & McDonald, A. J. (2023). Understanding internal cluster variability through subcluster metric analysis in a geophysical context. *Earth and Space Science*, *10*, e2022EA002373. https://doi.org/10.1029/2022EA002373

Received 7 APR 2022
Accepted 15 DEC 2022

Author Contributions:
Conceptualization: A. J. Schuddeboom, A. J. McDonald
Data curation: A. J. Schuddeboom
Funding acquisition: A. J. McDonald
Investigation: A. J. Schuddeboom
Methodology: A. J. Schuddeboom, A. J. McDonald
Software: A. J. Schuddeboom

Understanding Internal Cluster Variability Through Subcluster Metric Analysis in a Geophysical Context

A. J. Schuddeboom[1] **and A. J. McDonald**[1,2]

[1]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand, [2]Gateway Antarctica, University of Canterbury, Christchurch, New Zealand

Abstract Clustering algorithms are commonly used for inspecting the behavior of clouds in both model and satellite data sets. Often overlooked in cluster analysis is the variability that occurs within any clusters generated. This is particularly important in the geophysics where clusters are often generated with a focus on interpretability over mathematical optimization. Two metrics, the Davies-Bouldin index and the subsom entropy, are used to identify clusters with large internal variability. These metrics are applied to an example set of clusters from prior research that were generated using cloud top pressure-cloud optical thickness joint histograms from the Moderate Resolution Imaging Spectroradiometer data set. Applying these metrics to the clusters identifies one cluster in particular as a major outlier. Examining the calculations behind these metrics in more detail provides further information about the internal variability of the clusters. The clusters are also examined over several geographic regions showing mostly consistent behavior. There are, however, some large anomalies such as the behavior of the clear sky cluster or the behavior of several different clusters over the Arctic Ocean. To aide our interpretation of these results, two clusters are chosen for a detailed analysis of their subclusters. The geographic distributions and radiative properties of these subclusters are examined and clearly identify that subclusters have physically distinct behavior. This result illustrates that these metrics are capable of determining when a cluster contains physically distinct subclusters. This demonstrates the potential utility of these metrics if they were applied to other geophysical data sets.

1. Introduction

Unsupervised clustering is a popular technique for investigating the behavior of large, complex data sets and has been applied to a wide range of problems (Ambroise et al., 2000; Palomo et al., 2012). It has been a particularly useful tool in climate science research due to the size of data sets, high degree of spatial and temporal auto-correlation, and focus on physical phenomena that are easily categorizable (Cassano et al., 2007; Cavazos, 2000; Coggins et al., 2014; Gibson et al., 2017; Hewitson & Crane, 2002; Jakob, 2003). However, there is a tendency to assume that any clusters generated are highly representative of their constituent data without detailed examination of the behavior within these clusters. There are several reasons why a clustering scheme could lead to unrepresentative clusters including a poor choice of clustering algorithm, requesting an inadequate number of clusters or the underlying data could be poorly suited to clustering. Additionally, past research on clustering in the geophysics has often used a variety of alternative approaches for determining optimal cluster number (Cassano et al., 2007; Gibson et al., 2017; Harrington et al., 2016; Hewitson & Crane, 2002; Kidson, 2000; Sheridan & Lee, 2011). This is because many of these past papers were making conscious trade-offs between optimizing cluster numbers based on a mathematical operation and focusing on cluster interpretability. While these trade-offs often serve the purpose of the clustering well, they can lead to an increased internal variability within the clusters. Interestingly, some of these issues were discussed with much foresight in Vesanto and Alhoniemi (2000), particularly the balancing act between interpretability and cluster separation.

It is common for a set of clusters to be reused as the basis of follow-up research (Cassano et al., 2007; Coggins et al., 2014; Kidson, 2000). This approach has the advantage of being able to generate coherent bodies of work that allow researchers to better understand the associated phenomena. However, often these clusters will have unexamined biases. As such, poor results can be propagated forward without a serious re-examination of the underlying data. A key motivating factor for this research is determining a way to identify these biases without discarding the useful results from the earlier research.

Supervision: A. J. McDonald
Validation: A. J. Schuddeboom
Visualization: A. J. Schuddeboom
Writing – original draft: A. J. Schuddeboom, A. J. McDonald
Writing – review & editing: A. J. Schuddeboom, A. J. McDonald

Understanding the behavior of clusters becomes more complex when multiple data sets are simultaneously examined such as in climate model evaluation (Mason et al., 2015; Williams & Tselioudis, 2007). Working on multiple data sets is often an issue as applying clusters established using one data set to a second data set means that the resulting analysis can be impacted by biases in either data set and therefore without understanding the underlying data it becomes difficult to reach definitive conclusions. Similarly complex attribution issues arise if instead of using one set of clusters as a basis for examining other data, a combined set of clusters is generated using multiple data sets as in Mason et al. (2015). Additionally, changes in the underlying data set due to nonstationary phenomena such as the effects of climate change will also impact the interpretation. Detailed examination of the internal variability of the clusters can also provide valuable information to aide in the interpretation of the clusters (Cho et al., 2021).

For this research, we choose to focus on clusters that were generated using the self-organizing map (SOM) unsupervised learning algorithm. This algorithm has been used in a wide range of applications (Allinson & Ellis, 1992; Auger et al., 1992; Campelo et al., 2014), but is most commonly used as a clustering algorithm (Kohonen, 2013). Details about the design and operation of the SOM are given in Kohonen (1998) and Kohonen (2013). This paper focuses on internal cluster variability and primarily explores this variability through the use of subclustering or subsomming (also known as hierarchical somming/clustering). Subclustering is the process in which the clustering algorithm is reapplied to data that has already been clustered, resulting in the identification of representative subclusters within each of the original clusters. There are several variants of the SOM algorithm that make use of hierarchical clustering when generating clusters such as the ASSOM (Kohonen et al., 1997), GHSOM (Dittenbach et al., 2002), RSOM (Zhang & Yu, 2006), and TreeSOM (Samsonova et al., 2006). This approach has also been previously used to determine the optimal number of cloud clusters (Oreopoulos et al., 2016). However, rarely have these techniques been used generate subclusters on an existing cluster for the purpose of cluster evaluation which is what is done in this paper. Earlier examples of subclustering briefly being used in this capacity include McDonald et al. (2016) which creates subnodes to examine subtle variability with nodes and Jin et al. (2020) which used subregimes to split up an overly popular cluster. One major advantage, this approach has over defining new clusters is that it allows past research to be easily expanded upon.

The specific clusters used in this research are the cloud clusters developed using the SOM approach in Schuddeboom et al. (2018). These clusters were generated using cloud top pressure-cloud optical thickness (CTP-COT) joint histograms from the Moderate Resolution Imaging Spectroradiometer (MODIS) data set. This is based on a framework established by previous research including the SOM-based clustering in McDonald et al. (2016) as well as earlier research based on k-means clustering (Jakob, 2003; Leinonen et al., 2016; Oreopoulos et al., 2014). This approach for defining cloud clusters has been particularly useful for evaluating the quality of model cloud representation, as it allows for a cloud type-based evaluation of model cloud properties (Mason et al., 2015; Williams & Tselioudis, 2007; Williams & Webb, 2009). By focusing on specific cloud types, individual errors in the model associated with a cloud type can be identified. This is particularly valuable for model evaluation as it circumvents the issue of compensating errors, which occur when one error in a model is canceled out by another. These compensating errors are often very hard to identify but cloud clusters can provide unique insight of these errors. For example, Schuddeboom et al. (2019) uses cloud clusters to estimate the magnitude of compensating errors between different cloud types in the shortwave (SW) cloud radiative effect (CRE) for a set of simulations. This was refined in Schuddeboom and McDonald (2021) which extended the same approach to evaluating a range of CMIP6 models over the Southern Ocean region.

Using subclusters generated by the SOM algorithm allows for a deeper investigation of the behavior that occurs within these clusters. By investigating the constituent members of a cluster, it is possible to improve the understanding of the phenomena that drive the behavior of the cluster. Past research has explored these relationships implicitly through many techniques including dendrograms. The relationship between clusters and their constituents can be quantified explicitly in several different ways using subclusters. In this paper, the behavior of subclusters is investigated through the usage of two different metrics, the Davies-Bouldin (DB) index and the subsom entropy. The DB index (Davies & Bouldin, 1979) is a well-established metric that quantifies how distinct similar clusters are, while the subsom entropy is a metric defined in this paper that describes how evenly distributed the occurrence rates of the subclusters are. Other metrics for cluster evaluation besides the DB index are considered briefly, but their results are shown to be similar. The DB index and subsom entropy let us identify anomalous clusters for which the subclusters can then be examined in detail. While the approach that is developed in this

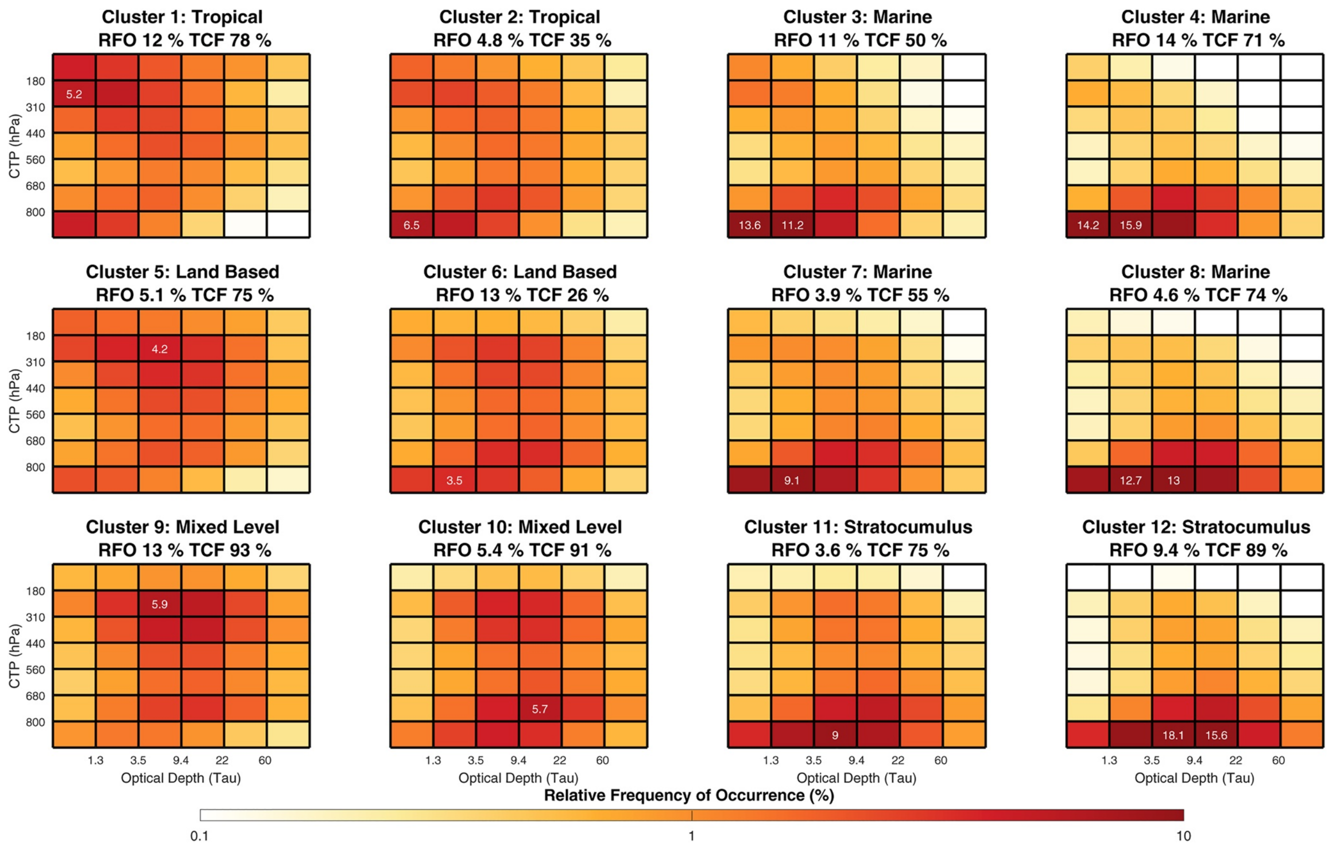

Figure 1. The Moderate Resolution Imaging Spectroradiometer (MODIS) cloud top pressure (CTP)-cloud optical thickness (COT) histogram clusters developed in Schuddeboom et al. (2018). The numbers in the subtitles of each cluster represent the relative frequency of occurrence (RFO) and the mean total cloud fraction (TCF) of the members of the cluster. When a given grid cell exceeds the limits of the color bar, it is displayed with a number over the grid cell that states the magnitude. Additionally, if none of the cells exceed the limits of the color bar, the highest occurrence cell is labeled with its magnitude.

paper are used with the SOM clustering scheme, they can be applied to subclusters generated by any clustering algorithm.

