Automated identification of characteristic droplet size distributions in stratocumulus clouds utilizing a data clustering algorithm

Nithin Allwayin[1], Michael Larsen[2], Alexander Shaw[3], and Raymond Shaw[1]

[1]Michigan Technological University
[2]Michigan Technological University,College of Charleston
[3]Brigham Young University

November 23, 2022

Abstract

Droplet-level interactions in clouds are often parameterized by a modified gamma fitted to a "global" droplet size distribution. Do "local" droplet size distributions of relevance to microphysical processes look like these average distributions? This paper describes an algorithm to search and classify characteristic size distributions within a cloud. The approach combines hypothesis testing, specifically the Kolmogorov-Smirnov (KS) test, and a widely-used machine-learning algorithm for identifying clusters of samples with similar properties: Density-based spatial clustering of applications (DBSCAN). The two-sample KS test does not presume any specific distribution, is parameter free, and avoids biases from binning. Importantly, the number of clusters is not an input parameter of the DBSCAN algorithm, but is independently determined in an unsupervised fashion. As implemented, it works on an abstract space from the KS test results, and hence spatial correlation is not required for a cluster. The method is explored using data obtained from Holographic Detector for Clouds (HOLODEC) deployed during the Aerosol and Cloud Experiments in the Eastern North Atlantic (ACE-ENA) field campaign. The algorithm identifies evidence of the existence of clusters of nearly-identical local size distributions. It is found that cloud segments have as few as one and as many as seven characteristic size distributions. To validate the algorithm's robustness, it is tested on a synthetic dataset and successfully identifies the predefined distributions at plausible noise levels. The algorithm is general and is expected to be useful in other applications, such as remote sensing of cloud and rain properties.

Automated identification of characteristic droplet size distributions in stratocumulus clouds utilizing a data clustering algorithm

Nithin Allwayin, [a] , Michael L. Larsen , [a,b] Alexander G. Shaw, [c] and Raymond A. Shaw [a]

[a] *Michigan Technological University, Houghton, Michigan*

[b] *College of Charleston, Charleston, South Carolina*

[c] *Brigham Young University, Provo, Utah*

Corresponding author: Michael L. Larsen, LarsenML@cofc.edu

Corresponding author: Raymond A. Shaw, rashaw@mtu.edu

ABSTRACT: Droplet-level interactions in clouds are often parameterized by a modified gamma fitted to a "global" droplet size distribution. Do “local” droplet size distributions of relevance to microphysical processes look like these average distributions? This paper describes an algorithm to search and classify characteristic size distributions within a cloud. The approach combines hypothesis testing, specifically the Kolmogorov-Smirnov (KS) test, and a widely-used machine-learning algorithm for identifying clusters of samples with similar properties: Density-based spatial clustering of applications (DBSCAN). The two-sample KS test does not presume any specific distribution, is parameter free, and avoids biases from binning. Importantly, the number of clusters is not an input parameter of the DBSCAN algorithm, but is independently determined in an unsupervised fashion. As implemented, it works on an abstract space from the KS test results, and hence spatial correlation is not required for a cluster. The method is explored using data obtained from Holographic Detector for Clouds (HOLODEC) deployed during the Aerosol and Cloud Experiments in the Eastern North Atlantic (ACE-ENA) field campaign. The algorithm identifies evidence of the existence of clusters of nearly-identical local size distributions. It is found that cloud segments have as few as one and as many as seven characteristic size distributions. To validate the algorithm’s robustness, it is tested on a synthetic dataset and successfully identifies the predefined distributions at plausible noise levels. The algorithm is general and is expected to be useful in other applications, such as remote sensing of cloud and rain properties.

SIGNIFICANCE STATEMENT: A typical cloud can have billions of drops spread over tens or hundreds of kilometers in space. Keeping track of the sizes, positions, and interactions of all of these droplets is impractical and, as such, information about the relative abundance of large and small drops is typically quantified with a "size distribution." But droplets in a cloud interact locally, so this work is motivated by the question of whether the cloud droplet size distribution is different in different parts of a cloud. A new method, based on hypothesis testing and machine learning, determines how many different size distributions a given cloud contains. This is important because the size distribution describes processes like cloud droplet growth and light transmission through clouds.

1. Introduction

A considerable portion of Earth's oceans is swathed by low-level stratocumulus clouds, enough to contribute to the planetary albedo significantly (Hahn and Warren 2007). Changes in the extent or coverage of these clouds can substantially impact global climate (Slingo 1990; Hartmann et al. 1992; Stephens 2005). Because droplet scales remain unresolved in climate and other coarse-resolution models, the processes involving drop-drop interactions are parameterized, often based on *in-situ* cloud observations. It is common to assume a functional form for cloud droplet size distributions in such numerical models, and similar assumptions are commonly made in remote sensing retrieval algorithms (Straka 2009; Shaw 2016; Igel and van den Heever 2017). Although several different forms including lognormal, exponential, and Weibull distributions have been used, most of the community has gravitated towards using a modified gamma distribution (Miles et al. 2000).

The work reported here was motivated by what started as a simple question: if we sample a small, localized volume of cloud, will the resulting droplet size distribution look like the macroscopically-averaged size distribution? Stated differently, do droplets interacting on microphysically-relevant scales "see" a gamma distribution? This leads naturally to hypothesis testing: what is the likelihood that a measured size distribution is a realization of a specified, theoretical size distribution? Or what is the likelihood that any two measured size distributions are sampled from the same distribution? As the work progressed, several related questions emerged. What scales must one average over in order to achieve convergence to a 'global' distribution? More intriguingly, might a seemingly

homogeneous cloud be described by a small number of clearly distinguishable, characteristic droplet size distributions throughout its interior? If so, are these distinguishable droplet size distributions localized within particular spatial parts of the cloud? Do the number of the characteristic droplet size distributions change from cloud to cloud or vary at different heights within the same cloud? To be clear, our intention in this paper is not to explore the physics of these interesting questions, but rather to introduce and illustrate the set of tools developed to identify characteristic cloud droplet size distributions from *in-situ* observations that could form the foundation for investigating the above questions. The tools bring together in a unique way methods of statistical hypothesis testing and of machine-learning-based cluster analysis.

From the observational side, disentangling sampling variability (each measurement only observes a certain number of drops), instrument uncertainty (all observed drops have a sizing uncertainty), and natural variability (the underlying drop size distribution may actually change from one part of a cloud to another) is challenging. Global and local processes make cloud systems inherently variable, and a significant challenge lies in quantifying this variability. These questions related to variability have been studied in the rain and cloud measurement communities and we have learned that, in general, spatial and temporal averaging can result in different statistical properties (Jameson et al. 2015a), atmospheric particulate data often do not pass tests for wide-sense stationarity or statistical homogeneity (Larsen et al. 2005; Larsen and O'Dell 2016; Jameson et al. 2018), and care must be taken in data analysis to ensure that samples are taken over appropriate spatial and temporal scales to optimize the trade-off between larger sampling volumes that minimize sampling variability and smaller sampling volumes that minimize artificial removal of natural variability (Jameson and Kostinski 2000; Jaffrain and Berne 2011; Jameson et al. 2015b; Larsen et al. 2018). The method introduced here attempts to identify droplet size distributions that are statistically similar, in spite of the natural and measurement uncertainties, by starting with the method of hypothesis testing. The method avoids the need to identify "appropriate" spatial or temporal averaging scales and instead identifies characteristic droplet size distributions that are not required to be spatially or temporally localized. Once the characteristic droplet size distributions are identified, it is then possible to explore whether they are more prevalent in certain spatial cloud regions or environmental conditions. This paper is focused on the first step, namely to identify the characteristic distributions.

The semi-parametric algorithm described here allows for the exploration of any *in-situ* data having spatially tagged information regarding particle (in this case cloud drop) detections. Specifically, we use data captured by the HOLODEC (Holographic Detector for Clouds) instrument (Fugal et al. 2004; Fugal and Shaw 2009; Spuler and Fugal 2011) during the ACE-ENA (Aerosol and Cloud Experiments in the Eastern North Atlantic) field project (Wang et al. 2021). In contrast to most cloud-sampling instruments that average over long distances to give a statistically significant distribution, HOLODEC samples all the droplets in a small volume (≈ 19 cm^3) to determine droplet positions and sizes within an individual hologram (Fugal et al. 2009). Thus, each HOLODEC sample contains a population of droplets and a corresponding, localized measurement of the droplet size distribution (Beals et al. 2015). The distance between these samples depends on aircraft speed, and is approximately 30 m for ACE-ENA.

The algorithm introduced here employs established statistical and machine learning tools, namely the Kolmogorov–Smirnov (KS) test and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). These tools are used to scan the ensemble of hologram volumes for similar size distributions, which are then grouped to form what we call "characteristic distributions", endemic to the cloud in question. Of particular note is that the method employed here does not make an *a priori* assumption regarding the functional form of the cloud droplet size distribution. The Kolmogorov–Smirnov (KS) test has been previously used with HOLODEC data to assess the spatial uniformity of droplets within a hologram or between neighboring holograms (Glienke et al. 2020). Here, we use machine learning to significantly expand on that work in order to not only identify regions where the size distribution is statistically similar, but also to identify the number of different size distributions and their associated locations within the cloud.

The remainder of this manuscript outlines the schema of the algorithm (section 2), presents sample results from when this algorithm is applied to HOLODEC data from the ACE-ENA campaign (section 3), explores the robustness of the algorithm by examining the characteristic size distributions revealed on synthetic data with prescribed statistical structure (section 4), and overarching results are discussed (section 5).

2. Method

Our method works by finding holograms with statistically similar size distributions and using the collection of these holograms to define a cluster. Specifically, note that this is not a cluster in space, but a cluster of hologram samples that have similar "characteristic" size distributions that may come from different regions within a cloud. The similarity between any two distributions is determined using the Kolmogorov-Smirnov (KS) test, and the grouping is done with the density-based clustering algorithm DBSCAN.

a. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is a non-parametric statistical test to determine if a sample probability distribution function could be a subset of a reference distribution (Kendall and Stuart 1979). The two-sample version of the test compares two measured distributions to determine if they could be from the same parent distribution. A significant advantage of the KS test is its dependence on cumulative distribution functions (CDFs) and therefore the avoidance of spurious results from arbitrary binning of data (Barlow 1993; Glienke et al. 2020). The test's key metric is the maximum distance between the two sample CDFs; the larger the distance between the CDFs the less likely the two distributions come from the same statistical distribution. The result of such a KS comparison is usually represented in a binary fashion, indicating either success (the distributions could be from the same parent distribution) or failure (the distributions likely are not from the same parent distribution). We use the built-in MATLAB function *kstest2* with its default alpha value of 0.05 to compare two measured distributions (95% confidence level). The MATLAB function returns 0 for success and 1 for failure. This is illustrated in Figure 1, where we see the CDFs of droplet size distributions from three holograms. The KS test compares the each pair of CDFs. CDFs 1 and 2 are very similar, so the KS test identifies them to be from the same parent distribution. On the other hand, CDF 3 is noticeably different than CDFs for holograms 1 and 2; the maximum difference between CDFs 1/3 and 2/3 is much larger, so the KS test gives a "failure" result indicating that hologram 3 has a different size distribution than hologram 1 and 2.