2. Data and Methods

2.1. Data

The cloud clusters developed in Schuddeboom et al. (2018) are used for the analysis in this paper. These clusters were generated by applying the SOM algorithm to CTP-COT joint histograms from the MODIS data set. MODIS is an instrument aboard the Terra and Aqua satellites that provides 1° × 1° resolution measurements of various cloud properties by passively sensing radiances at different wavelengths (King et al., 2003; Platnick et al., 2003). In particular, we use data from the collection six data set (Platnick et al., 2017) covering the year 2007. As 2007 is a strong La Niña year that led to another strong La Niña year, it could be suspected that this would bias our results. However, the interannual variability was low enough that any bias would be small. While 1 year is a very limited time period, our past work has only found minor variations in the clusters when examined on a decadal time scale. Additionally, the volume of data over a single year is still very large with over 30 million samples per year. This presents some issues with accurate estimation of uncertainties that could potentially be addressed with a subsampling strategy; however, we consider this out of scope in the present study. There are also several well-known biases in the MODIS data set including handling of partially cloudy (PCL) pixels, limited sampling and broken cloud but we do not think that these will bias our analysis.

The clusters generated from the MODIS data are shown in Figure 1. A full description of the generation of these clusters is given in Schuddeboom et al. (2018). Note that the histograms presented in this figure are the representative nodes of the SOM and not the mean histogram of all of the cluster constituents. As such, they are likely

10.1029/2022EA002373

to appear less similar to their constituents because of the neighborhood effects of the SOM algorithm. The cloud fraction values that are used in this paper are also taken from the MODIS data set.

One aspect which could impact the interpretation of these clusters is the close relationship between several of the clusters. This will be discussed where relevant throughout the manuscript but to aide the reader unfamiliar with our previous papers, we have also included a figure which shows the Pearson correlation coefficient between each of the cluster histograms in Figure S1 in Supporting Information S1. Caution should be used when interpreting these values as they are simply the correlation coefficient between the representative histograms. These values are intended just a brief summary of these relationships, a more comprehensive examination would look at geographic distributions and relationship to physical properties as in Schuddeboom et al. (2018). Figure S1 in Supporting Information S1 shows many very strong correlations between the histograms with the strongest relationships seen between neighboring nodes, highlighting many of the relationships established in Schuddeboom et al. (2018). The fact that the histograms generally consist of a few dominant cells and then generally small values in the other cells leads potentially to inflated correlation values.

We choose to use the clusters from Schuddeboom et al. (2018) because they are representative of clustering approaches used in the geophysics, they clearly describe the underlying data and because these clusters have established links to physical descriptions of clouds. It could be argued that cluster accuracy would be improved with additional preprocessing such as removing outliers based on silhouette values. However, as stated above, the clusters from Schuddeboom et al. (2018) were developed in a manner consistent with other clustering research in the geophysics. As the purpose of this paper is to demonstrate the methodology used, the exact set of clusters used should not strongly impact the results but ensuring that the clusters used are similar to those used in other research is important. Given the motivations for this research, using a data set based on complex real-world data instead of any simplified test data set will also be important due to differences in variability within clusters.

Each cluster from Schuddeboom et al. (2018) was split into six subclusters. These subclusters are generated by applying the SOM algorithm to the constituents of the predefined clusters. Several different configurations and numbers of subclusters were tried. Ultimately six subclusters appeared to capture a wide range of behavior in the MODIS data while keeping each of the subclusters relatively distinct from one another. A plot which shows the sensitivity of the DB index and SSE values to subcluster number is included in Figure S2 in Supporting Information S1. It shows that while there is variability due to the cluster number, the variations appear to be relatively minor particularly in normalized SSE where the majority of clusters show differences between the smallest and largest values of <10%. This suggests that the choice of cluster number will impact our results but the impacts will be relatively minor. It also suggests that an approach with a different number of subclusters for each cluster might be more suitable; however, we opt to keep things simple by using a single fixed number. A three by two grid is used for our subclustering, which ensures two modes of variability can be covered by the subclusters. Additionally, later analysis requires the MODIS data to be subset into different regions. The regions analyzed in this paper are the same as those used in Schuddeboom et al. (2018), which are in turn based on those used in Leinonen et al. (2016) and are also shown in a figure that is included as Figure S3 in Supporting Information S1.

In addition to cloud data taken from the MODIS data set, radiative fluxes from the Clouds and the Earth's Radiant Energy System (CERES) synoptic 1° (SYN1deg) product are also used (Doelling et al., 2013; Smith et al., 2011; Wielicki et al., 1996). CERES is a three channel radiometer that is carried aboard several different satellites including both the Aqua and Terra satellites which hold the MODIS instruments. The CERES instruments passively records radiance measurements which are then processed to form higher level data sets, such as SYN1deg. In particular, this paper uses SW and longwave (LW) top of atmosphere fluxes for clear sky and overcast conditions to generate SW and LW CRE values following the process described in Schuddeboom et al. (2018).

2.2. Methodology

In many applications of unsupervised learning algorithms, it can be difficult to interpret the behavior of the resulting clusters. When the underlying data set is suitable, this issue can be circumvented by using a small number of clusters. However, with most data sets, a large number of clusters is needed to capture the variability in the system. This results in many researchers making a trade-off between the mathematical quality of larger cluster numbers and the interpretability of smaller cluster numbers. In many cases, the large number of clusters makes subjective evaluation impossible, so metric-based evaluation is required. This work focuses on exploring two such metrics that can be used to understand the variability within clusters. First is the DB index which is well established in prior cluster research. This index is commonly used to determine the optimal number of clusters for a given data set and is used in this work to represent how distinct the subclusters are from each other. The second metric used is the subsom entropy which is a new metric defined in this paper based on the well-established concept of entropy. The subsom entropy analyzes the different occurrence rates of the subclusters which is often overlooked by established cluster variability metrics. This can have a large impact on the interpretation of clusters due to the complementary role that occurrence rates can play in aiding the interpretation provided by other forms of analysis like the DB and Calinski-Harabasz (CH) indices.

The DB index was first defined in Davies and Bouldin (1979) to determine the optimal number of clusters for a given data set. The basis of this paper was that by generating several sets of clusters, each with a different number of clusters, and calculating the Davies-Bouldin index, the set of clusters that minimizes this index would be the most representative of the underlying data. This index is used in a fundamentally different capacity in this study, instead of calculating the index once for the full set of clusters it is calculated for every cluster using its subclusters. Other metrics that provide a similar function to the DB index were also briefly examined with their results included in Section 5 and Supporting Information S1. These include the Dunn index, Silhouetting, and the CH index (Calinski & Harabasz, 1974; Halkidi et al., 2002a, 2002b; Rousseeuw, 1987; Saini et al., 2019). By calculating the DB index for each cluster, the internal variability within the cluster is quantified. The index is calculated using the following series of equations:

$$\mathrm{DB} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{1}$$

$$R_i = max_{j\neq i}\frac{S_i + S_j}{M_{ij}} \tag{2}$$

where DB is the Davies-Bouldin index, R_i is an intermediary variable used in the calculation and i and j are iterating variables that correspond to the subclusters. N is the number of subclusters associated with the given cluster, S_i is a measure of the dispersion of subcluster i and M_{ij} is a measure of distance between the subclusters i and j. Specifically, we use the MATLAB package SOM Toolbox to calculate the Davies-Bouldin index which uses the following equations for S_i:

$$S_i = \left\{\frac{1}{T_i}\sum_{j=1}^{T_i} |X_{j,i} - Y_i|^q\right\}^{\frac{1}{q}} \tag{3}$$

$$M_{ij} = \left\{\sum_{k=1}^{L} |Y_{i,k} - Y_{j,k}|^p\right\}^{\frac{1}{p}} \tag{4}$$

where T_i is the number of constituent vectors allocated to subcluster i, $X_{j,i}$ is the jth constituent vector of the cluster i and Y_i is the mean of all of the constituent vectors of subcluster i. L is the number of elements in each of the vectors and the variable k in Equation 4 is used to iterate over each element within the vector $Y_{i,k}$. p and q define the distance metric used. In this work, we use the SOM Toolbox default values of $p = q = 2$ which corresponds to the Euclidean distance. We did experiment with other value of p and q; however, the impacts were minimal on our data set.

Summarizing, S_i describes how variable the data aggregated into a subcluster is while M_{ij} describes how distinct the subcluster centroid is from the other subcluster centroids. Therefore, R_i is a ratio of the variability within the subclusters to the distinctness of its subclusters. This means that a large R_i value indicates that the subclusters within the cluster are unrepresentative of the underlying data either due to the subclusters containing highly variable data or being too similar to each other. This is why minimizing the R_i value can be used to determine the optimal number of clusters. As only the maximum value pair is used to calculate R_i, each subcluster is compared only to its most similar subcluster. In addition to the DB index, which is primarily used as a measure of subcluster distinctness, S and M values are also directly used to evaluate the behavior of the clusters later in this paper.

While the DB index summarizes the relationships between the subclusters of a given cluster, it has a clear limitation in that it does not consider the occurrence rates of the subclusters. This means the DB index can be unrepresentative as very rare clusters can ultimately control the results. Therefore, we introduce the subsom entropy which quantifies how even the occurrence rates are across different subclusters. Entropy is often used as a statistical measure of randomness in a system which consists of several microstates (Shannon, 2001) and has also been used for many different purposes in several studies in the atmospheric sciences (Bannon, 2015; Krützmann et al., 2008; McDonald & Cairns, 2020) and in past clustering-based studies (Campelo et al., 2014; De Mántaras, 1991; Halkidi et al., 2002b). By understanding that the occurrence rate of subclusters are equivalent to the probability of the occurrence of a given microstate, entropy can be used as a measure of the evenness of the distribution of subcluster occurrence rates. As such, the subsom entropy allows us to incorporate occurrence rate into our analysis. As suggested above, this aides in the interpretation of many other metrics such as the DB index and can provide some additional insight into the cluster variability. Specifically, we define the subsom entropy by the equation

$$\mathrm{SSE}_j = -\sum_i \mathrm{RFO}_i \ln\left(\mathrm{RFO}_i\right) \tag{5}$$

where SSE_j is the subsom entropy of cluster j and RFO_i is the occurrence rate of subcluster i given that the corresponding cluster j occurs. This equation has the same form as the standard statistical mechanics and information theory equations for entropy. The RFO values in this equation must be scaled to a fraction between 0 and 1 (technically not including 0 where the natural log is undefined). The simplest interpretation of the subsom entropy is that the smaller the value the more unevenly distributed the subcluster RFO is. Therefore, a small subsom entropy indicates that a given cluster is dominated by one or two subclusters.

While both of these metrics are useful for analyzing clusters individually, as discussed earlier they are more powerful when used in conjunction. For example, if a cluster has a low DB index and a low SSE, it must have relatively representative subclusters with uneven occurrence rates. This corresponds to a cluster with several physically meaningfully distinct states with one or two of the states being more common than the others. On the other hand, a cluster with a high DB index and a low SSE would have unrepresentative subclusters dominated by a few subclusters. This could mean that there is very little variance in the constituents of the cluster or potentially indicates some underlying issue with the data being clustered. Alternatively, a cluster with a high DB index and high SSE value would contain unrepresentative subclusters with no single subcluster dominating. This describes a situation where the cluster is unlikely to be a physically coherent cluster. These clusters would be the strongest targets for further analysis that is focused on understanding data that is poorly captured by the set of clusters. Finally, a low DB and high SSE cluster would be highly representative of the underlying data.

3. Results

The clusters developed in Schuddeboom et al. (2018) are examined by the creation of subclusters for each of these clusters. The DB index and SSE values were calculated for each of the MODIS cloud clusters using these subclusters. The DB value is calculated for each cluster using the calculation in Equations 1–4 and the SSE is calculated using the subcluster RFOs. These values are shown in Figure 2 for each of the clusters. The majority of clusters are grouped relatively tightly with the exception of cluster 6 which has the largest DB and smallest SSE values. Our previous work labeled cluster 6 as the clear sky cluster and identified it as a significant problem for model simulation (Schuddeboom et al., 2018). Besides cluster 6, the next largest outlier is cluster 2. This is mostly because cluster 2 shows the second largest DB value which suggests that its subclusters are relatively unrepresentative. A potential argument could be made for interpreting these clusters in smaller groupings; however, the differences between these groupings would be very small and mostly based on DB index alone. Alternative versions of this plot using the Dunn index, CH index, and Silhouette values are included in Supporting Information S1. These show quite different results highlighting that choice of metric will impact the results.

To investigate what is driving the variation in the DB values, Figure 3 shows the mean, minimum, and maximum S_i, M_{ij}, and R_i as defined in Equations 2–4. Examining the S values, it is clear that three of the clusters (clusters 1, 5, and 9) stand out with large mean and maximum values. However, these clusters all correspond to small to average DB index values due to their large M values. Additionally, clusters 2 and 6, which have the largest DB values, show smaller than average S values, but the smallest M values. Clusters 10, 11, and 12 all show similar

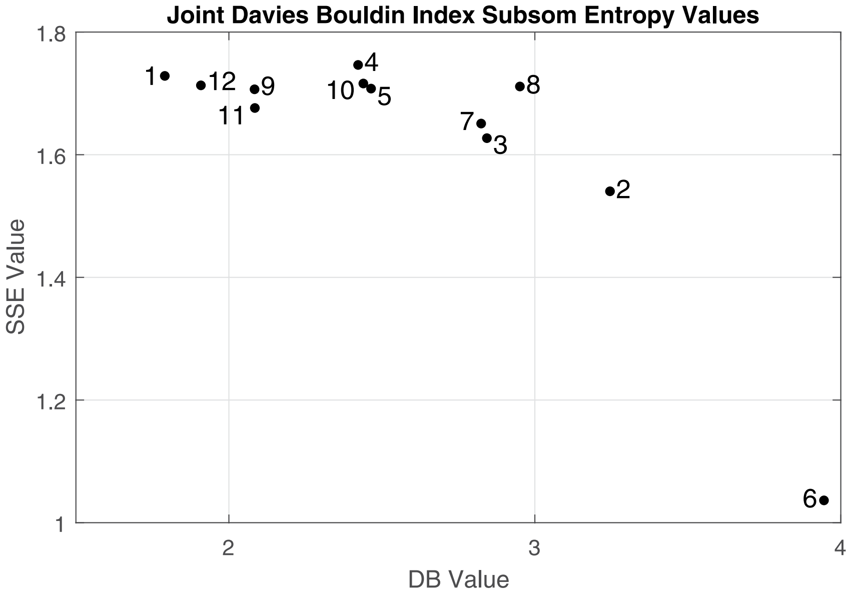

Figure 2. The global Davies-Bouldin index and subsom entropy values for each of the MODIS-derived cloud clusters. Note the restricted scale used in the axis to better highlight the differences between the clusters.

patterns to clusters 1, 5, and 9 with above average S values and large M values. Due to this combination of S and M values, clusters 10, 11, and 12 have some of the lowest DB values. Clusters 3, 4, 7, and 8 all show extremely small S and smaller than average M values. This results in above average DB values, meaning less representative subclusters. Also clear from Figure 3 is that the M values generally show a larger variability than the S values and therefore have a greater role in determining a clusters DB index.

Examination of the ranges of values in Figure 3 can also provide insight into the subclusters. For instance, clusters 2 and 6 both show a large range of S values and relatively tightly packed M values. This means that the corresponding subclusters are similar to one another, but have a wide range of behavior in their constituent histograms. Clusters 1, 5, and 9 show interesting behavior in their ranges, with large ranges in both S and M but not in R. This must be due to the subclusters with large S values always having a corresponding large M value to keep the ratio for R similar. While the R values generally show small ranges, clusters 3, 6, and 8 have a large range of R values. For clusters 3 and 8, this is likely due to the limited variation in S while in cluster 6 it is due to relatively tightly constrained M values.

Also included in Figure 3 is the distribution of Pearson correlation coefficient values calculated between each cluster and its constituent histograms. This approach has been used previously to examine internal cluster variability (Gibson et al., 2017; McDonald & Parsons, 2018) fulfilling a similar role to the S value. The results show that while there is some agreement on how representative a cluster is between the correlation coefficient and the DB index values; in general, they are quite different. For example, both the DB index and correlation coefficient results show that cluster 6 is potentially unrepresentative of its constituents; however, the correlation coefficient identifies cluster 5 as unrepresentative while the DB index does not. This could be partly due to the correlation coefficient calculation examining the entire cluster, while the S value is calculated by examining each of the subclusters independently. As such, cluster 5 results could be explained by having distinct subclusters that are considerably more representative of their constituents than the cluster. There does appear to be some relationship between a clusters maximum S value and correlation coefficients, with the largest S values showing the lowest correlation coefficient values. This is possibly due to high S values requiring at least one subcluster where the constituents are significantly different from their representative histogram which would naturally imply a smaller correlation coefficient.

The results of Figure 3 can be used to more clearly interpret the results from Figure 2. Clusters 1, 5, and 9 have large S values which indicate that when their subclusters are examined they have a relatively wide range of constituent histograms. Additionally, the M values suggest that each of these subclusters are distinct from one

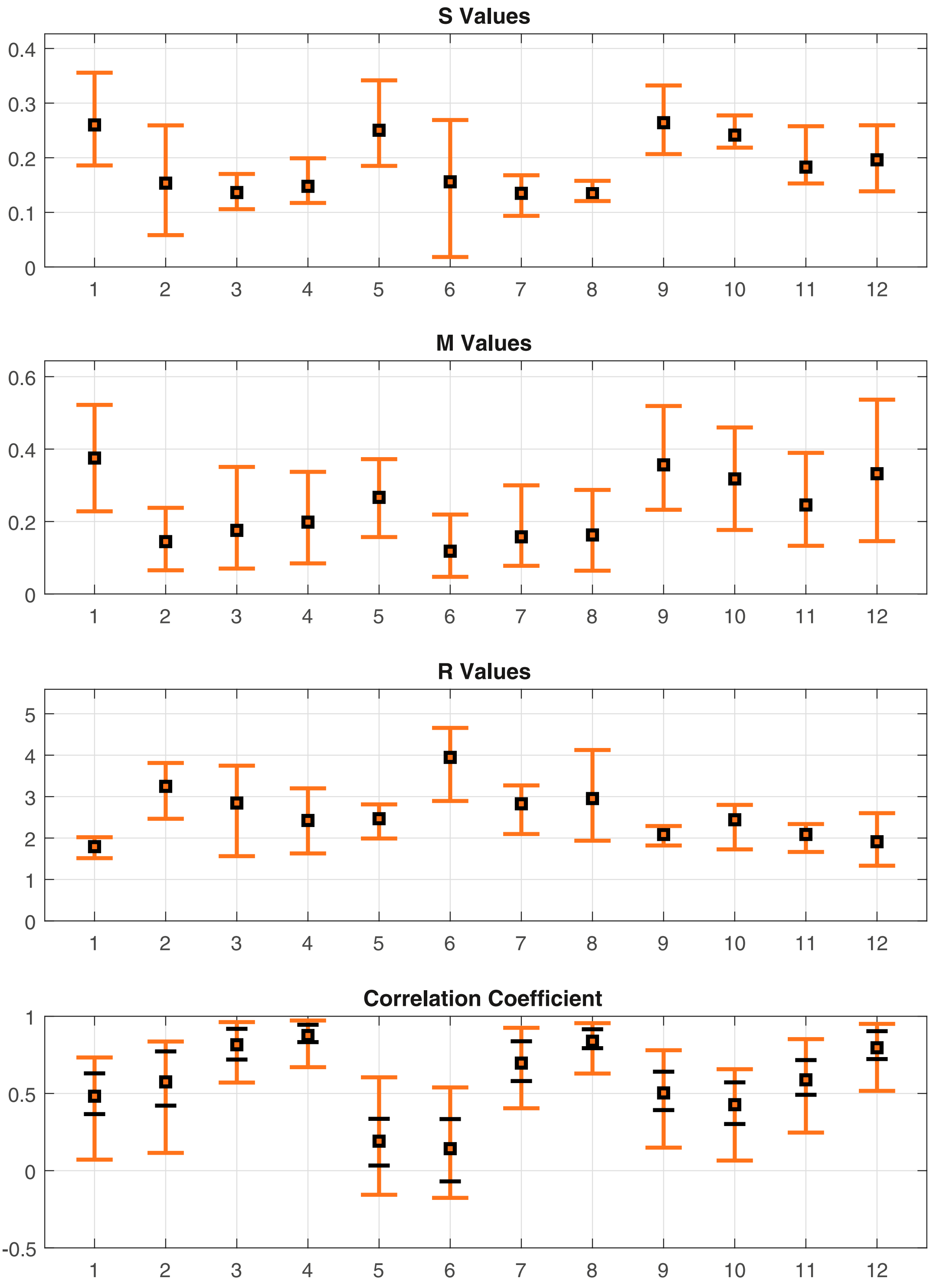

Figure 3. The breakdown of the global Davies-Bouldin index into S_i, $M_{i,j}$, and R_i for each of the different clusters. Also included is the Pearson correlation coefficient between each cluster and its constituent histograms. The black box represents the mean values for the cluster. For the first three subplots, the orange bars show the minimum and maximum values. For the correlation coefficient plot, the inner black bars indicate the 25th and 75th percentiles and the outer orange bars indicate the 5th and 95th percentiles.