We employ the KS test to compare the cloud droplet diameter distributions from all hologram pairs in a sequence of holograms measured during a cloud transect with the HOLODEC sensor. All data from the sequence of holograms have similar characteristics (e.g., all measurements have

the same lower droplet diameter cutoff of 10 μm, the utilized sample volume for each hologram is the same, and any instrumental imperfections are expected to be consistent from hologram to hologram). The distributions from each hologram are compared against those from all other holograms, including itself.

Previous work using the KS test has noted that sample size determines the step size of the empirical CDF and therefore KS testing can be very sensitive to the sample sizes of the two distributions (e.g., see the discussion of Figure 4 in Glienke et al. (2020)). To avoid this issue, our analysis fixes the number of droplets in each hologram to a uniform cutoff value. This cutoff is set to be 70 percent of the mean number of droplets per hologram; all the holograms with droplet numbers less than the cutoff are removed from the KS testing process. For all holograms that have a number of drops that exceed this cutoff value, we sample (without replacement) droplets from each hologram to the cutoff value. This gives a consistent data size and CDF resolution for all KS tests. To minimize the associated sampling uncertainties, we create an ensemble of such samples and conduct the KS test for each ensemble member. The average of the results of the KS test for the ensemble members gives the final result. This converts the otherwise binary output to a value between 0 and 1, namely the fraction of ensemble KS tests for which the null hypothesis was rejected. A number close to 0 indicates that the two holograms have drop size distributions that likely come from the same parent distribution, whereas a number close to 1 suggests the holograms have drop size distributions that are unlikely to be drawn from the same parent distribution. Thus, for a set of n holograms, we will have n^2 of these 0-1 outputs, each of which is the result of an ensemble average of inter-comparisons between the empirical hologram drop size distributions sampled to the cutoff value.

If these KS test results from each hologram are arranged as an array, we can construct a matrix of size $n \times n$ indicating a measure of the likelihood of dissimilarity between the associated size distributions between the holograms in the associated row and column of the matrix. In this work, we call this the KS matrix, and Figure 2 (a) depicts a cartoon of such a matrix for $n = 25$ synthetic holograms. The data for the matrix is drawn from three different distributions with 13, 6, and 1 holograms belonging to each of the different distributions. The other 5 holograms have a random distribution and constitute "noise" holograms that are not drawn from the three pre-assigned parent distributions. The ensemble size for the sub-samples is 1000, and thus each cell in this matrix is

from a thousand KS tests and has a value in the range of 0 to 1 (with ensemble-based resolution of 1/1000). Visually, we can clearly identify some holograms with similar distributions from the KS matrix by looking for rows or columns with similar visual structure to other rows or columns. These holograms constitute a cluster, and the set of such clusters form the "characteristic distributions" in the cloud segment. It is sub-optimal and non-objective to detect all such clusters visually; they can be better classified using unsupervised algorithmic clustering techniques.

b. Density-based spatial clustering of applications with noise

To identify clusters within our droplet-size distribution measurements, we implement the Density-based spatial clustering of applications with noise (DBSCAN) algorithm. Many popular clustering algorithms rely on the user to specify the number of clusters as an input parameter. For our data, pre-assignment of the number of clusters biases the process and thus we require an algorithm that determines the number of clusters in a dataset and assigns its members to these clusters, making DBSCAN a natural choice. Here, each of these members can be imagined as points in space. This space is abstract and depends on the metric used to identify the clusters. DBSCAN is not purely non-parametric; the user is required to input a value (epsilon), determining how close a point must be to a cluster for it to be included in the cluster. It does not imply here that the points be spatially close but rather the user must specify a metric to compute this "closeness". Additionally, the user specifies the minimum number of points (min-points) required to define a single cluster, thereby helping to eliminate spurious clusters (Ester et al. 1996; Gan et al. 2020). These inputs are called hyperparameters, henceforth referred to as parameters.

Abstractly, the algorithm works by choosing an arbitrary point and scanning the space to count all points within the "closeness" limit. If this number is greater than or equal to the minimum number needed to form a cluster, they are all classified as a cluster. If not, the starting point is identified as noise (e.g. not a member of any cluster). Once a cluster is identified, the same process is repeated for all other points in the cluster. Each point now acts as the defining point in the cluster, adding additional points to the clusters if the criteria for the closeness and minimum number of points are met. Once all the points in the cluster are identified, the algorithm randomly chooses another (unclassified) point, and the whole process repeats until all the points are either classified or remain outside of any cluster as outliers. These outliers constitute the "noise".

Note that in this context, a "point" refers to a hologram in the segment. By grouping different points we seek to find sets of holograms that have similar size distributions. To assign these clusters using DBSCAN we use the average KS score between two hologram pairs as our metric defining the "closeness" between two points. For a hologram, these closeness values with the other holograms are illustrated in the corresponding row of the KS matrix, with a value close to 0 indicating that the hologram pairs have very similar underlying drop size distributions. Analogous to the original method, DBSCAN works by choosing a random point or hologram, represented by the corresponding row in the KS matrix. The elements of this row fills the space for the cluster search, with "closeness" indicated by the KS similarity scores of these row elements. The closeness or minimum distance between the specified hologram and the other holograms in the analysis can thus have a KS score similarity value between 0 and 1. If the specified hologram has at least the minimum number other holograms separated from it with a KS score similarity value below the DBSCAN closeness parameter, these other holograms, along with the starting hologram, are identified as a cluster. Every row corresponding to the other holograms in this cluster iterate through a turn as the defining point. Then, a new as-yet unclustered hologram is chosen to potentially start a new cluster. This process repeats until all the holograms are classified and the resulting clusters represent collections of holograms having similar droplet-size distributions within the cloud.

The results of DBSCAN depend on the two user-defined parameters. Details of how these parameters are chosen for our application are explained in subsection d. Figure 2 (b) shows the results of DBSCAN applied to the sample synthetic KS Matrix defined earlier. The two parameters, epsilon (defining the minimum "closeness") and min-points (defining the minimum number of points required to define a cluster), are 0.1 and 5, respectively. Two clusters of sizes 10 and 5 are identified with 100 percent accuracy. The number of the holograms in the identified clusters is less than the actual number of holograms associated with each class of drop size distributions (13, 6, and 1 as stated earlier) because some points in the sample space are removed because the associated hologram did not meet the cutoff criteria for the minimum number of drops in the hologram, and a distribution containing only 1 hologram falls below the minimum number of points parameter set in the algorithm (5). A more detailed validation of the algorithm is performed in section 4.

c. Algorithm

The algorithm is constructed specifically for the data from the HOLODEC instrument, but it can readily be adapted to similar observations where a distribution of a random variable is measured at regular spatial or temporal intervals. The holograms from HOLODEC give size distributions for the cloud droplets at different points in the cloud. Many clouds have droplet concentrations that fluctuate greatly, resulting in some holograms having drop counts significantly below the ensemble average and thus removed from the analysis. The algorithm is most logically implemented in clouds that visually appear homogeneous – for example, using data taken from flight segments with near constant altitude and approximately steady number concentrations. A full description of the algorithm as applied to *in situ* flight data follows.

1. The holograms from a cloud segment at a constant altitude and having relatively steady number concentration are selected. Each hologram provides a size distribution of cloud droplets. Let this number of holograms be n.

2. A cutoff is defined for the minimum sample size required for each size distribution to be included in further analysis. This is set to 70 percent of the mean number of droplets in a hologram, obtained by dividing the total droplets in the entire segment by the total number of holograms. All the holograms having fewer detected drops than this cutoff (say x) are removed from the analysis.

3. Now, we select the first hologram in the data-set that meets the cutoff criteria as the primary hologram. It is then repeatedly sub-sampled to create an ensemble, each having the same number of cloud droplets (equal to the cutoff). Similar ensembles are created for all other holograms in the segment. Each ensemble member of the primary hologram is now compared using the KS test to ensemble members of all holograms with drop numbers exceeding the cutoff, including itself. The KS results from the ensemble comparison between every hologram pair are averaged to get a mean KS score. We now have a vector with length $(n-x)$ summarizing the KS results for the primary hologram, with each element in the vector taking on a value in the range of 0 (meaning the KS test implies the size distribution matches for all members of the ensembles for both holograms) to 1 (meaning the KS test implies the size distribution does not match for any members of the first hologram's ensemble when compared

to the second hologram's ensemble). These $(n-x)$ values are used to populate a row of length n in the KS matrix where the remaining elements (matching the primary hologram to the holograms that did not meet the cutoff) are given a non-numerical place-holder. This helps to visualize the relationships between spatial locations of holograms and the similarity of their associated size distributions.

4. The next hologram that meets the cutoff is then chosen as the primary hologram. The previous step repeats to generate the next row in the matrix; the process continues until the entire KS matrix is populated. Holograms that don't meet the cutoff have all n entries represented by a non-numerical place holder. A value in the matrix closer to 0 indicates that the two holograms (associated with the row and column indexes of the matrix) have populations from very similar size distributions, whereas a value close to 1 indicates clearly different drop size distributions. The non-numerical values denote holograms discarded from the clustering analysis.

5. The resultant KS matrix is fed into DBSCAN for cluster identification. The algorithm works on the KS space to group holograms into clusters. The user-defined parameters are chosen manually to get the results presented here.

d. Determination of the DBSCAN input parameters

The results of DBSCAN depend on its two input parameters and their specified values therefore can be adjusted according to the user's focus. In this study, we choose to select the combination of parameters that maximizes the number of clusters, while still maintaining what is considered a reasonable cluster size. It is done this way to explore all possible differences in cloud size distributions and hence might be sufficient to serve as an upper bound to the number of characteristic distributions in the cloud. To help understand how much the results of the algorithm vary with the input parameters, we performed a sensitivity test. In our analysis, we found that DBSCAN exhibits higher sensitivity to "min-points", which is found to be directly related to the detectability of the smallest clusters. On the other hand, the results were fairly insensitive to the "epsilon" parameter, which can be ascribed to the consistency of the hologram ensemble comparisons of the KS Matrix. They have an average value close to 0 for the pass cases and a value close to 1 for the failures, making clear cut distinction between the results. Hence, there is less sensitivity on "epsilon" values as they depend on the scores in this KS space. This practically reduces the problem at hand to

determining a single parameter (min-points). We choose to fix the value of "epsilon" at 0.1 and change "min-points" in steps of 5. The lower limit to "min-points" is set to 10. This is done to prevent a few holograms dominating the results and this cutoff roughly corresponds to about 1 percent of the holograms in the flight transect. More details on implementation of the algorithm with different input parameters are included in the supplement.