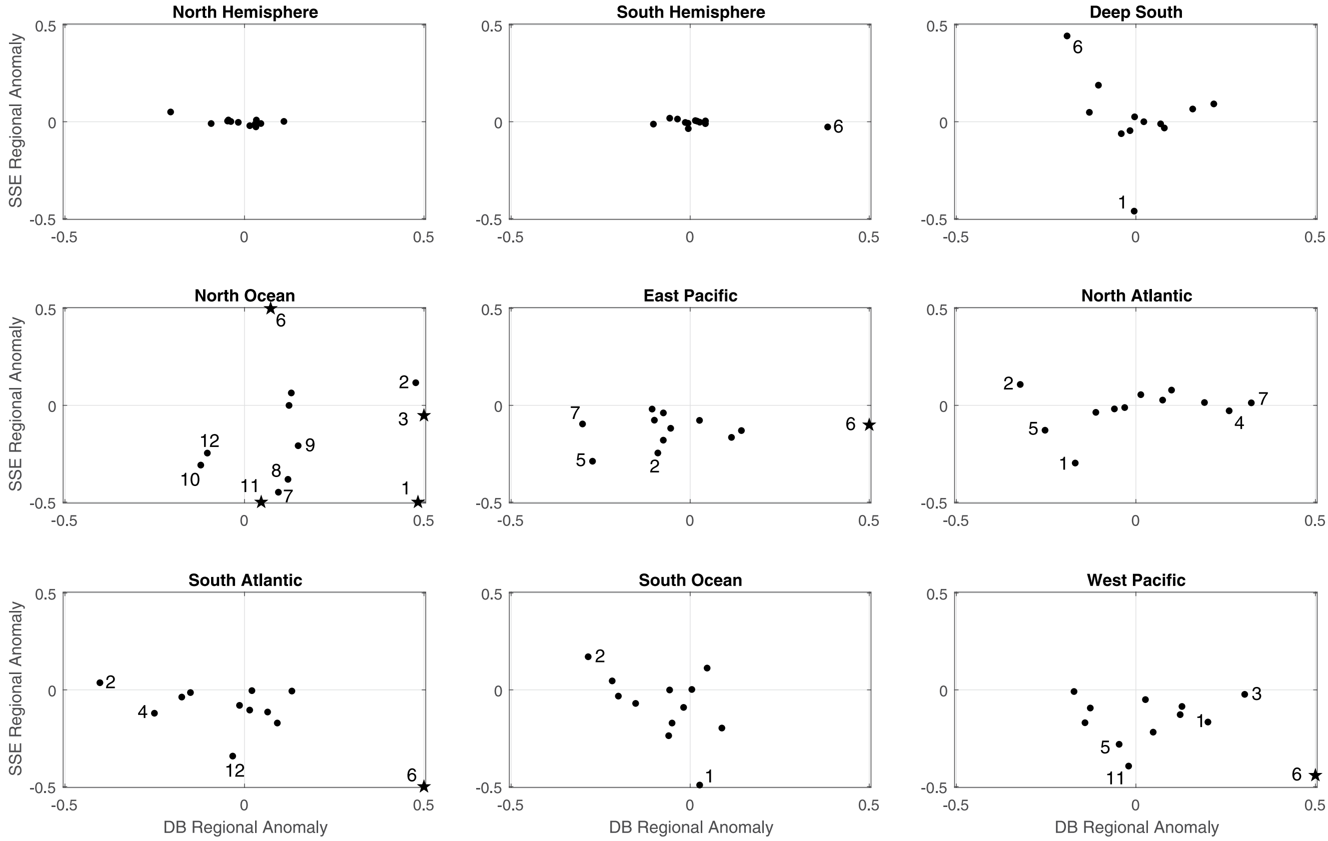

Figure 4. The region-specific Davies-Bouldin index and subsom entropy values for each cluster. These values are presented as anomalies from the global mean values. Any clusters that are a distance greater than 0.25 from the origin are labeled with their cluster numbers. Additionally, star symbols are plotted on the closest point when a cluster exceeds the bounds of the figure. The regions plotted are the regions identified Schuddeboom et al. (2018).

another. Combining these shows that clusters 1, 5, and 9 have distinct subclusters, but these still fail to capture all of the behaviors in the constituent histograms. Similar conclusions can be reached about cluster 10, 11, and 12, but due to their lower *S* values, it appears these clusters have subclusters that are slightly better at representing their constituents. On the other hand, clusters 2 and 6 show a wide range of constituent histograms within each subcluster, but considerably less distinct subclusters. When the low SSE values of these clusters, shown in Figure 2, are also considered, it suggests that these clusters are represented by a dominant subcluster which may be impacting the *S* values because all of the other subclusters are given equal weighting. The remaining clusters all show low *S* and *M* values.

The information presented in Figure 2 describes global behavior of these clusters, but may also mask large regional variability. To investigate the regional variation in the behavior of the clusters, the values for the DB index and the SSE are plotted for different regions of the globe in Figure 4. These plots are shown as anomalies from the global values in Figure 2 to better illustrate regional variation. A version of this plot using the regional values instead of the anomalies is included in Supporting Information S1. The nine regions examined in this figure are those defined in Schuddeboom et al. (2018) and a corresponding figure showing these regions is included in Supporting Information S1. In general, the region results are similar to the global results shown in Figure 2. There are, however, some regions such as the North Ocean and the North Atlantic which have many clusters that stand out as anomalies. There also appear to be some clusters that are more likely to be outliers like clusters 1, 11, and 12.

The region which is the largest outlier in Figure 4 is the North Ocean region. This region shows SSE and DB values that contrast strongly with the global results. These differences are larger in the SSE than the DB index. This region shows major deviations in many clusters including clusters 2 and 6 which do not have the same anomalous subsom entropy that they do globally. Clusters 1 and 11 have the smallest SSE values, showing smaller

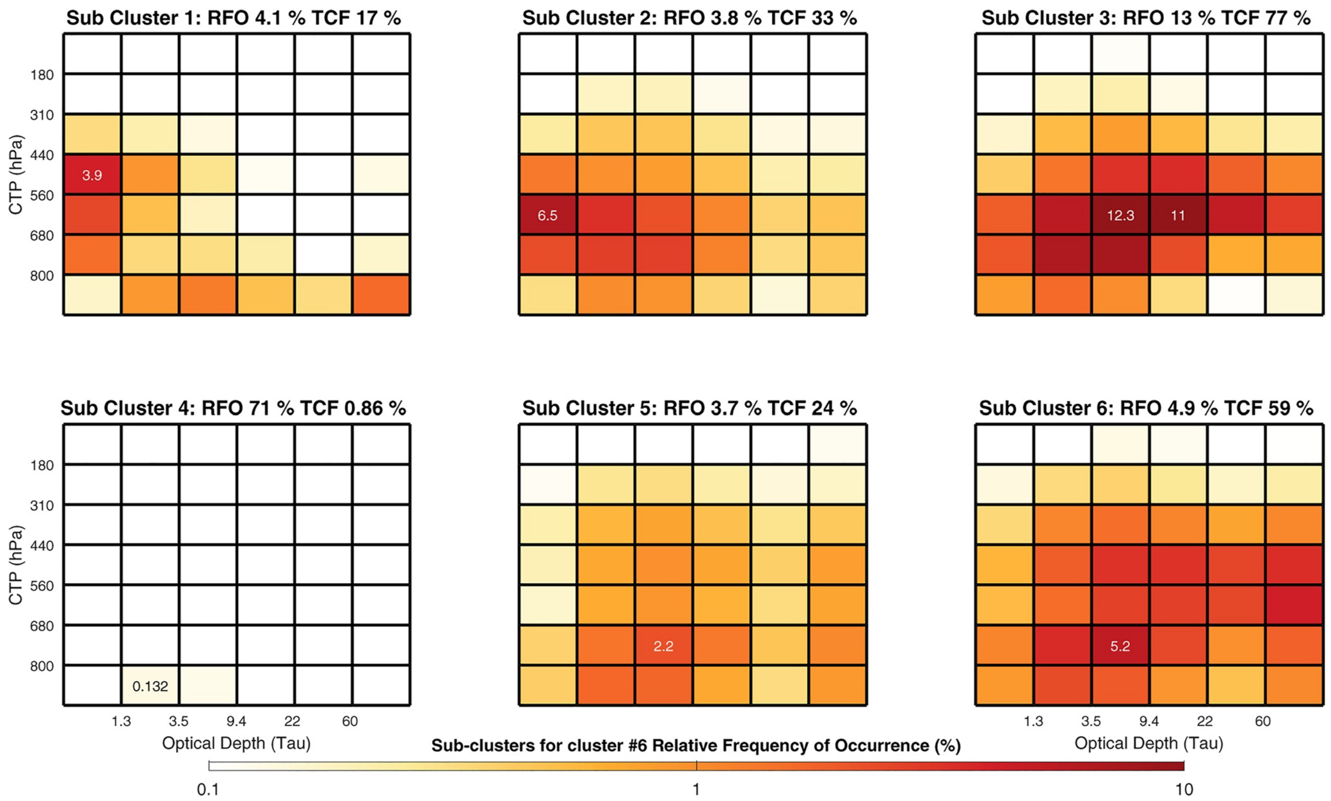

Figure 5. The mean cloud top pressure-cloud optical thickness (CTP-COT) histograms for each of the subclusters associated with cluster 6. The numbers in the subtitles of each cluster represent the relative frequency of occurrence (RFO) and the mean total cloud fraction (TCF) of the members of the cluster. When a given grid cell exceeds the limits of the color bar, it is displayed with a number over the grid cell that states the magnitude. Additionally, if none of the cells exceed the limits of the color bar, the highest occurrence cell is labeled with its magnitude.

values than the global values for cluster 6. The differences shown in this region could be at least partially due to the North Ocean region being the smallest region examined. This region also has several other clusters which display notable behavior, such as clusters 1, 2, and 3 showing abnormally large DB index values.

4. Case Studies

Cluster 6 has been identified as the strongest outlier in Figures 2–4. Our past research established this cluster as associated with relatively clear skies concentrated over deserts. To better understand this cluster, a more detailed examination of its subclusters is undertaken. First, the subcluster histograms associated with cluster 6 are shown in Figure 5. As the mean histograms of the subclusters are used in the metric analysis, they are shown here instead of the nodes of the SOM. The results shown earlier in Figure 1 identify cluster 6 as having low cloud fraction and a very diffuse overall structure. It is immediately clear that all of the subcluster histograms differ significantly from the representative cluster. The cloud types represented by these subclusters also differ considerably with the most common and second most common clusters showing an almost 80% difference in cloud fraction. The subclusters show a clear ordered structure along both axes, which is expected due to the SOM process. This is visible along the rows in the form of an increase in cloud fraction and a shift to higher optical thicknesses as we move from left to right. Comparing the rows shows that the subclusters in the bottom row have a similar structure to those in the top row but with lower cloud fractions. As expected from the low subsom entropy value of the cluster, the RFO of the different subclusters is very unevenly distributed, with the vast majority of the constituents associated with the lowest cloud fraction subcluster (subcluster 4).

Detailed examination of subcluster histograms in Figure 5 allows us to better understand the cluster 6 results shown in Figure 3. For instance, the average *M* value for cluster 6 is the smallest of all the clusters. This means that the average Euclidean distance between the subclusters is less than any other cluster. While the clusters

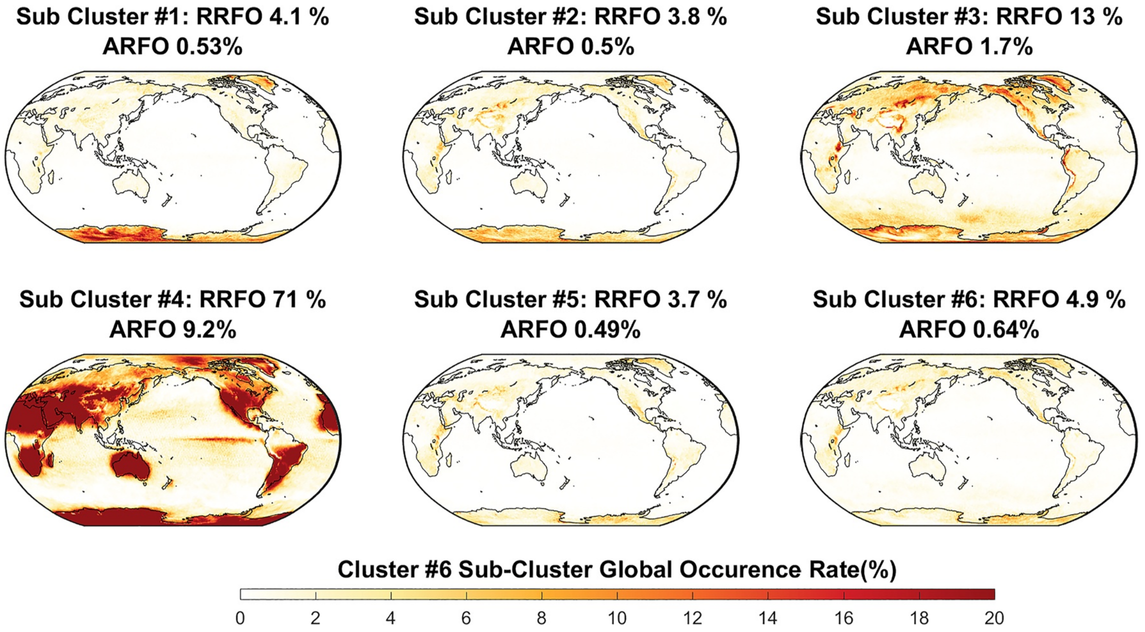
Sub Cluster #1: RRFO 4.1 %
ARFO 0.53%
Sub Cluster #2: RRFO 3.8 %
ARFO 0.5%
Sub Cluster #3: RRFO 13 %
ARFO 1.7%
Sub Cluster #4: RRFO 71 %
ARFO 9.2%
Sub Cluster #5: RRFO 3.7 %
ARFO 0.49%
Sub Cluster #6: RRFO 4.9 %
ARFO 0.64%
Cluster #6 Sub-Cluster Global Occurence Rate(%)
0
2
4
6
8
10
12
14
16
18
20

Figure 6. The geographic distributions of the occurrence rate for each of the subclusters associated with cluster 6. The numbers in the subtitles of each cluster represent the regional relative frequency of occurrence (RRFO) and the absolute relative frequency of occurrence (ARFO) of the subclusters.

in Figure 5 appear distinct from each other it is important to appreciate that the Euclidean distances between subclusters 1, 2, 4, and 5 will all be small due to the low cloud fraction values. As an aside, the approach taken in Doan et al. (2021), which uses structural similarity to allocate members to clusters, shows potential as a possible remedy for this issue. Having a low cloud fraction will reduce the M value because a cluster with many high cloud fraction subclusters can show slight structural differences in the histogram which will result in relatively large Euclidean distances. The S values show a very large range of values including at least one very low and high S valued subcluster. Examination of the individual subcluster S values shows the lowest valued subcluster is subcluster 4 which is unsurprising given the low cloud fraction and likely abundance of similar clear sky cases. The remaining subclusters all show much larger S values with subcluster 3 displaying the largest value. Given that subcluster 4 has a much higher occurrence rate than the other subclusters, an argument could be made that their should be some form of cluster weighting for S value calculation. This does show the value of using the subsom entropy as just using the DB index would not include any occurrence rate information.

The geographic distributions of the occurrence rate for each of the subclusters of cluster 6 are shown in Figure 6. The regional and absolute occurrence rates are identified, where the regional value is defined as the occurrence rate of a subcluster given that the corresponding cluster occurs, while the absolute values are the occurrence rate given that a valid measurement is made. This means that the absolute occurrence rate is the same as the earlier definition of RFO but is defined separately here for clarity. As identified in Schuddeboom et al. (2018) cluster 6 is confined almost entirely to land and is particularly prevalent over deserts. Subcluster 4 dominates the cluster and matches well with the established geographic distribution of cluster 6. Given the extremely low cloud fraction values associated with subcluster 4, it appears that the constituents of cluster 6 often have lower cloud fractions than the representative cluster. The other notable subcluster in Figure 6 is subcluster 3. This subcluster primarily occurs in the midlatitudes in both hemispheres. As these regions are characterized by persistent and dense cloud cover, it is unsurprising that the cloud fraction of subcluster 3 is considerably higher. There is also a substantial presence of this cluster over mountainous regions which could possibly correspond to erroneous retrievals. Given this difference it is surprising that this cloud would be grouped in cluster 6. This is likely a result of the constituents of this subcluster not being well captured by any of the other cluster and therefore being assigned to cluster 6 out of a lack of suitable alternative.

To further analyze the behavior of the subclusters of cluster 6, we examine the cloud properties of each of the subclusters. Figure 7 shows the distribution of SW CRE, LW CRE, and the cloud fraction for each subcluster. Also included is the Pearson correlation coefficient between the representative subcluster and its constituent

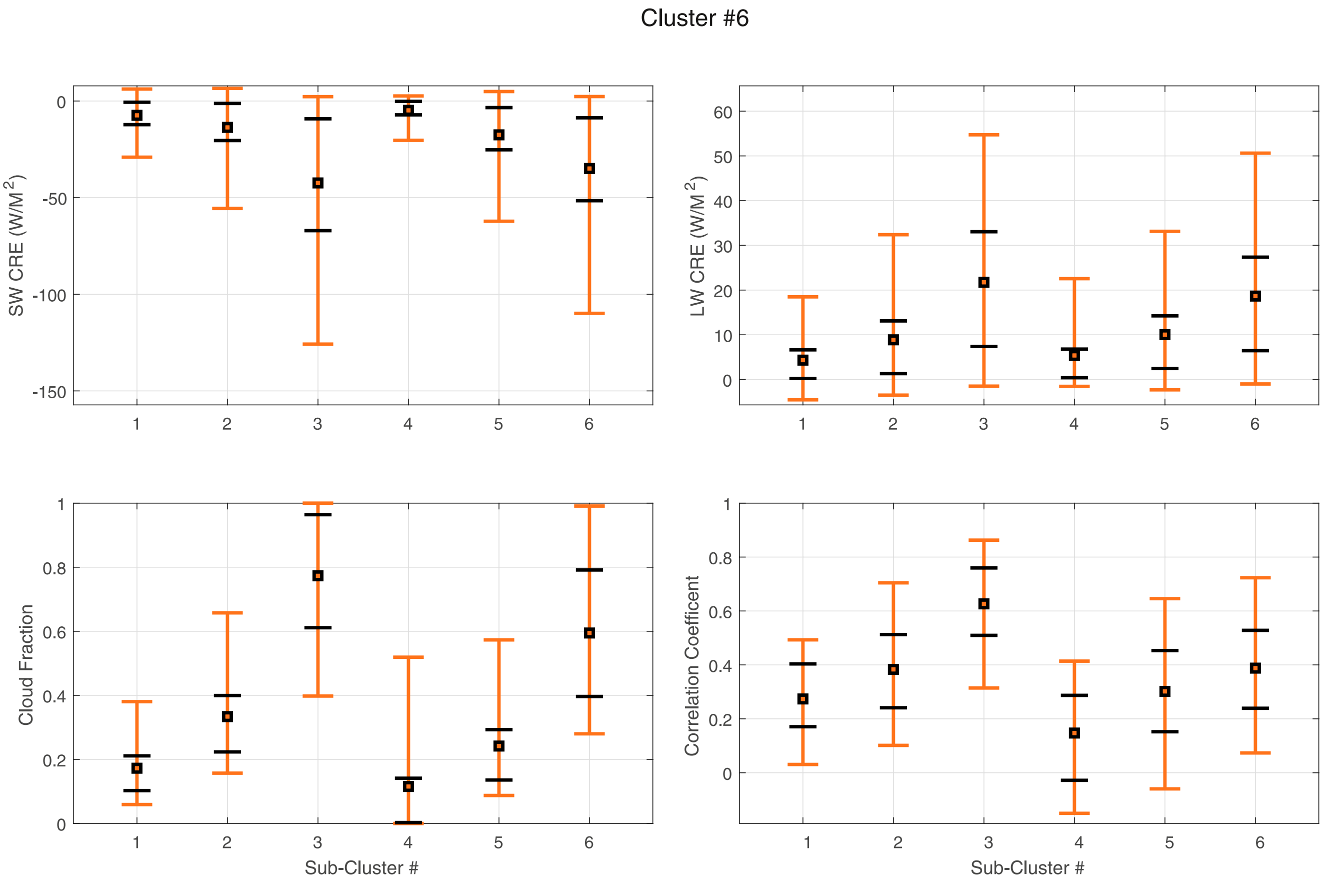

Figure 7. The distributions of SW CRE, LW CRE, and cloud fraction of the subclusters from cluster 6. The Pearson correlation coefficient between the representative subcluster and the constituent histograms is also included. The mean value is indicated by the black box. The inner black bars mark the 25th and 75th percentiles, while the outer orange bars mark the 5th and 95th percentiles.

histograms. Immediately apparent is that the subclusters can be grouped into three distinct groups, with each corresponding to a pair of subclusters along a column in Figure 5. First, looking at the SW and LW CRE, the subclusters 1 and 4 pair show a small mean with a small range, the subclusters 2 and 5 pair show slightly larger mean and range and the subclusters 3 and 6 pair show a much larger mean and range. These same patterns are also observed for the cloud fraction and correlation coefficient values. The SW CRE, LW CRE, and cloud fraction values support the earlier identification of subcluster 4 as mostly clear skies and subcluster 3 as an optically thicker midlatitude cloud. Over a quarter of the subcluster 4 measurements are associated with zero or near zero cloud fraction and over three quarters fall well under a value of 0.2. Clearly the results presented in Figure 7 suggest that subclusters 3 and 4 represent fundamentally different types of clouds. The wide differences in ranges covered by these distributions suggest a larger difference between these subclusters than would be implied from the mean values alone.

The Pearson correlation coefficient data in Figure 7 are more complicated to interpret. These values show the range of correlation coefficients between the representative histogram and its constituent members. Examination of the two dominant subclusters (subclusters 3 and 4) shows drastically different behavior. Subcluster 3 displays strong agreement between the representative cluster and its constituents with a mean correlation coefficient greater than 0.6 and even the 5th percentile is close to 0.3. Subcluster 4 shows completely different results with a mean value below 0.2 and over a quarter of measurements below 0. However, these values will be biased due to the low cloud fraction values of the subcluster. This occurs because when the cloud fraction of a histogram is lower, a smaller change to the histogram has a much larger effect on the correlation values. The other variables in Figure 7 show that even though the subcluster 4 constituents do differ from the representative histogram they are still physically coherent.