3. Using the Algorithm with Real Data

a. Dataset:ACE-ENA

The data set used in our study is derived from the HOLODEC deployment during the ACE-ENA (Aerosol and Cloud Experiments in the Eastern North Atlantic) campaign (Wang et al. 2021). The ACE-ENA campaign aimed at studying the low-level stratocumulus clouds near the Azores islands (Portugal) in the Atlantic Ocean. An extensive set of instruments on board the G1 aircraft, operated as part of the ARM (Atmospheric Radiation Measurement) Aerial Facility, were used to make various measurements of the cloud and the boundary layer over two Intensive Operational Periods (IOP). The G1 flight moved at an approximate speed of 100 m/s. As HOLODEC has a hologram acquisition frequency of 3.3 Hz, this would mean that the holograms obtained are about 30 meters apart. Therefore, a set of such holograms would capture the local variability across a large section of the cloud. In our analysis, we choose data from research flight on July 18, 2017 (RF18) from IOP 1. Datasets from different horizontal legs of the flight are selected to capture the vertical variability within the cloud. On this basis, we identified five such segments – two near the cloud top (S1 & S2), two in the mid cloud regions (S3 & S4), and one near the base of the cloud (S5). These segments and the corresponding droplet count histograms are shown in Figure 3. They correspond to altitudes of approximately 950, 850 and 750 meters, respectively. Detailed information about the different segments are given in Table 1.

b. Results

Cloud data from the HOLODEC probe are analyzed with the algorithm to look for the characteristic distributions. The initial step was to generate the KS matrices for all the cloud segments. DBSCAN is then employed to identify the different hologram clusters. The results of this classification are summarized in Table 2. While one segment had one characteristic distribution, most

had more than one identified, and one segment near cloud top had as many as 7 characteristic size distributions. A sizeable number of holograms from these segments are also identified as noise. These noise holograms have distributions that are different from those of the identified clusters.

An illustration of this result for Segment 1 is shown in Figure 4. The algorithm identifies 7 clusters for the chosen parameters. Out of these 7, 3 clusters have over 35 holograms, while the other smaller clusters have around ten holograms. The PDFs of the smaller clusters closely resemble the nearby bigger clusters and might be related to their closeness in spatial locations. The PDF of all the droplets from the entire segment (from holograms that satisfy the threshold limit) is also shown. The main clusters identified for this segment can also be obtained with reasonable accuracy with a visual inspection of the KS matrix. However, this is not possible for all segments. For example, the KS matrix for segment S4 is less sparse, and clusters cannot be easily made out visually. The algorithm successfully identifies two clusters with 362 and 50 holograms, respectively, which can be seen in Figure 5. Only a single cluster is found for segment S2, as seen in Figure 6, indicating that most of the holograms in this segment are similar to each other. In both these cases, the PDFs for the entire segment closely resembles that of the major cluster. This is expected as the primary cluster covers the bulk of the holograms for the segments. Two and four clusters are identified for the segments S3 & S5 respectively, the results of which can be found in more detail in the supplement. The parameters for DBSCAN are chosen by iterating through different sets of values for "min-points" as discussed before. The value of "epsilon" is fixed at 0.1. For the segment S1, maximum clusters are found for a "min-points" value of 10. Clusters are also identified for other values, generally decreasing in number for an increase in "min-points". This is because some clusters have only 10-20 members and hence are not detected as the minimum number of points is increased. For segment 2, there is only one cluster and it is insensitive to the parameters. For segments 3 and 4: there are only certain sets of values that give multiple clusters. The number of clusters identified for the segment S5 increases with "min-points" first and then decrease. The maximum value is found for a minimum number of 15.

Maximizing the number of clusters allows us to look for all reasonable differences in the size distributions. These differences between the clusters can be seen from their average PDFs for different segments. Figures 4(c), 5(c), and 6(c) illustrate this clearly. These can be compared to the average PDF of the entire segment(dashed black line) for all the holograms above the cutoff. The

standard deviation of the average PDFs is comparatively small, indicating that the size distribution of droplets from the holograms in each cluster are very similar.

We also fit the PDFs from each cluster to a modified gamma function. Modified gamma distributions are selected because of their wide use in representing cloud droplet populations in modeling and remote sensing communities. The gamma distribution is defined by two degrees of freedom, the shape(k) and scale(θ) parameters and is given by

$$f(d) = \frac{1}{\Gamma(k)} \left(\frac{d}{\theta}\right)^{k-1} d \, \exp\left(-\frac{d}{\theta}\right), \tag{1}$$

where d is the diameter of the droplets. The scale parameter (θ) has dimensions of length. The shape parameter (k) is non dimensional and determines how broad the distribution is. To avoid binning the diameters while creating a PDF, we fit the empirical CDFs of the hologram distributions to the CDF of the gamma distribution, given by

$$F(d) = \frac{1}{\Gamma(k)} \gamma\left(k, \frac{d}{\theta}\right). \tag{2}$$

Here Γ and γ are the upper and lower incomplete gamma functions, respectively. The shape and scale parameters obtained from these fits are shown in Figures 4(d), 5(d) and 6(d). Their mean values for the different clusters can be found in Table 2; the mean value for the full flight leg is also shown by the large dot in Figures 4(d), 5(d) and 6(d). There is a distinction between the gamma parameters for the different clusters, which might be difficult to infer from a scatter plot of these gamma parameters alone without the assistance provided by the clustering algorithm. The results of the fits to the data supplied here are generally within the range observed by Miles et al. (2000). The differences can be attributed to the detectability ranges of instruments used in the studies. Here HOLODEC is limited in resolution to reliable detection of cloud droplets larger than 10 μm diameter. This means that the mode diameter for our data will be larger and hence can help explain the difference in the size and shape parameters with those from Miles et al. (2000).

4. Validation with synthetic data

It is imperative to verify the reliability and robustness of the algorithm to validate the correctness of the results we obtained. For this purpose, we create a synthetic dataset that mimics the droplet

size distributions from HOLODEC. This synthetic dataset mimics a cloud transect with a specific set of holograms. Each such hologram contains simulated information regarding detected droplet sizes that are then used to form a cloud size distribution. For this synthetic data, we model all the distributions found in the cloud transect as modified gammas. However, one could in principle choose any other distribution or groups of distributions and the results are not expected to sensitively depend on the chosen distributions.

a. Synthetic data

Once modified gamma distributions are assumed for the droplets, we begin by mimicking the data for each individual hologram. For the synthetic dataset, we define three clusters with corresponding gamma parameters. The scale and shape parameters are 40, 60, & 80 μm and 0.6, 0.5, & 0.4 respectively. In addition to these clusters, we also define a gamma parameter space, and the draws from that space constitute the members of what will become the noise holograms. Here noise means that the distribution does not belong to any of the predefined clusters. However, as this noise parameter space also includes the region of scale and shape parameter space that generates the different clusters, a random draw from this space could generate a hologram that belongs to one of the clusters. All the holograms in the transect are assigned either belonging to one of the clusters or as a noise hologram. The number of holograms assigned to a defined cluster is chosen from a random draw. The droplet count in each hologram is also chosen randomly from a Gaussian distribution with mean and standard deviation similar to what we have for the actual data used in this study. To create the synthetic data set for a hologram, we first randomly chose its gamma parameters and the number of droplets. Random drop sizes are then drawn from a gamma distribution with these parameters. For a noise hologram, the gamma parameters are randomly chosen from the defined parameter space. As a final step, we remove all the droplets below 10 microns to set resolution limits similar to HOLODEC. This process is then repeated for all the holograms to create the synthetic dataset.

b. Results

The algorithm has been designed in such a way that we expect that it would be able to identify the three clusters present, even in the presence of noise and independently of the spatial locations

of both the noise and cluster holograms in the synthetic data set. We, therefore, create three different datasets with different numbers of noise holograms. They are labeled as SD1, SD2, and SD3, respectively, with the number of noise holograms increasing in each synthetic dataset. The respective proportion of clusters and noise are given in Table 3. These datasets are then processed using the algorithm after selecting the DBSCAN input parameters. For all cases, the predefined cluster members and parameters are identified with great accuracy. Thus, the results of the algorithm are not greatly dependent on the noise level. Note that the number of elements in each cluster is lower than the numbers defined in the dataset. This is because all the holograms with droplets less than the cutoff threshold are removed. Figure 7 outlines these results for the segment SD1. The data when fitted to the modified gamma distributions give the same mean shape and size parameters as the predefined values. A very few of the noise holograms are also recognized as part of the clusters. This is expected as the defined noise parameter space is relatively narrow, so some of the randomly drawn noise parameters might be similar to those defined for the clusters. The percentage of noise holograms in respective clusters increases with the number of noise holograms, as seen in SD2 and SD3. Notably, none of the holograms belonging to a cluster is erroneously classified into another cluster for all these cases.

There are, however, a few points that need to be addressed. The classification depends on the selection of the input parameters. If they are too lenient, say, for example, if the minimum number of cluster elements is too low, then more and more noise elements can induce the creation of spurious small clusters consisting of noise holograms. This means that the small number of random noise holograms that have similar parameters are recognized as a cluster. This can been seen for the segment SD2 given in Figure 8. Two smaller clusters are also identified in addition to the defined clusters. This is in no way a defect of the algorithm but caused by the high chance of creating new clusters from the narrow noise space. Similarly, when the PDFs of the clusters are close enough, and the noise holograms are drawn from a very narrow space between those clusters, the holograms with the two predefined gammas, along with the noise hologram, may be identified as a single cluster by the algorithm. Individually, the two predefined holograms might be close enough to the noise hologram to give a KS pass result causing all of them to be labelled as a single cluster. This is caused by the sensitivity of KS results and may lead to underestimation of the number of clusters present unless the parameters are properly selected.

5. Discussion

We started with a simple question: Do the "local" cloud size distributions match the "global" mean. The answer being usually no, we further expanded the question to see whether there is similarity between "local" size distributions. In other words, can we determine a characteristic set of droplet size distributions to describe a cloud? In our pursuit to answer this question, we developed a technique that determines the similarity between different distributions and then categorizes them into distinct sets. For the HOLODEC dataset from the ACE-ENA campaign, we identified the existence of these characteristic distributions for transects at different vertical levels for the research flight on July 18, 2017. These distributions perhaps could be thought of as analogous to basis sets of a coordinate system in linear algebra. We, however, make no effort in this paper to explore the physics behind their existence. Also whether the apparent success of the algorithm in exploring this dataset can be generalized will require looking at data from additional ACE-ENA flights and data gathered in other experiments. These more intriguing questions will be looked at in more detail in subsequent works.