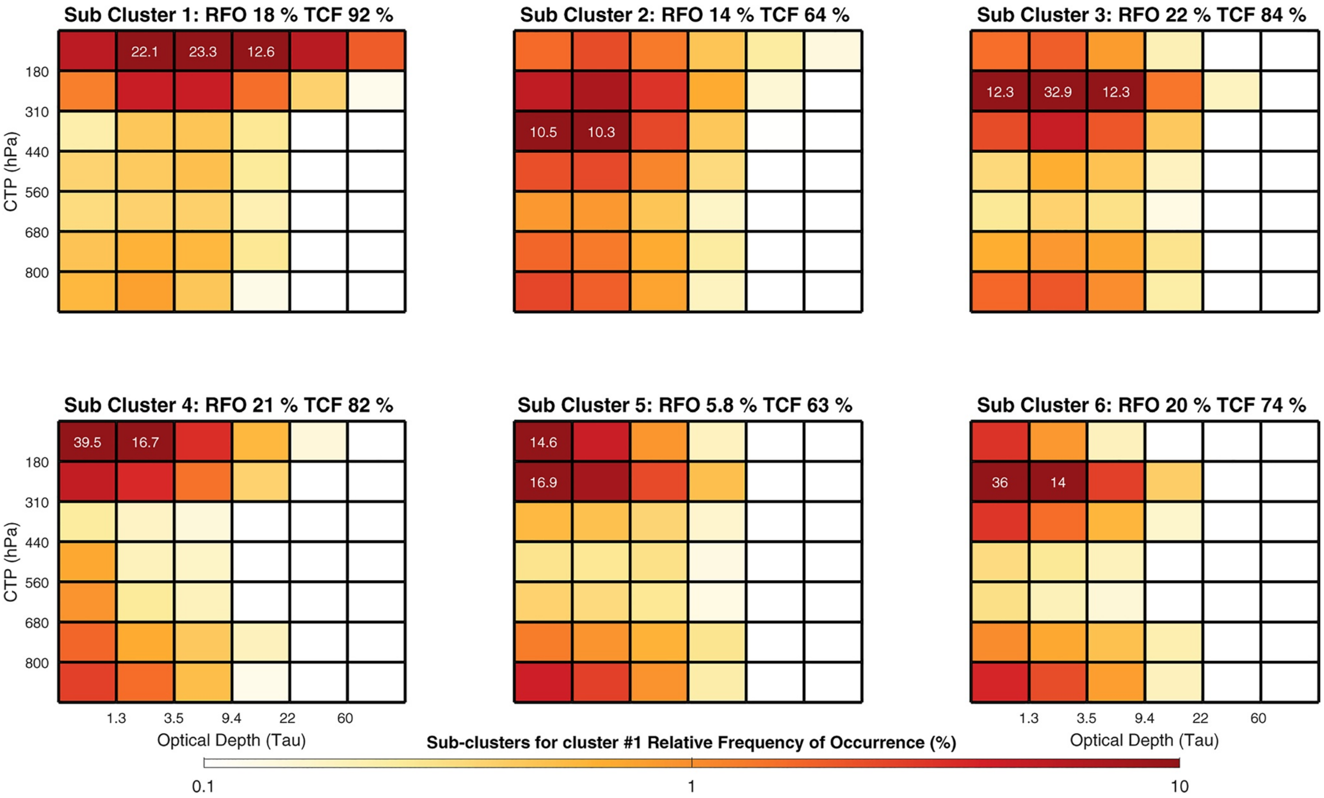

Figure 8. The mean cloud top pressure-cloud optical thickness (CTP-COT) histograms for each of the subclusters associated with cluster 1. The numbers in the subtitles of each cluster represent the relative frequency of occurrence (RFO) and the mean total cloud fraction (TCF) of the members of the cluster. When a given grid cell exceeds the limits of the color bar, it is displayed with a number over the grid cell that states the magnitude. Additionally, if none of the cells exceed the limits of the color bar, the highest occurrence cell is labeled with its magnitude.

As a point of comparison to cluster 6, we also examine the subclusters of cluster 1 which shows the smallest DB values and amongst the largest subsom entropy values. This cluster is associated with high altitude generally optically thin cloud concentrated over the Tropics, particularly over South East Asia. The joint histograms for each of the subclusters within cluster 1 are shown in Figure 8. All of these histograms consist of optically thin high-altitude cloud above optically thin lower level cloud. The subclusters appear to slightly differ in altitude and optical thickness while the form of the histogram remains largely unchanged. It is also worth noting that the cloud fraction values shown in subclusters 2 and 5 are smaller than the other subclusters. Compared to the equivalent cluster 6 histograms shown in Figure 5, the differences between the structure of the subcluster histograms in Figure 8 appear small. However, recall that the M values shown in Figure 3 were larger for cluster 1 than cluster 6. This may seem like a contradiction but is in fact due to increases in cloud fraction causing the Euclidean distances between subclusters to increase with small structural differences. This does not bias the DB index results as it should impact S and M values equally. The occurrence rates of the subclusters are also very similar which is expected given the large subsom entropy.

The analysis of cluster 6 in Figure 7 is recreated for the subclusters of cluster 1 as Figure 9. This figure shows that the subclusters of cluster 1 have very similar physical properties to one another with subcluster 1 being the only subcluster that is notably distinct. These results are in stark contrast to the cluster 6 results. The cluster 1 subclusters show much more overlap in SW CRE, LW CRE, and cloud fraction distributions and have a much stronger correlation coefficients between the subcluster histograms and their constituents. This highlights the effectiveness of the subsom entropy and Davies-Bouldin index at describing the relationship between a given cluster and its subclusters. Interestingly, the ranges shown for all of the variables in Figure 9 are larger than the ranges in Figure 7. This suggests that the subclusters of cluster 1 are more restrictive in the behavior shown in other variables than the subclusters of cluster 6. This suggests that in a very particular way cluster 1 shows less physical cohesion than cluster 6.

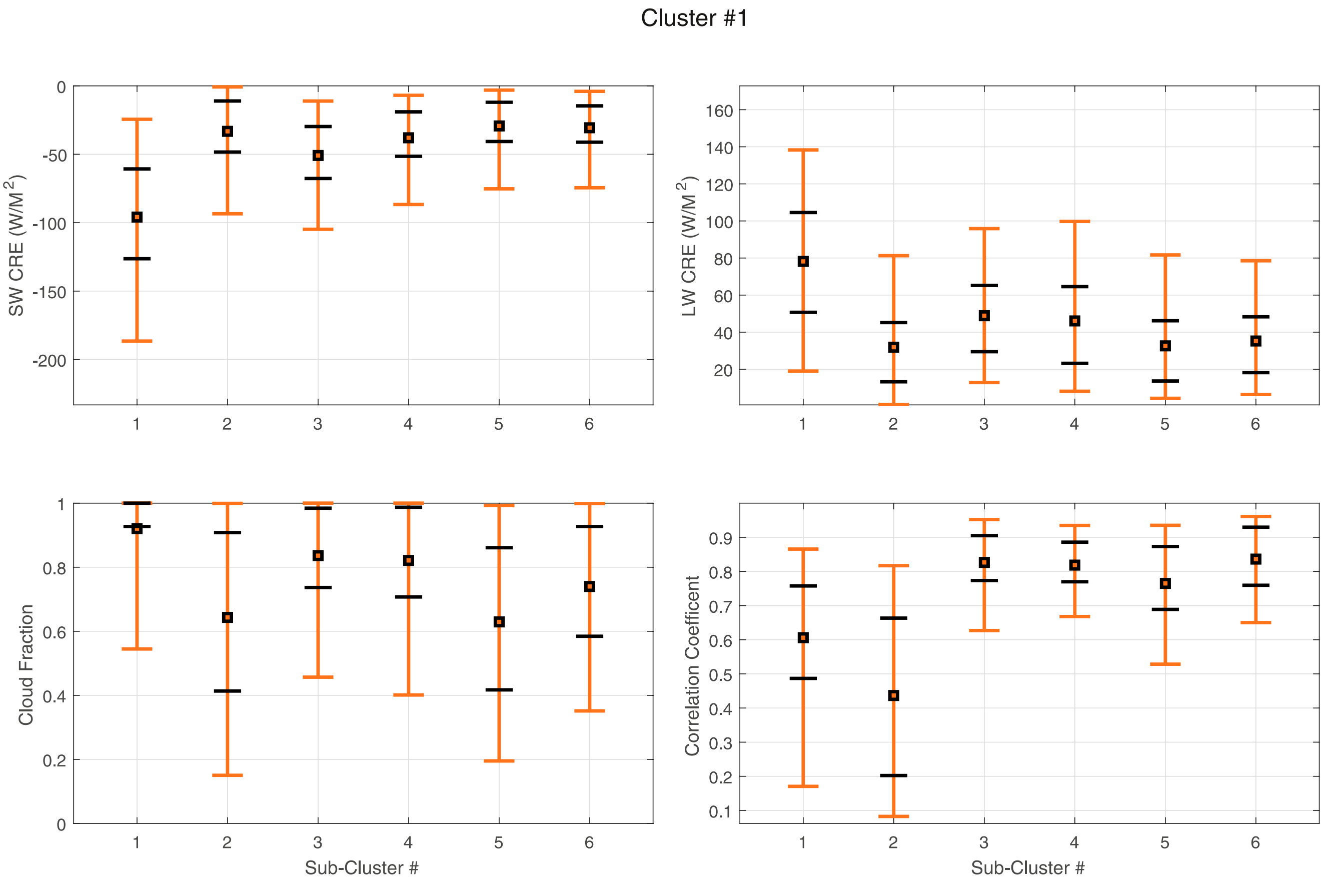

Figure 9. The distributions of shortwave (SW) cloud radiative effect (CRE), longwave (LW) CRE, and cloud fraction of the subclusters from cluster 1. The Pearson correlation coefficient between the representative subcluster and the constituent histograms is also included. The mean value is indicated by the black box. The inner black bars mark the 25th and 75th percentiles, while the outer orange bars mark the 5th and 95th percentiles.

5. Discussion

While the majority of the results presented above are simple to interpret, some of these figures could be masking more complex relationships. For example, the process of interpreting the cluster-specific correlation coefficient values in Figure 3 is complicated by a limited understanding of how the correlation coefficient relates to the other variables. To examine how these values relate to the *S*, *M*, and *R* variables, scatter plots of these values are included in Supporting Information S1. Also included in each of these figures is the Spearman rank-order correlation coefficient between *S*, *M*, and *R*. These figures show no clear relationships between these variables except for possibly the cluster correlation and *S*. The Spearman values for the relationship between ranks of the cluster correlation values and *S* values indicate a relationship between the ranks of these variables. The spearman values are not an ideal approach considering that small perturbations to some of these values could lead to changes in rank. However, even if these changes in rank would occur they would not be sufficient to produce evidence of a relationship between these variables.

Additionally, the apparent relationship between the *S* and *M* values is also investigated. A scatter plot between these variables is also included in Supporting Information S1. This relationship shows a Spearman rank-order correlation coefficient of $\rho = 0.78$ and a relatively small *p*-value of $p = 0.004$ suggesting a direct relationship in the ranks of the two variables is plausible. This relationship appears to be more impacted by the limitations of the spearman coefficient with many values closely grouped and therefore ranks potentially interchangeable. Although, visually the evidence for a (likely nonlinear) relationship appears strong.

The clusters from Schuddeboom et al. (2018) were used extensively in this paper, but Schuddeboom et al. (2018) also groups these clusters into subjective types which are unexamined here. Interestingly there appears to be a relationship between the cloud types and the *S*, *M*, and DB values in Figure 3. The marine (clusters 3, 4, 7, and

8), mixed-level (clusters 9 and 10), and stratocumulus (clusters 11 and 12) cloud types all show coherent behavior between all of their member clusters. However, the Tropical (clusters 1 and 2) and Land-based (clusters 5 and 6) cloud types display clear differences between their associated clusters. This is somewhat expected as these two types show larger differences in their histograms and were identified in Schuddeboom et al. (2018) as less physically coherent than the other types. Interestingly, the correlation coefficient values show greater agreement with this typing than the *S*, *M*, and DB values. While it is hard to draw definitive conclusions about this relationship, it does show that the ability for these clusters to represent their constituents is directly related to the physical characteristics of the clouds.

Due to their relative complexity, it is probable that the clusters with the greatest internal variability would also be the most challenging for models to simulate. Some insight into this relationship can be gained by comparing the results of this paper to the previous research in Schuddeboom et al. (2018) which identifies the quality of cluster simulation in the atmosphere only component of the HadGEM3 climate model (Hewitt et al., 2011). This previous work identifies clusters 4, 6, 9, 10, and 12 as particularly difficult for the model to capture. By examining how these clusters are shown in this work, we can get some insight into if there is a relationship between model biases and cluster variability. Cluster 6 has been clearly identified by both the DB and SSE as an outlier cluster. Clusters 9, 10, and 12 all show normal SSE and DB values, but clearly have anomalously large *S* and *M* values. Cluster 4, however, is not identified by any of our analysis as a notable cluster. Overall, this suggests that there may be some relationship between our metrics and the quality of model simulation, but the strength of this relationship is unclear. Further investigation is required to properly quantify how these values relate to errors in model simulations.

One potential criticism of this research is the exclusive use of the DB index to characterize the variability of the clusters. To explore how dependent these results are on this metric, we also examine the behavior of three other metrics; the Dunn index, CH index, and the mean Silhouette values (Calinski & Harabasz, 1974; Dunn, 1973; Halkidi et al., 2002a; Saini et al., 2019). Unfortunately, it is not feasible to properly calculate these metric values given the large sample size of our data set, so instead these metrics are estimated by randomly sampling 5,000 points from each of the subclusters. To ensure the comparison is fair between clusters two further modifications are applied to these metrics. First, the DB index is inverted so that larger values indicate less variability as in the other metrics. Second, the CH index and Silhouette values have to be normalized to account for differences in occurrence rates between the subclusters to ensure that calculation is not biased due to the sampling process. Without this normalization, the subsampling process effectively assumes that each of the subclusters is equally likely to occur. The impact of this normalization is large with the differences in silhouette value shown in Figure S10 in Supporting Information S1.

The results from each of the different metrics is plotted in Figure 10. A cursory examination shows similarities between the DB index and CH index, but radically different behaviors in the results from the Dunn index and Silhouette values. The Dunn index appears to be a major outlier as it is often determined by the largest outliers within the subcluster as opposed to the other metrics which are impacted by all of the members of the subcluster. The DB and CH indices show good general agreement on which clusters have large values and which have small values. The results from the silhouette coefficient analysis appears to show relatively good clustering for cluster 2 and flat values for all other clusters. Further interpretation of these values is complex especially the whole distribution of silhouette values is normally considered not just the mean. Detailed examination not included here shows that the value for cluster 6 is due to very high silhouette values associated with subcluster 4. In general, the computational costs suggest that the silhouette value limits our ability to compare results with the DB index but the CH index could possibly be used.

One additional issue that become apparent during the course of this research is the dependence of these results upon the usage of Euclidean distance. However, the Euclidean distance is slightly inappropriate for this analysis due to the inability to account for spatial structures in the joint histograms. This impacts the metrics used to evaluate the clusters as well as the clusters themselves (as the SOM is defined using Euclidean distance). One approach that could resolve this issue is the S-SOM developed in Doan et al. (2021) which uses structural similarity in place of Euclidean distance.

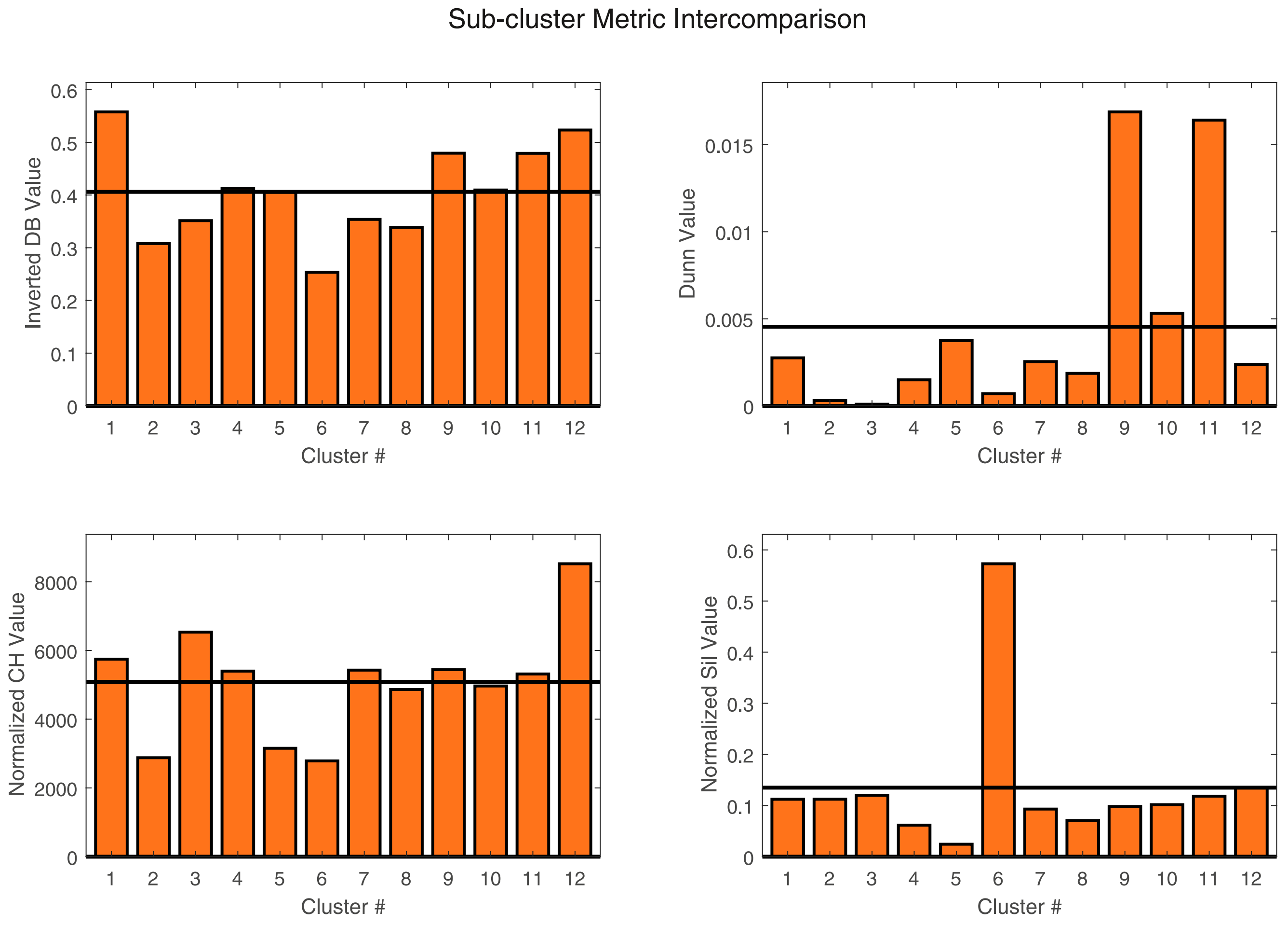

Figure 10. The values for the DB index, Dunn index, CH index, and mean silhouette values. Each of these variables is calculated using the subclusters as outlined for the DB index in the methodology section. Due to computational restrictions, these values are just estimates of the metric value. The black line indicates the mean value for the corresponding metrics over all of the clusters.

6. Conclusions

The Davies-Bouldin index and subsom entropy metrics are used to explore the variability within a set of cloud clusters. These metrics are used to examine subclusters generated from a set of clusters defined using CTP-COT joint histograms from the MODIS data set in Schuddeboom et al. (2018). The SOM algorithm was then reapplied to these clusters to develop a set of six subclusters for each of the established clusters. A particular effort was made to ensure that the methods used to investigate variability were scalable to systems with a larger number of clusters. This is achieved by working in a framework of first identifying any unusual clusters and then examining them individually. The full set of subclusters were examined using the Davies-Bouldin index and subsom entropy and several clusters were identified as outliers. In particular, clusters 2 and 6 (which correspond to tropical cloud and clear skies, respectively) were found to be the largest outliers. These results were then examined more closely by splitting the Davies-Bouldin index into its components, *S* and *M*, and then examining directly. This provided a more detailed understanding of the subclusters and the nature of the variability within the clusters. This was particularly valuable for understanding clusters 1, 2, 5, 6, and 9 (associated with tropical cloud, land-based cloud, and mixed layer cloud) each of which showed some evidence of complex internal variability within the originally defined cluster. These results were also partitioned into different regions of the globe showing a general agreement with the global results. Some clusters were identified as showing regionally distinct behavior over particular regions such as clusters 1 and 11 and also some regions which showed major deviations in behavior such as the North Ocean region.

The metric analysis consistently identified cluster 6 as a major outlier. Globally this cluster shows both the largest Davies-Bouldin index value and the smallest subsom entropy value. To better understand the variability in cluster 6 results, the subcluster histograms and geographic distributions of subcluster occurrence rates were examined. Subcluster 4 of cluster 6 was shown to be dominant subcluster accounting for 71% of the cluster 6 occurrences with subcluster 3 the next most common at 13%. These subclusters are associated with radically different types of clouds, with a difference in average cloud fraction of over 70%. The geographic distributions of the subcluster occurrence rates were examined and showed that subcluster 4 was dominant over desert regions, behaving exactly as expected given the low cloud fraction associated with the cluster. The second most frequent subcluster of cluster 6 (subcluster 3) was concentrated over the midlatitudes and mountainous regions. This is somewhat unexpected given the high cloud cover over this region but could be the result of these histograms not having a suitable place in any other cluster. The distributions of SW CRE and LW CRE were also examined and showed large differences between subclusters 3 and 4. This independently shows that these two subclusters represent very different types of clouds. In addition to cluster 6, the subclusters of cluster 1 were examined using the same approaches and as expected from our metrics showed much smaller variations than in cluster 6.

The metrics used in this paper were able to identify clusters that have large internal variability. More detailed analysis was then used to better understand the nature of this variability. Combined with the development of subclusters, this presents a new approach for understanding an established set of clusters. This introduced framework is likely the most useful result of this paper. While there has been some past research focused on understanding clusters through subclusters, it has generally been limited and often only for the purpose of cluster number identification. Similarly, using the DB index in the manner it is used in this paper is a significant deviation from most past usage.

There are several clear possibilities for future development and refinement of this approach. For instance, these metrics would be particularly useful on complex data sets with a larger number of identified clusters. Examples of this include the intercomparison of several climate models as described in Schuddeboom and McDonald (2021) or the organization of large document databases as in Kohonen (2013). The metrics would be particularly useful for these applications because it could allow the rapid identification of clusters that show unusual behavior, a process that would normally require significant manual investigation with a large data set. Another extension would be introducing a second variable that is different from what is used to generate the original clusters and then using this second variable to generate the subclusters. This could aide in understanding the strength of different physical relationships to the cloud clusters. A clear example of this approach would be building on the work linking vertical velocity measurements with the cloud clusters in McDonald and Parsons (2018) by examining clusters generated from CTP-COT histograms and then generating subclusters using vertical velocity. We are also interested in the potential that fuzzy clustering approaches could have in this space.

Data Availability Statement

The cloud clusters that were used as the basis for this research are accessible at https://doi.org/10.5281/zenodo.1202280 (Schuddeboom, 2018). The CERES data were obtained from https://ceres.larc.nasa.gov/ (Doelling, 2013), and the MODIS data were obtained from https://modis.gsfc.nasa.gov/ (Platnick et al., 2015).

Acknowledgments
The authors would like to thank the reviewers for improving the quality of this manuscript and acknowledge the teams behind the CERES and MODIS data sets. This work was supported by funding from the Deep South National Science challenge.