Here we have introduced and illustrated the method developed for this purpose. The method employs standard statistical and machine learning tools. The two-sample Kolmogorov-Smirnov test is used for identifying statistically-similar droplet size distributions from pairs of holograms. It has the advantage of being free of any assumed distribution or binning of data. The DBSCAN algorithm, commonly used in the machine learning community, is adapted to identify clusters based on the results of the Kolmogorov-Smirnov hypothesis test. It has the advantage of not having a specified number of clusters or the requirement that the clusters are spatially connected.

Of particular interest is the nature of the identified clusters. We have emphasized that spatial correlation for the holograms in a cluster is not mandatory. Clusters identified for segments S1 and S5 have neighboring holograms, but also contain members that are spatially distant. These clusters also vary in prevalence, having just over ten holograms in some to the largest having over 500 holograms. No significant dependence is found for the clustering with altitude for this flight. One segment (S1) from the cloud top showed significant clustering, whereas the other segment (S2) was the only one with a single cluster. The cloud's mid regions (S3 & S4) had very similar distributions for most of the holograms, which formed the primary cluster and then also had a smaller subsidiary cluster. The cloud base region (S5) showed some periodic nature to one of its

TABLE 1. Information about the different cloud segments on which the algorithm is used. These segments are chosen from the research flight on July 18, 2017 from the ACE-ENA campaign.

Segment	No of holograms	Mean Altitude (m)	Std Altitude (m)	Mean droplet count	Std droplet count
S1	512	978.38	2.78	801.80	367.30
S2	1024	944.93	3.24	598.47	209.02
S3	1024	856.30	10.25	726.70	218.24
S4	1024	811.72	2.83	621.59	249.95
S5	1024	756.82	3.43	379.93	282.11

clusters. In all these cases, we were able to fit these distributions to gamma functions and generate the scale and shape parameters. There is a distinct difference in these parameters for different clusters, further highlighting the need to engage in this sort of analysis, rather than assuming that a cloud-averaged size distribution well characterizes the full cloud. For example, this variability has direct implications for autoconversion rates (Zhang et al. 2019).

The reliability and robustness of the algorithm are also verified using a synthetic dataset that mimics multiple elements of our field data. Synthetic data sets with three predefined clusters corresponding to different noise levels were generated; the algorithm successfully identified all three clusters in all cases. Some cases also identify additional clusters corresponding to the noise holograms with very similar parameters. Importantly, there was no case of misclassification of a cluster element to a different cluster. We take this as evidence that the algorithm appears reliable and is able to successfully complete an unsupervised classification of the hologram data.

In practice, this algorithm has much broader applicability and can be used to determine and classify the similarities between different data samples representable as CDFs. An application similar to this work, for which the algorithm might be appropriate, is to classify remotely sensed cloud droplet or rain size distributions. Similarly, the Doppler spectrum from each of a series of radar pulses would be a possible candidate for this type of cluster identification. A further example application would be a time series of spectral irradiance, from which each sample gives a distribution that could be converted to a CDF. Our experience is that the approach described here has the advantage of being free of imposed assumptions about distribution type or shape, and of finding clusters with relatively minor oversight from the user. It is therefore likely that its application could extend to problems outside the scope of the atmospheric and climate sciences.

TABLE 2. The results after implementation of the algorithm on the data from various cloud segments. The corresponding input parameters are also included. Additionally, the fitted generalized gamma parameters for each of the identified clusters are presented.

Segment	DBSCAN parameters: Epsilon - 0.1	Cluster Properties			Mean Gamma Parameters	
	Min-points	Clusters	Cluster size	Noise	Shape parameter	Scale parameter (μm)
S1	10	7	77, 12,	171	39.82, 36.91,	0.45, 0.46,
			67, 12,		32.30, 55.19,	0.63, 0.39,
			38, 10,		52.95, 47.04,	0.42, 0.49,
			11		30.18	0.72
S2	10	1	792	51	37.93	0.63
S3	15	2	715,18	120	57.91, 56.38	0.41, 0.45
S4	40	2	362,50	373	36.41, 42.23	0.50, 0.46
S5	15	4	255, 48,	168	45.01, 22.45,	0.34, 0.76,
			116,17		42.45, 45.52	0.51, 0.49

TABLE 3. Details of the synthetic dataset to check the efficacy of the algorithm. It includes information about the pre-defined clusters and its comparison to the clusters identified by the algorithm.

Segment	Pre-defined Cluster Properties		Number of clusters identified	DBSCAN clusters: Fraction from		
	Cluster elements	Noise holograms		Original cluster	Noise holograms	Other clusters
SD1	664,310,34	16	3	0.993	0.007	0
				1	0	0
				0.903	0.097	0
SD2	518,274,86	146	5	0.946	0.054	0
				0.965	0.035	0
				0.914	0.086	0
				0	1	0
				0	1	0
SD3	506,114,47	357	3	0.776	0.224	0
				1	0	0
				0.366	0.644	0

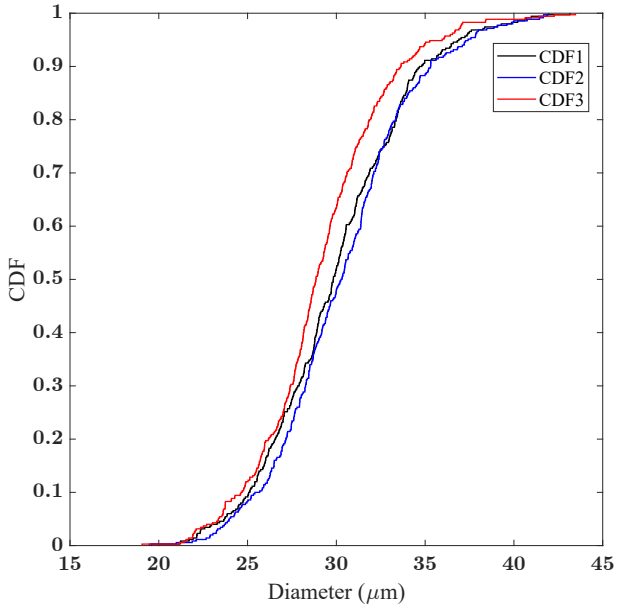


FIG. 1. Cumulative distribution functions (CDFs) of three diameter distributions. Distributions 1 and 2 are similar enough to pass the KS test while the distribution 3 is different from distrbutions 1 and 2. The corresponding KS distances for the 1/2, 1/3 and 2/3 pairs are 0.0686, 0.1371 and 0.1714 respectively.

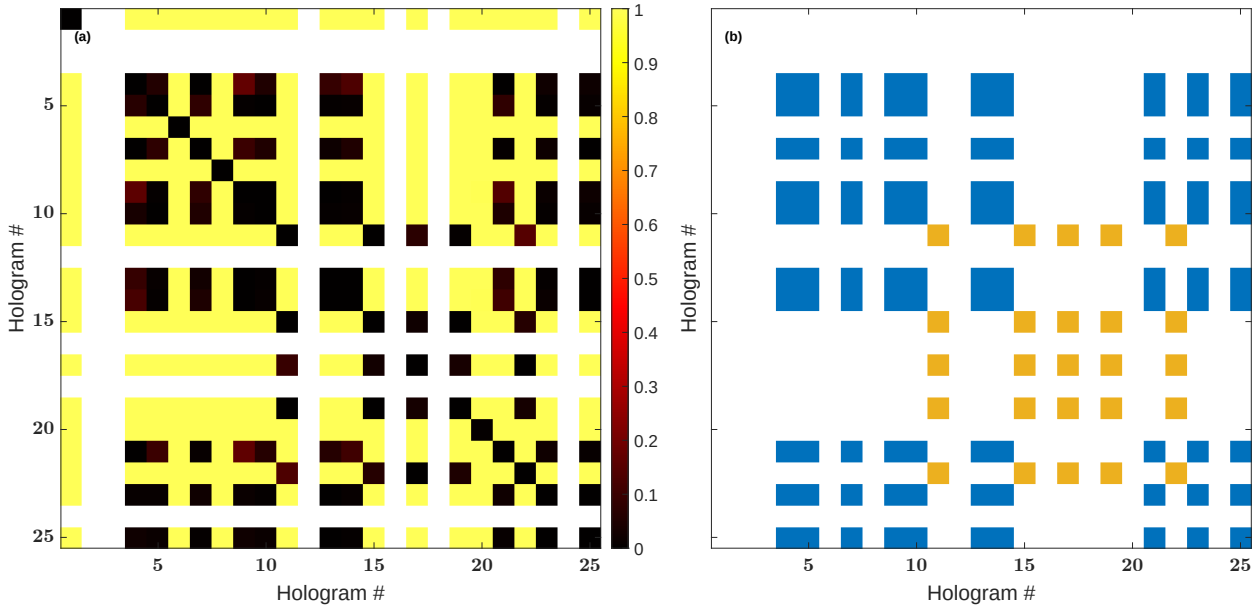


FIG. 2. (a) Illustration of a KS Matrix from 25 synthetic holograms. The cell values range from 0 to 1. A value close to 0 (black) indicates that the size distributions are largely indistinguishable and the values close to 1 (yellow) means they are clearly different. The holograms below the cutoff limit are whitened. Note that the diagonal of the matrix shows values close to zero as expected. (b) The clusters identified using DBSCAN. The clusters are depicted by different colours. Here blue and yellow represent two clusters of sizes 10 and 5 holograms respectively

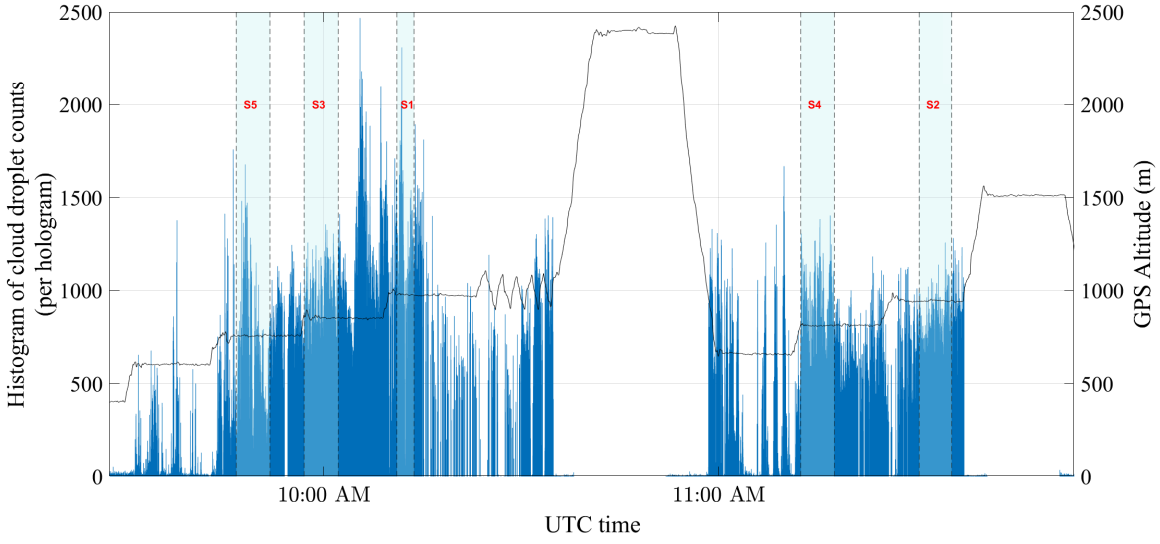
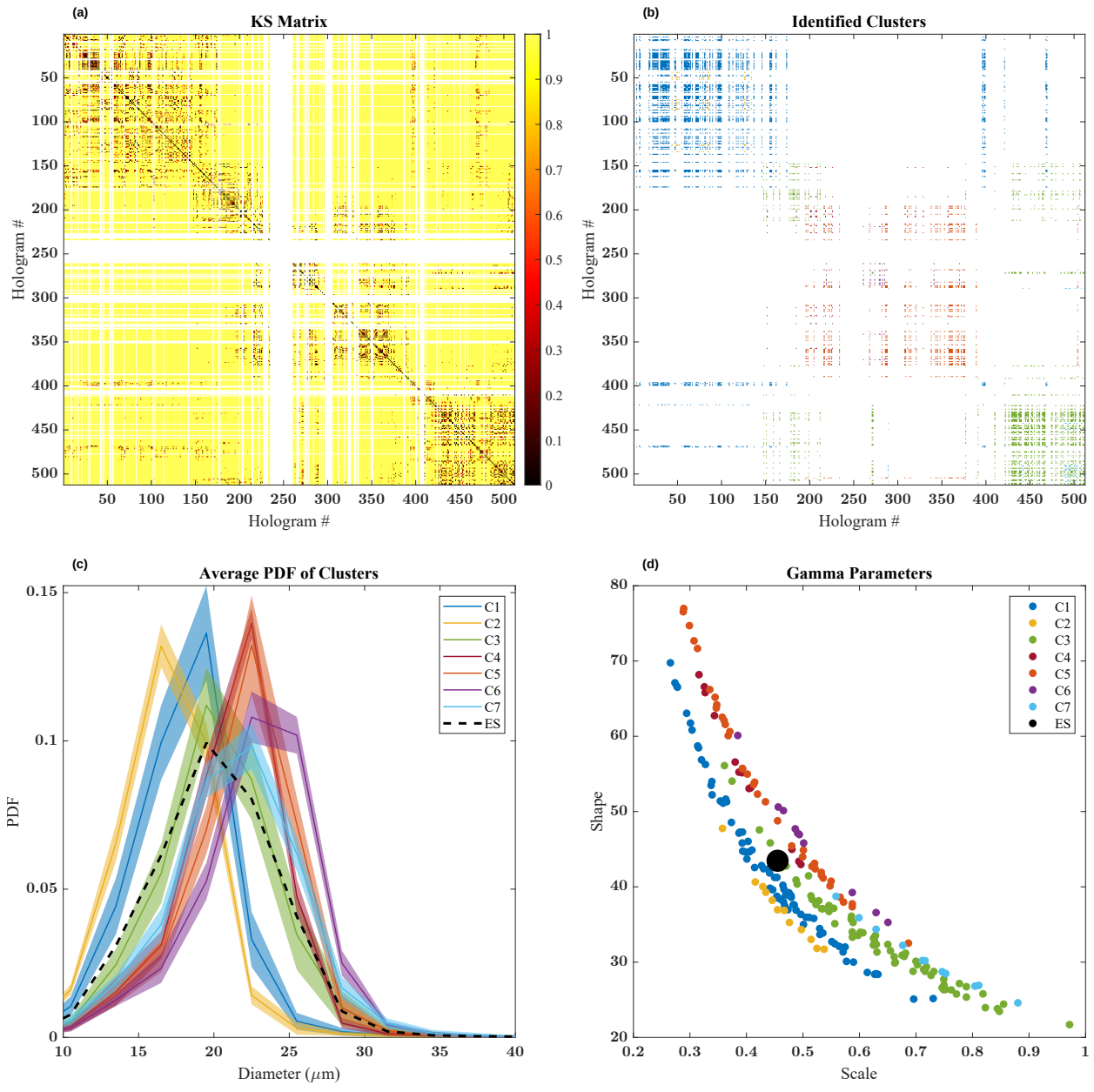
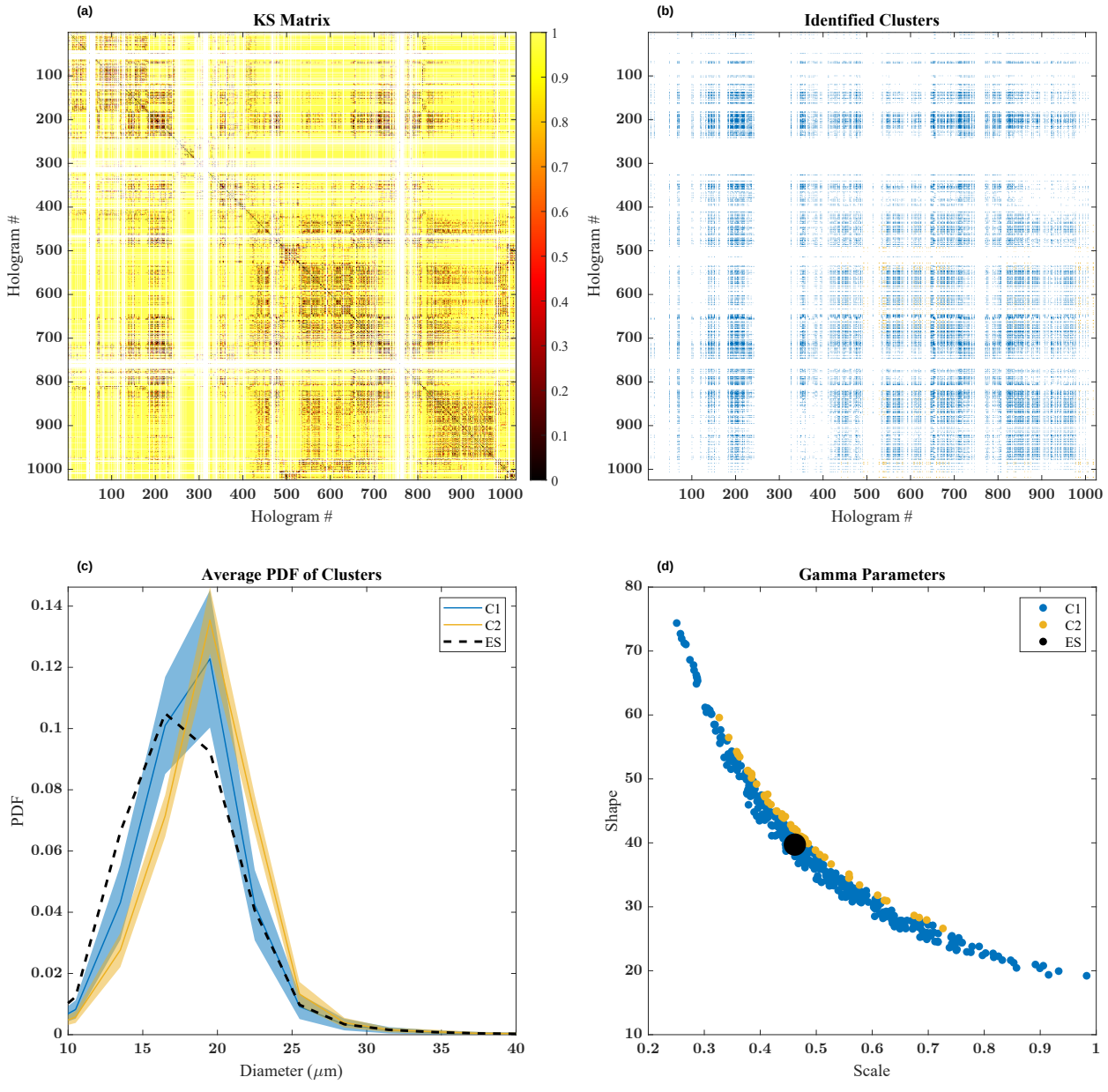


FIG. 3. (left) Time series of the cloud droplets histogram counts (per hologram) from the research flight on July 18, 2017 (IOP 1 RF18). The different flight segments (S1,S2,S3,S4 & S5) are shown in shaded regions. (right) The measured onboard flight altitude for the G1 aircraft for the same flight.



498 FIG. 4. Results from Segment S1 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm.
499 The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion
500 represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms
501 above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for the holograms
502 in different clusters. The large black dot gives the shape and size parameter for the entire segment.



503 FIG. 5. Results from Segment S4 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm.
504 The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion
505 represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms
506 above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for the holograms
507 in different clusters. The large black dot gives the shape and size parameter for the entire segment.