References

Allinson, N. M., & Ellis, A. W. (1992). Face recognition: Combining cognitive psychology and image engineering. *Electronics & Communication Engineering Journal*, *4*(5), 291–300. https://doi.org/10.1049/ecej:19920050

Ambroise, C., Sèze, G., Badran, F., & Thiria, S. (2000). Hierarchical clustering of self-organizing maps for cloud classification. *Neurocomputing*, *30*(1–4), 47–52. https://doi.org/10.1016/S0925-2312(99)00141-1

Auger, J.-M., Idan, Y., Chevallier, R., & Dorizzi, B. (1992). Complementary aspects of topological maps and time delay neural networks for character recognition. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks* (Vol. 4, pp. 444–449). IEEE. https://doi.org/10.1109/IJCNN.1992.227304

Bannon, P. R. (2015). Entropy production and climate efficiency. *Journal of the Atmospheric Sciences*, *72*(8), 3268–3280. https://doi.org/10.1175/JAS-D-14-0361.1

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Campelo, A. D. S., Farias, V. J. C., Tavares, H. R., & da Rocha, M. P. D. C. (2014). Self-organizing maps and entropy applied to data analysis of functional magnetic resonance images. *Applied Mathematical Sciences*, *8*(100), 4953–4969. https://doi.org/10.12988/ams.2014.310585

Cassano, J. J., Uotila, P., Lynch, A. H., & Cassano, E. N. (2007). Predicted changes in synoptic forcing of net precipitation in large Arctic River basins during the 21st century. *Journal of Geophysical Research*, *112*, G04S49. https://doi.org/10.1029/2006JG000332

Cavazos, T. (2000). Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *Journal of Climate*, *13*(10), 1718–1732. https://doi.org/10.1175/1520-0442(2000)013<1718:USOMTI>2.0.CO;2

Cho, N., Tan, J., & Oreopoulos, L. (2021). Classifying planetary cloudiness with an updated set of MODIS cloud regimes. *Journal of Applied Meteorology and Climatology*, *60*(7), 981–997. https://doi.org/10.1175/JAMC-D-20-0247.1

Coggins, J. H. J., McDonald, A. J., & Jolly, B. (2014). Synoptic climatology of the ross ice shelf and ross sea region of Antarctica: k-means clustering and validation. *International Journal of Climatology*, *34*(7), 2330–2348. https://doi.org/10.1002/joc.3842

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning*, *6*(1), 81–92. https://doi.org/10.1023/A:1022694001379

Dittenbach, M., Rauber, A., & Merkl, D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*, *48*(1–4), 199–216. https://doi.org/10.1016/S0925-2312(01)00655-5

Doan, Q.-V., Kusaka, H., Sato, T., & Chen, F. (2021). S-SOM v1.0: A structural self-organizing map algorithm for weather typing. *Geoscientific Model Development*, *14*(4), 2097–2111. https://doi.org/10.5194/gmd-14-2097-2021

Doelling, D. R. (2013). CER_SYN1deg-Day_Terra-Aqua-MODIS_Edition4A. *NASA Atmospheric Science Data Center (ASDC)*. https://doi.org/10.5067/Terra+Aqua/CERES/SYN1degDay_L3.004A

Doelling, D. R., Loeb, N. G., Keyes, D. F., Nordeen, M. L., Morstad, D., Nguyen, C., et al. (2013). Geostationary enhanced temporal interpolation for ceres flux products. *Journal of Atmospheric and Oceanic Technology*, *30*(6), 1072–1090. https://doi.org/10.1175/JTECH-D-12-00136.1

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Gibson, P. B., Perkins-Kirkpatrick, S. E., Uotila, P., Pepler, A. S., & Alexander, L. V. (2017). On the use of self-organizing maps for studying climate extremes. *Journal of Geophysical Research: Atmospheres*, *122*, 3891–3903. https://doi.org/10.1002/2016JD026256

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002b). Cluster validity methods: Part I. *SIGMOD Record*, *31*(2), 40–45. https://doi.org/10.1145/565117.565124

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002a). Clustering validity checking methods: Part II. *SIGMOD Record*, *31*(3), 19–27. https://doi.org/10.1145/601858.601862

Harrington, L. J., Gibson, P. B., Dean, S. M., Mitchell, D., Rosier, S. M., & Frame, D. J. (2016). Investigating event-specific drought attribution using self-organizing maps. *Journal of Geophysical Research: Atmospheres*, *121*, 766–812. https://doi.org/10.1002/2016JD025602

Hewitson, B. C., & Crane, R. G. (2002). Self-organizing maps: Applications to synoptic climatology. *Climate Research*, *22*(1), 13–26. https://doi.org/10.3354/cr022013

Hewitt, H. T., Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R. S., Keen, A. B., et al. (2011). Design and implementation of the infrastructure of HadGEM3: The next-generation Met Office climate modelling system. *Geoscientific Model Development*, *4*(2), 223–253. https://doi.org/10.5194/gmd-4-223-2011

Jakob, C. (2003). An improved strategy for the evaluation of cloud parameterizations in GCMS. *Bulletin of the American Meteorological Society*, *84*(10), 1387–1402. https://doi.org/10.1175/BAMS-84-10-1387

Jin, D., Oreopoulos, L., Lee, D., Tan, J., & Kim, K. M. (2020). Large-scale characteristics of tropical convective systems through the prism of cloud regime. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD031157. https://doi.org/10.1029/2019JD031157

Kidson, J. W. (2000). An analysis of New Zealand synoptic types and their use in defining weather regimes. *International Journal of Climatology*, *20*(3), 299–316. https://doi.org/10.1002/(SICI)1097-0088(20000315)20:3<299::AID-JOC474>3.0.CO;2-B

King, M., Menzel, W., Kaufman, Y., Tanre, D., Bo-Gao, C., Platnick, S., et al. (2003). Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from MODIS. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(2), 442–458. https://doi.org/10.1109/TGRS.2002.808226

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1–3), 1–6. https://doi.org/10.1016/S0925-2312(98)00030-7

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. https://doi.org/10.1016/j.neunet.2012.09.018

Kohonen, T., Kaski, S., & Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, *9*(6), 1321–1344. https://doi.org/10.1162/neco.1997.9.6.1321

Krützmann, N. C., McDonald, A. J., & George, S. E. (2008). Identification of mixing barriers in chemistry-climate model simulations using Rényi entropy. *Geophysical Research Letters*, *35*, L06806. https://doi.org/10.1029/2007GL032829

Leinonen, J., Lebsock, M. D., Oreopoulos, L., & Cho, N. (2016). Interregional differences in MODIS-derived cloud regimes. *Journal of Geophysical Research: Atmospheres*, *121*, 11648–11665. https://doi.org/10.1002/2016JD025193

Mason, S., Fletcher, J. K., Haynes, J. M., Franklin, C., Protat, A., & Jakob, C. (2015). A hybrid cloud regime methodology used to evaluate Southern Ocean cloud and shortwave radiation errors in ACCESS. *Journal of Climate*, *28*(15), 6001–6018. https://doi.org/10.1175/JCLI-D-14-00846.1

McDonald, A. J., & Cairns, L. H. (2020). A new method to evaluate reanalyses using synoptic patterns: An example application in the ross sea/ross ice shelf region. *Earth and Space Science*, *7*, e2019EA000794. https://doi.org/10.1029/2019EA000794

McDonald, A. J., Cassano, J. J., Jolly, B., Parsons, S., & Schuddeboom, A. (2016). An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states. *Journal of Geophysical Research: Atmospheres*, *121*, 13009–13030. https://doi.org/10.1002/2016JD025199

McDonald, A. J., & Parsons, S. (2018). A comparison of cloud classification methodologies: Differences between cloud and dynamical regimes. *Journal of Geophysical Research: Atmospheres*, *123*, 11173–11193. https://doi.org/10.1029/2018JD028595

Oreopoulos, L., Cho, N., Lee, D., & Kato, S. (2016). Radiative effects of global MODIS cloud regimes. *Journal of Geophysical Research: Atmospheres*, *121*, 2299–2317. https://doi.org/10.1002/2015JD024502

Oreopoulos, L., Cho, N., Lee, D., Kato, S., & Huffman, G. J. (2014). An examination of the nature of global MODIS cloud regimes. *Journal of Geophysical Research: Atmospheres*, *119*, 8362–8383. https://doi.org/10.1002/2013JD021409

Palomo, E. J., North, J., Elizondo, D., Luque, R. M., & Watson, T. (2012). Application of growing hierarchical SOM for visualisation of network forensics traffic data. *Neural Networks*, *32*, 275–284. https://doi.org/10.1016/j.neunet.2012.02.021

Platnick, S., King, M. D., Ackerman, S. A., Menzel, W. P., Baum, B. A., Riédi, J. C., & Frey, R. A. (2003). The MODIS cloud products: Algorithms and examples from terra. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(2 Part 1), 459–472. https://doi.org/10.1109/TGRS.2002.808301

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., et al. (2017). The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(1), 502–525. https://doi.org/10.1109/TGRS.2016.2610522

Platnick, S., Hubanks, P., Meyer, K., & King, M. D. (2015). MODIS atmosphere L3 daily product. NASA MODIS adaptive processing system. *Goddard Space Flight Center*. Retrieved from https://modis-atmos.gsfc.nasa.gov/products/daily

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saini, N., Saha, S., & Bhattacharyya, P. (2019). Automatic scientific document clustering using self-organized multi-objective differential evolution. *Cognitive Computation*, *11*(2), 271–293. https://doi.org/10.1007/s12559-018-9611-8

Samsonova, E. V., Kok, J. N., & Ijzerman, A. P. (2006). TreeSOM: Cluster analysis in the self-organizing map. *Neural Networks*, *19*(6–7), 935–949. https://doi.org/10.1016/j.neunet.2006.05.003

Schuddeboom, A. J. (2018). Modis and GA 7.0 cluster analysis results. *Zenodo*. https://doi.org/10.5281/zenodo.1202280

Schuddeboom, A. J., & McDonald, A. J. (2021). The Southern Ocean radiative bias, cloud compensating errors, and equilibrium climate sensitivity in CMIP6 models. *Journal of Geophysical Research: Atmospheres*, *126*, e2021JD035310. https://doi.org/10.1029/2021JD035310

Schuddeboom, A. J., McDonald, A. J., Morgenstern, O., Harvey, M., & Parsons, S. (2018). Regional regime-based evaluation of present-day general circulation model cloud simulations using self-organizing maps. *Journal of Geophysical Research: Atmospheres*, *123*, 4259–4272. https://doi.org/10.1002/2017JD028196

Schuddeboom, A. J., Varma, V., McDonald, A. J., Morgenstern, O., Harvey, M., Parsons, S., et al. (2019). Cluster-based evaluation of model compensating errors: A case study of cloud radiative effect in the Southern Ocean. *Geophysical Research Letters*, *46*, 3446–3453. https://doi.org/10.1029/2018GL081686

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE-Mobile Computing and Communications Review*, *5*(1), 3–55. https://doi.org/10.1145/584091.584093

Sheridan, S. C., & Lee, C. C. (2011). The self-organizing map in synoptic climatological research. *Progress in Physical Geography*, *35*(1), 109–119. https://doi.org/10.1177/0309133310397582

Smith, G., Priestley, K., Loeb, N., Wielicki, B., Charlock, T., Minnis, P., et al. (2011). Clouds and Earth Radiant Energy System (CERES), a review: Past, present and future. *Advances in Space Research*, *48*(2), 254–263. https://doi.org/10.1016/j.asr.2011.03.009

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600. https://doi.org/10.1109/72.846731

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B., Smith, G. L., & Cooper, J. E. (1996). Clouds and the Earth's Radiant Energy System (CERES): An Earth observing system experiment. *Bulletin of the American Meteorological Society*, *77*(5), 853–868. https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2

Williams, K. D., & Tselioudis, G. (2007). GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Climate Dynamics*, *29*(2–3), 231–250. https://doi.org/10.1007/s00382-007-0232-2

Williams, K. D., & Webb, M. J. (2009). A quantitative performance assessment of cloud regimes in climate models. *Climate Dynamics*, *33*(1), 141–157. https://doi.org/10.1007/s00382-008-0443-1

Zhang, L., & Yu, H. (2006). RSOM algorithm for radar target recognition. *Paper presented at CIE International Conference of Radar Proceedings*. https://doi.org/10.1109/ICR.2006.343241