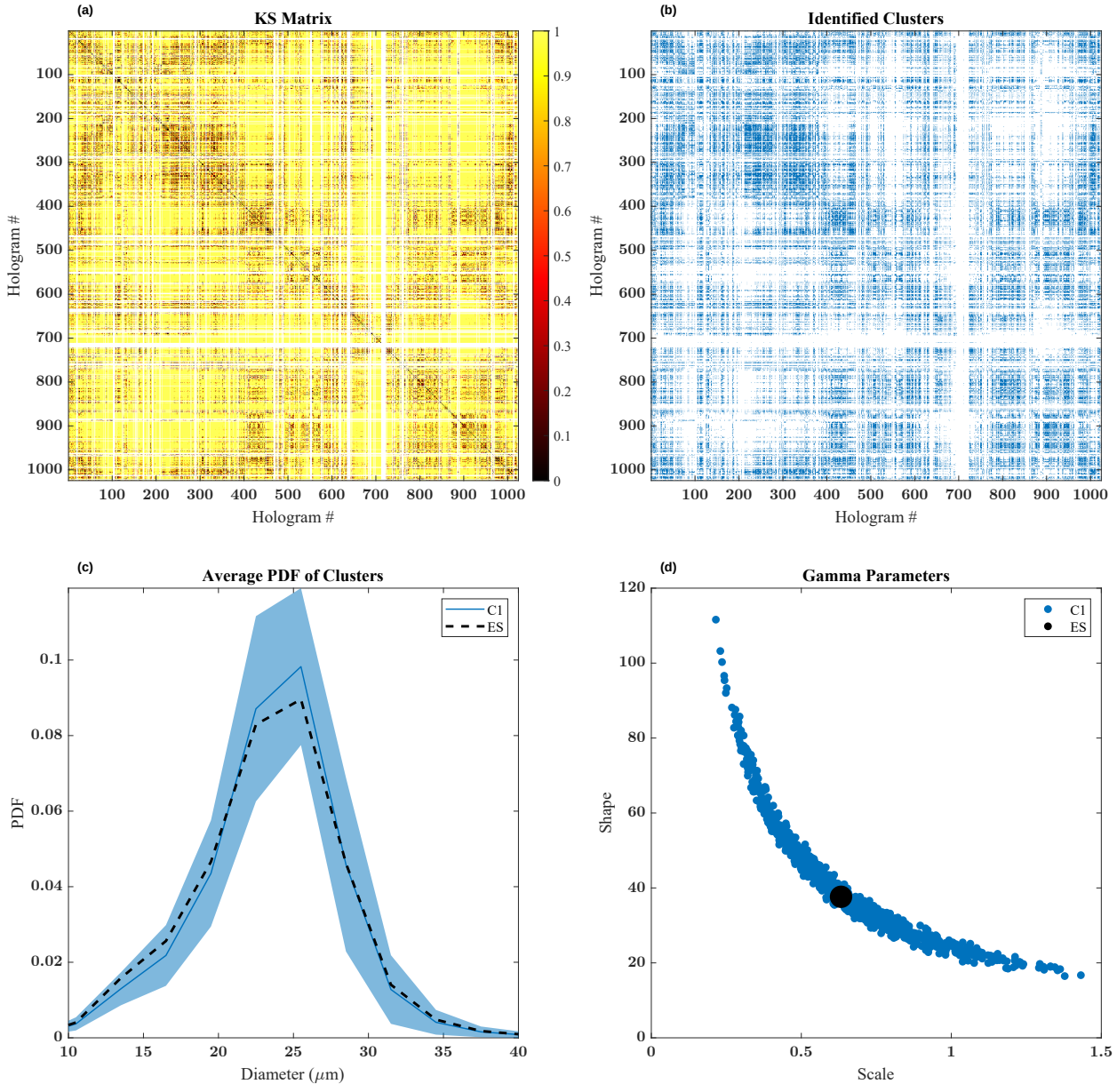
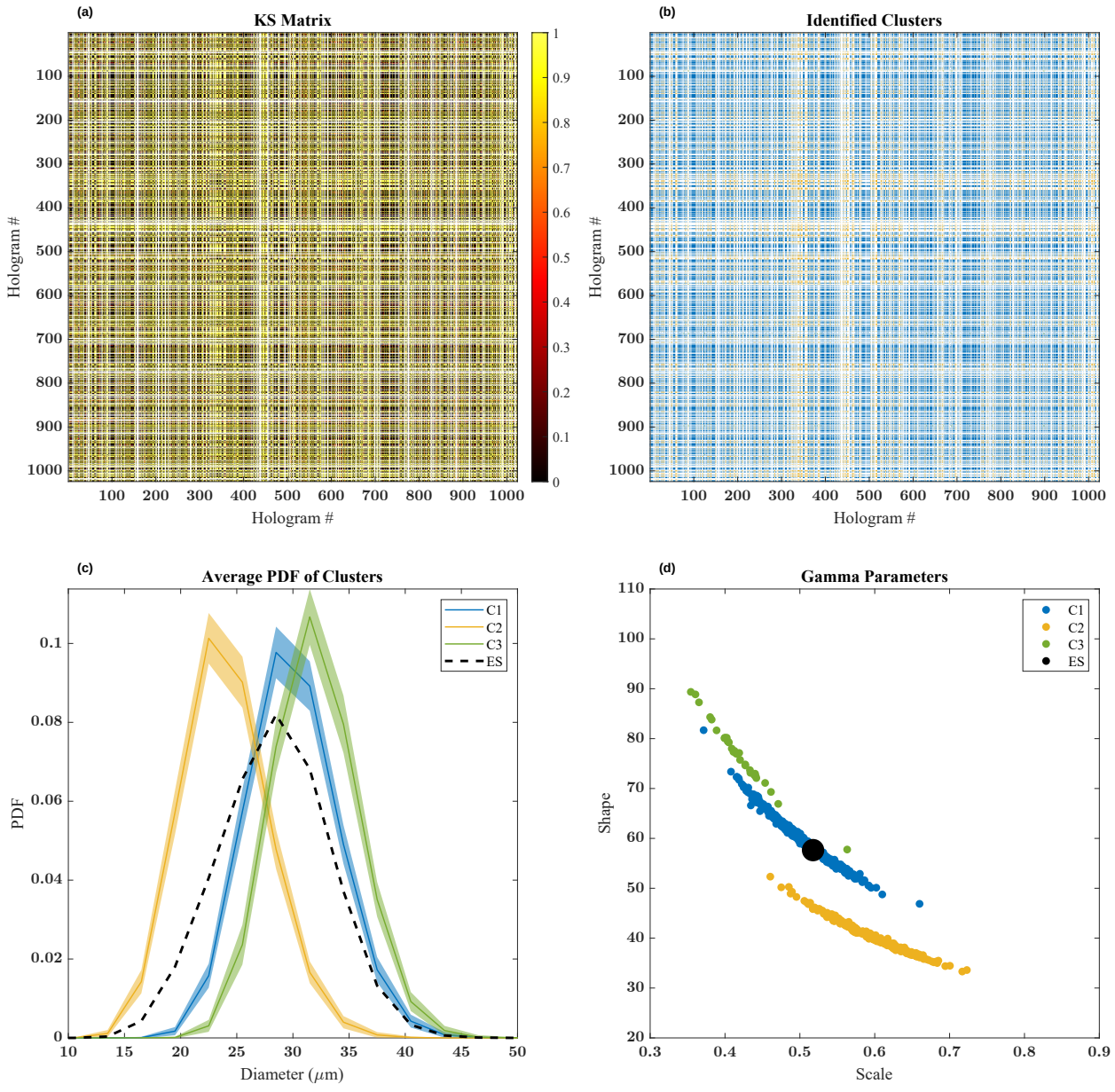


FIG. 6. Results from Segment S2 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for the holograms in different clusters. The large black dot gives the shape and size parameter for the entire segment.



513 FIG. 7. Results from synthetic data SD1 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm.
514 The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion
515 represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms
516 above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for the holograms
517 in different clusters. The large black dot gives the shape and size parameter for the entire segment.

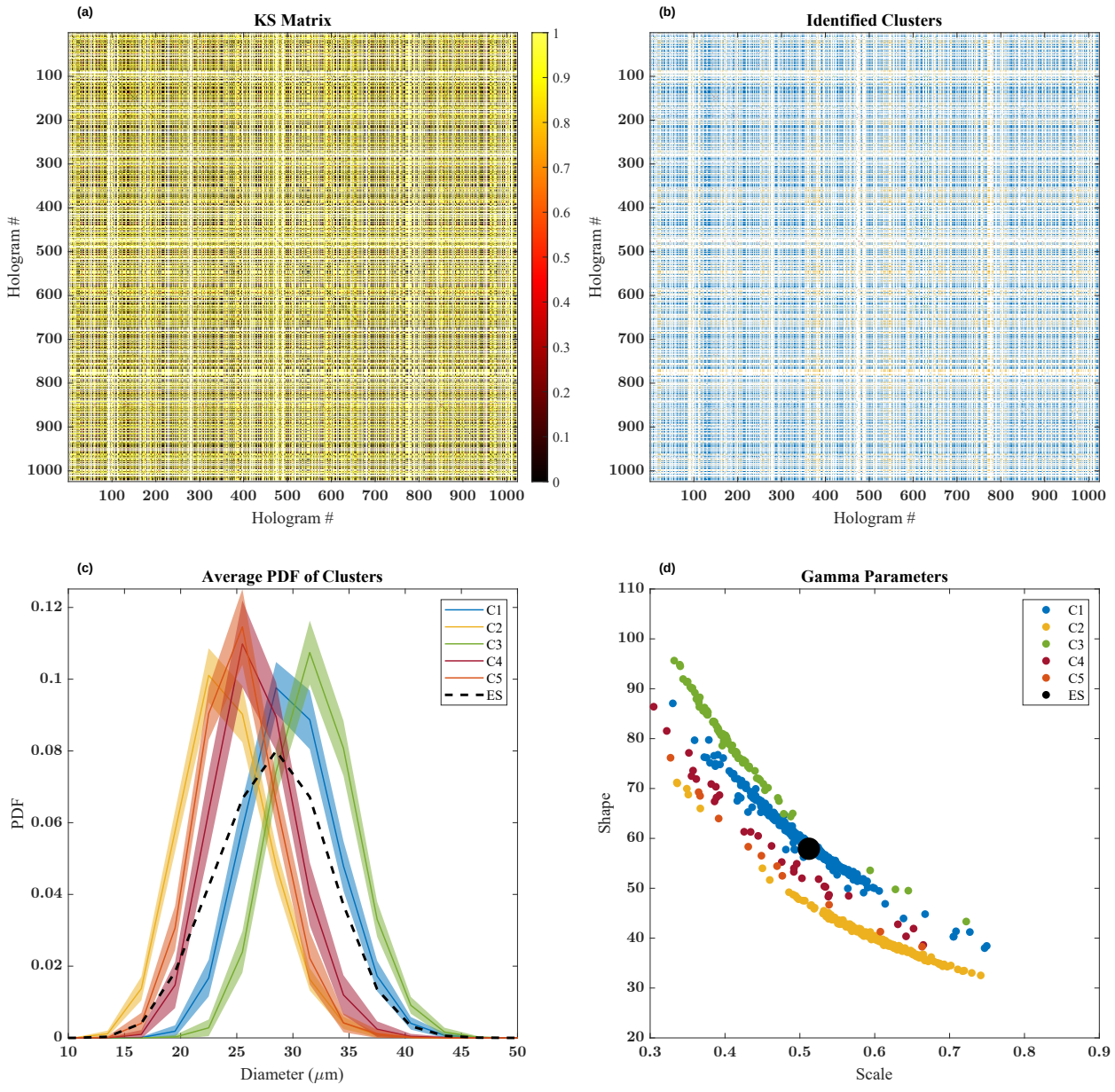


FIG. 8. Results from synthetic data SD2 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for different clusters. The large black dot gives the shape and size parameter for the entire segment

Acknowledgments. This work was supported by U.S. Department of Energy Office of Science Award DE-SC0020053 and through National Science Foundation award AGS-2001490. We thank Dr. Susanne Glienke and the staff of the ARM Aerial Facility for their role in obtaining the data during the ACE-ENA project.

Data availability statement. The HOLODEC data used in this study can be downloaded from "https://www.arm.gov/research/campaigns/aaf2017ace-en".

References

Barlow, R. J., 1993: *Statistics: a guide to the use of statistical methods in the physical sciences*, Vol. 29. John Wiley & Sons.

Beals, M. J., J. P. Fugal, R. A. Shaw, J. Lu, S. M. Spuler, and J. L. Stith, 2015: Holographic measurements of inhomogeneous cloud mixing at the centimeter scale. *Science*, **350 (6256)**, 87–90.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, and Coauthors, 1996: A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, Vol. 96, 226–231.

Fugal, J., and R. Shaw, 2009: Cloud particle size distributions measured with an airborne digital in-line holographic instrument. *Atmospheric Measurement Techniques*, **2 (1)**, 259–271.

Fugal, J. P., T. J. Schulz, and R. A. Shaw, 2009: Practical methods for automated reconstruction and characterization of particles in digital in-line holograms. *Measurement Science and Technology*, **20 (7)**, 075 501.

Fugal, J. P., R. A. Shaw, E. W. Saw, and A. V. Sergeyev, 2004: Airborne digital holographic system for cloud particle measurements. *Applied optics*, **43 (32)**, 5987–5995.

Gan, G., C. Ma, and J. Wu, 2020: *Data clustering: theory, algorithms, and applications*. SIAM.

Glienke, S., A. B. Kostinski, R. A. Shaw, M. L. Larsen, J. P. Fugal, O. Schlenczek, and S. Bormann, 2020: Holographic observations of centimeter-scale nonuniformities within marine stratocumulus clouds. *Journal of the Atmospheric Sciences*, **77 (2)**, 499–512.

548 Hahn, C. J., and S. G. Warren, 2007: *A gridded climatology of clouds over land (1971-96) and*
549 *ocean (1954-97) from surface observations worldwide*. Oak Ridge National Laboratory, Carbon
550 Dioxide Information Analysis Center . . . .

551 Hartmann, D. L., M. E. Ockert-Bell, and M. L. Michelsen, 1992: The effect of cloud type on
552 earth's energy balance: Global analysis. *Journal of Climate*, **5 (11)**, 1281–1304.

553 Igel, A. L., and S. C. van den Heever, 2017: The importance of the shape of cloud droplet size
554 distributions in shallow cumulus clouds. part ii: Bulk microphysics simulations. *Journal of the*
555 *Atmospheric Sciences*, **74 (1)**, 259–273.

556 Jaffrain, J., and A. Berne, 2011: Experimental quantification of the sampling uncertainty associated
557 with measurements from parsivel disdrometers. *Journal of Hydrometeorology*, **12**, 352–370.

558 Jameson, A., and A. Kostinski, 2000: Fluctuation properties of precipitation. part vi: Observations
559 of hyperfine clustering and drop size distribution structures in three-dimensional rain. *Journal*
560 *of the Atmospheric Sciences*, **57**, 373–388.

561 Jameson, A., M. Larsen, and A. Kostinski, 2015a: Disdrometer network observations of finescale
562 spatial-temporal clustering in rain. *Journal of the Atmospheric Sciences*, **72**, 1648–1666.

563 Jameson, A., M. Larsen, and A. Kostinski, 2015b: On the variability of drop size distributions
564 over areas. *Journal of the Atmospheric Sciences*, **72**, 1386–1397.

565 Jameson, A., M. Larsen, and A. Kostinski, 2018: On the detection of statistical heterogeneity in
566 rain measurements. *Journal of Atmospheric and Oceanic Technology*, **35**, 1399–1413.

567 Kendall, M., and A. Stuart, 1979: *The Advanced Theory of Statistics*, Vol. Vol. 2. Macmillan, 723
568 pp.

569 Larsen, M., A. Kostinski, and A. Tokay, 2005: Observations and analysis of uncorrelated rain.
570 *Journal of the Atmospheric Sciences*, **62**, 4071–4083.

571 Larsen, M., and K. O'Dell, 2016: Sampling variability effects in drop-resolving disdrometer
572 observations. **121**, 11 777–11 791.

Larsen, M., R. Shaw, A. Kostinski, and S. Glienke, 2018: Fine-scale droplet clustering in atmospheric clouds: 3d radial distribution function from airborne digital holography. *Physical Review Letters*, **121**, 204 501.

Miles, N. L., J. Verlinde, and E. E. Clothiaux, 2000: Cloud droplet size distributions in low-level stratiform clouds. *Journal of the atmospheric sciences*, **57 (2)**, 295–311.

Shaw, M. A., 2016: Testing lidar-radar derived drop sizes against in situ measurements. M.S. thesis, Michigan Technological University.

Slingo, A., 1990: Sensitivity of the earth's radiation budget to changes in low clouds. *Nature*, **343 (6253)**, 49–51.

Spuler, S. M., and J. Fugal, 2011: Design of an in-line, digital holographic imaging system for airborne measurement of clouds. *Applied optics*, **50 (10)**, 1405–1412.

Stephens, G. L., 2005: Cloud feedbacks in the climate system: A critical review. *Journal of climate*, **18 (2)**, 237–273.

Straka, J. M., 2009: *Cloud and precipitation microphysics: principles and parameterizations*. Cambridge University Press.

Wang, J., and Coauthors, 2021: Aerosol and cloud experiments in the eastern north atlantic (ace-ena). *Bulletin of the American Meteorological Society*, 1–51.

Zhang, Z., H. Song, P.-L. Ma, V. E. Larson, M. Wang, X. Dong, and J. Wang, 2019: Subgrid variations of the cloud water and droplet number concentration over the tropical ocean: satellite observations and implications for warm rain simulations in climate models. *Atmospheric Chemistry and Physics*, **19 (2)**, 1077–1096.

Automated identification of characteristic droplet size distributions in stratocumulus clouds utilizing a data clustering algorithm - Supplement

Nithin Allwayin, [a] , Michael L. Larsen , [a,b] Alexander G. Shaw, [c] and Raymond A. Shaw [a]

[a] *Michigan Technological University, Houghton, Michigan*

[b] *College of Charleston, Charleston, South Carolina*

[c] *Brigham Young University, Provo, Utah*

Corresponding author: Michael L. Larsen, LarsenML@cofc.edu

Corresponding author: Raymond A. Shaw, rashaw@mtu.edu

1. Selection of DBSCAN input parameters: Additional Information

DBSCAN input parameters are selected to maximize the number of clusters in each segment. As discussed in the main paper, the results are less sensitive to "epsilon" values and hence it is chosen to be a constant in our analysis. The "min points" values are iterated in steps of 5 for an "epsilon" value of 0.1. The lower cutoff value for "min points" is set as 10 in the paper. For more complete illustration, we plot the dependence of the number of identified clusters on the minimum number of points in a cluster in Figure 1. This demonstrates that even for smaller "min points", the algorithm identifies finite number of characteristic distributions.

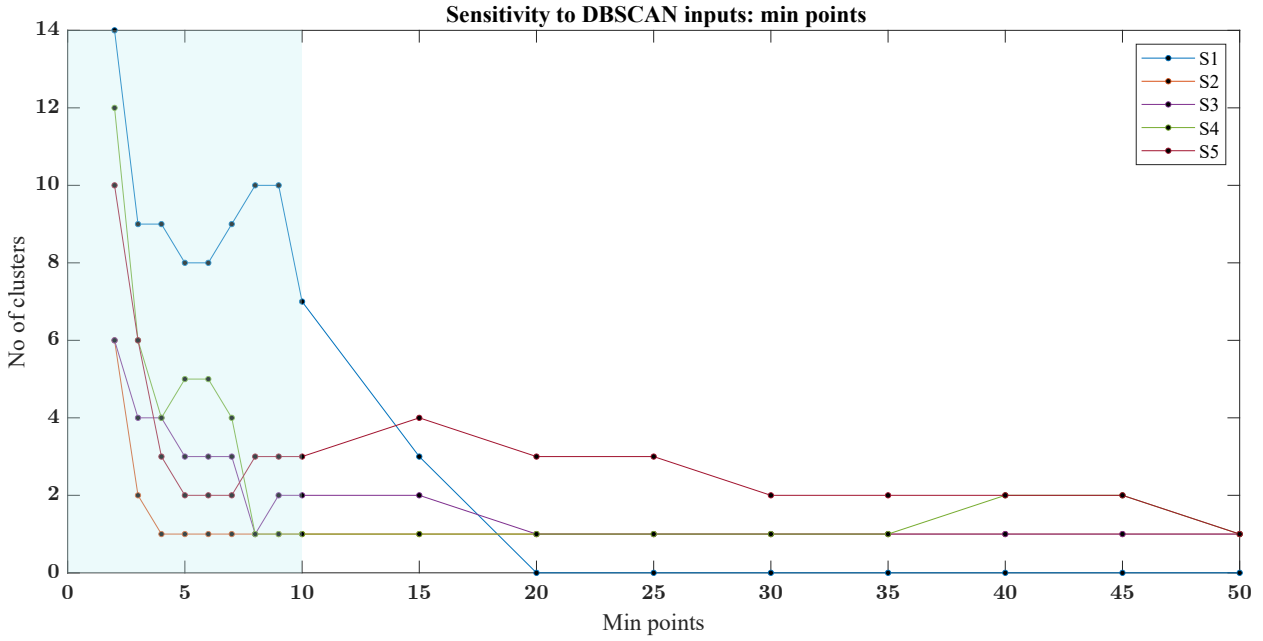


FIG. S 1. Sensitivity of the DBSCAN results to the input parameter: "min points". The number of clusters for the different cloud segments are shown. Here the value of "epsilon" is fixed to be 0.1.

2. Results for segment S3 & S5

Two clusters are identified for the segment S3. This segment is from the mid cloud region. The larger cluster has 715 members and the smaller one has 18. The results are shown in Figure S 2. Segment S5 is the transect near the cloud base. Four clusters with sizes 255,48,116 and 17 are identified for this segment and are shown in Figure S 3.

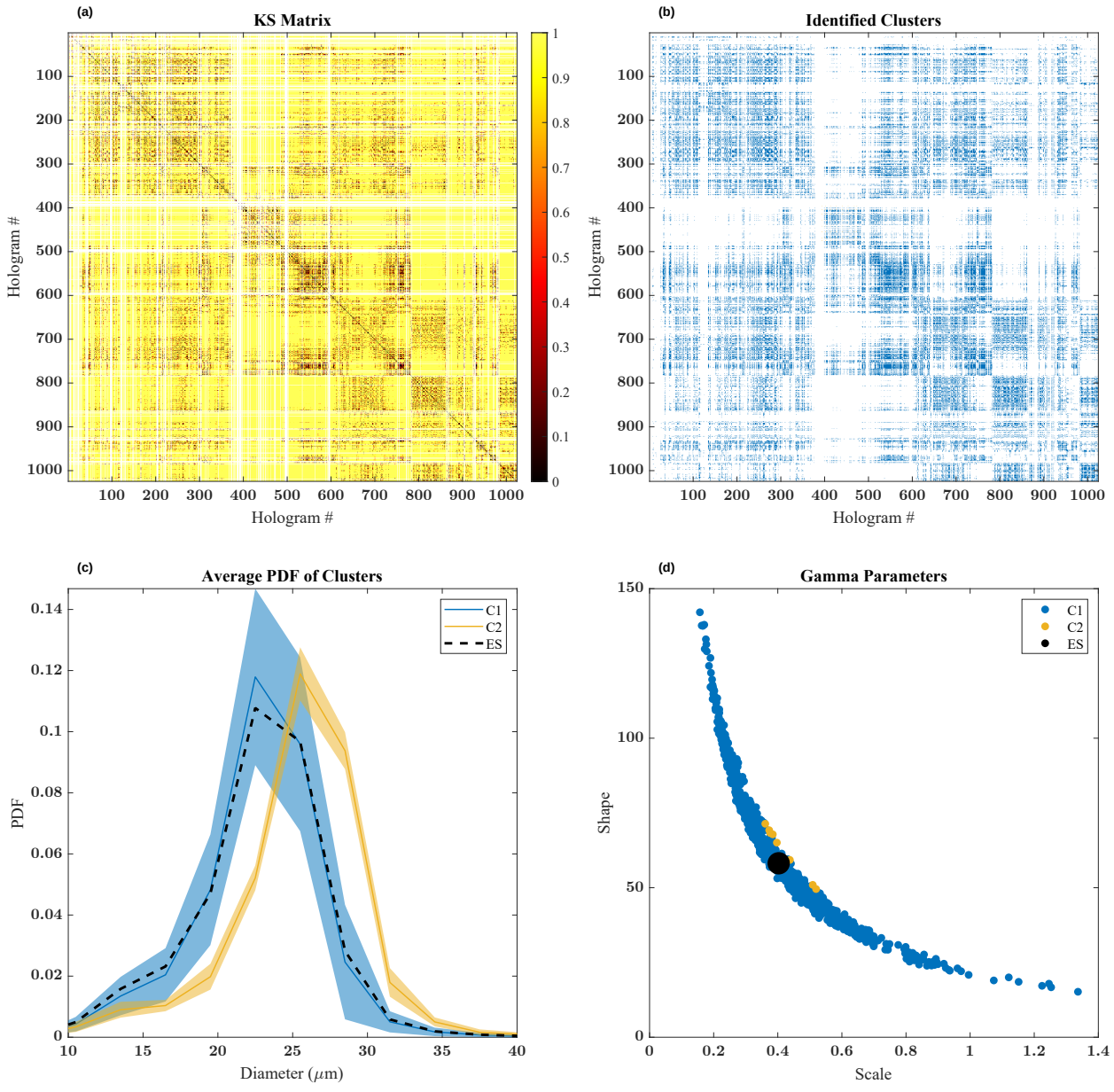


FIG. S 2. Results from Segment S3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for different clusters. The large black dot gives the shape and size parameter for the entire segment.

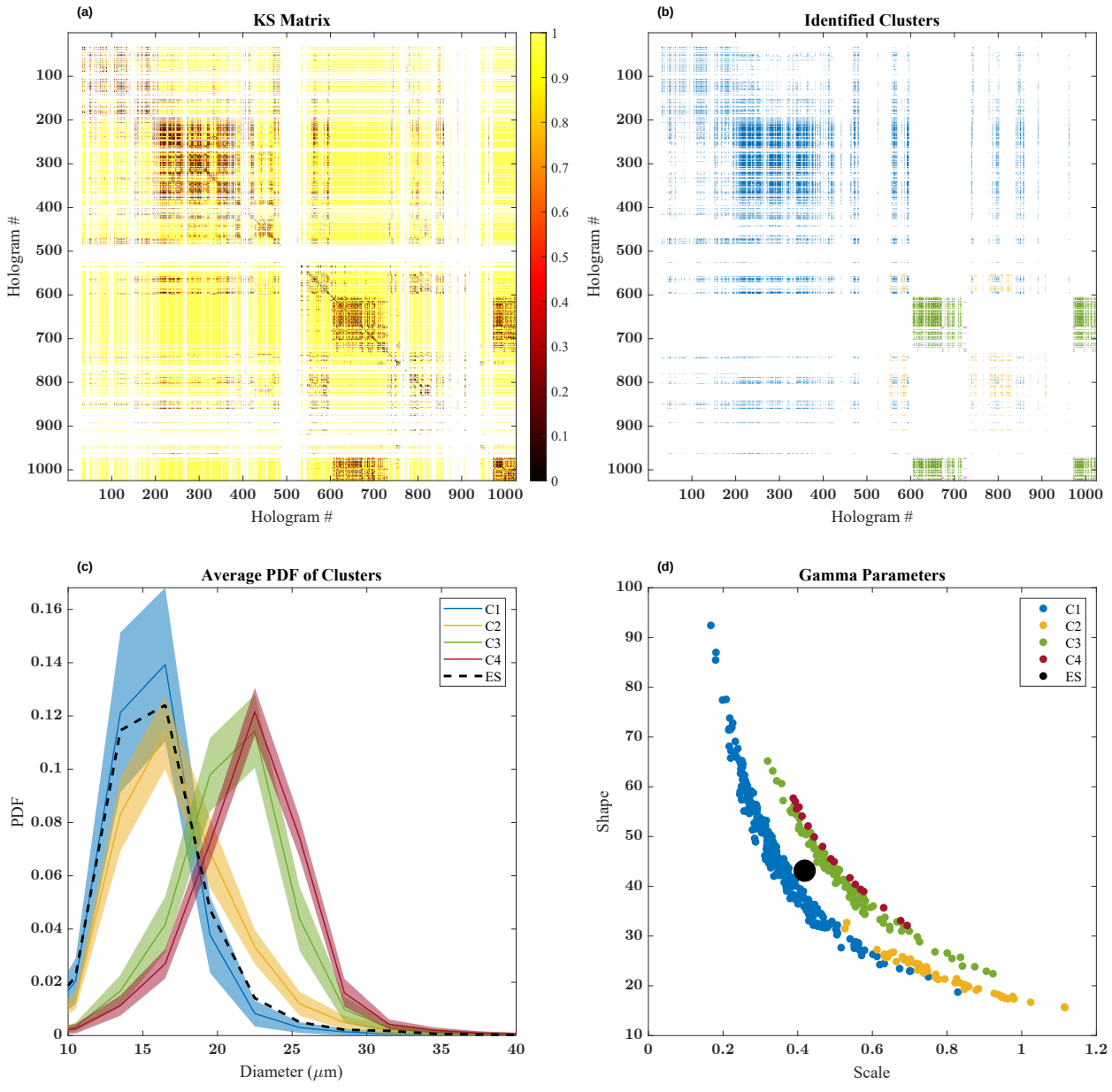


FIG. S 3. Results from Segment S5 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for different clusters. The large black dot gives the shape and size parameter for the entire segment.

3. Results of synthetic holograms-Segment SD3

The results for the set of synthetic holograms labelled SD3 is illustrated in Figure S 4. This set is one with the largest noise content among the three synthetic data sets.

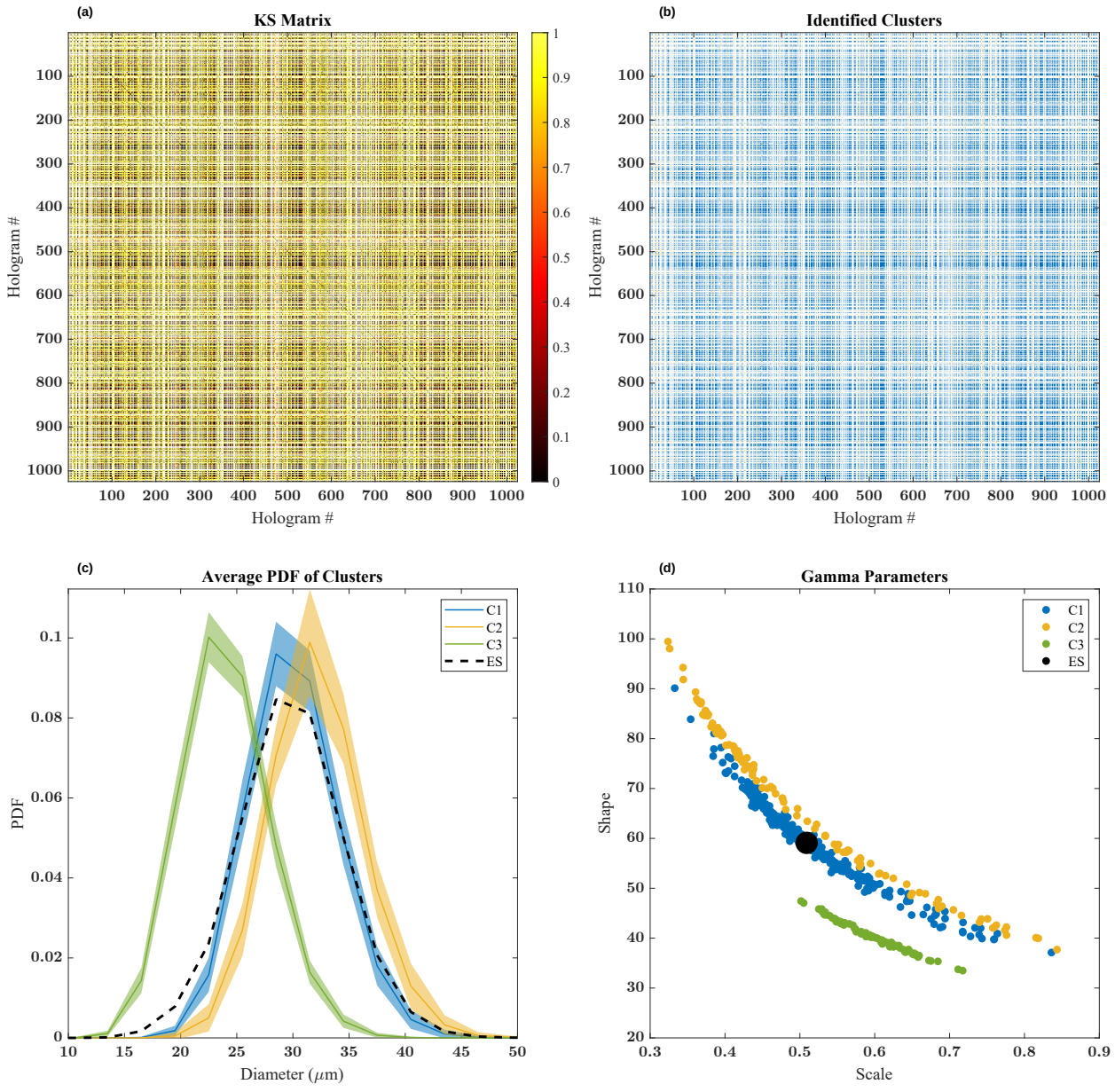


FIG. S 4. Results for the synthetic hologram set: SD3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and size parameters of the modified gamma distribution for different clusters. The large black dot gives the shape and size parameter for the entire segment.