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*

(Manuscript received 11 January 2022, in final form 3 May 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 class of machine learning algorithms for identifying clusters of samples with similar properties: density-based spatial clustering of applications with noise (DBSCAN) is used as the specific example for illustration. 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-type algorithms 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 the 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." Droplets in a cloud interact locally, however, 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 are contained in a given cloud. This is important because the size distribution describes processes such as cloud droplet growth and light transmission through clouds.

KEYWORDS: Cloud droplets; Cloud microphysics; In situ atmospheric observations; Machine learning; Microscale processes/variability

1. Introduction

A considerable portion of Earth's oceans are 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 on the basis of 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 toward 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 to achieve convergence to a global distribution? More intriguing, might a seemingly homogeneous cloud be described by a small number of clearly distinguishable, characteristic droplet size distributions throughout its interior?

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-22-0003.s1.

Corresponding author: Michael L. Larsen, larsenml@cofc.edu; Raymond A. Shaw, rashaw@mtu.edu

DOI: 10.1175/AIES-D-22-0003.1 e220003

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, despite 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 semiparametric algorithm described here allows for the exploration of any in situ data having spatially tagged information about particle (in this case cloud drop) detections. Specifically, we use data captured by the Holographic Detector for Clouds (HOLODEC) instrument (Fugal et al. 2004; Fugal and Shaw 2009; Spuler and Fugal 2011) during the Aerosol and Cloud Experiments in the Eastern North Atlantic (ACE-ENA) field project (Wang et al. 2022). 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 ($\approx$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 a class of data clustering algorithms that does not require number of clusters as an input parameter. Specifically, the results are based on "density-based spatial clustering of applications with noise" (DBSCAN), but the method can be extended to related approaches like "ordering points to identify the clustering structure" (OPTICS). 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 about the functional form of the cloud droplet size distribution. The 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 paper 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 discusses overarching results (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 KS test, and the grouping is done with the density-based clustering algorithm; for the main results we use DBSCAN, and in the discussion section we show the extension to a related clustering algorithm OPTICS.

a. Kolmogorov–Smirnov test

The KS test is a nonparametric 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

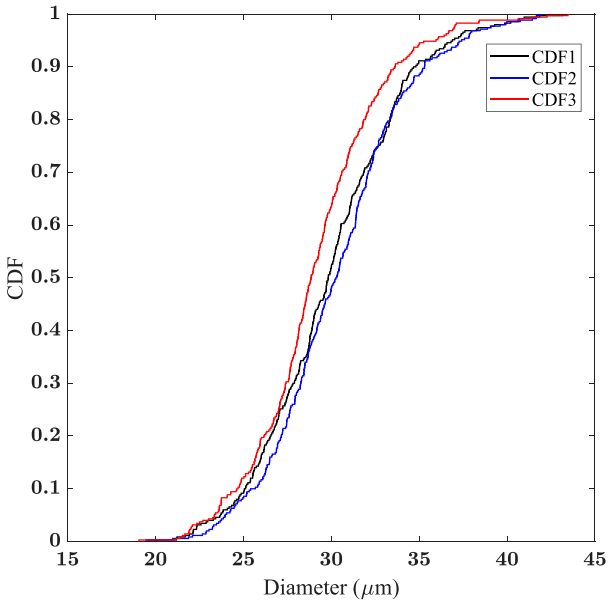

FIG. 1. Cumulative distribution functions of three diameter distributions. Distributions 1 and 2 are similar enough to pass the KS test, whereas distribution 3 is different from distributions 1 and 2. The corresponding KS distances for the 1–2, 1–3, and 2–3 pairs are 0.0686, 0.1371, and 0.1714, respectively.

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 Fig. 1, where we see the CDFs of droplet size distributions from three holograms. The KS test compares 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 holograms 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 with 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 Fig. 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% 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 intercomparisons between the empirical hologram drop size distributions sampled to the cutoff value. A representative histogram of these n^2 outputs bounded between 0 and 1 is shown in section 7 of the online supplemental material.

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 Fig. 2a 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 preassigned parent distributions. The ensemble size for the subsamples is 1000, and thus each cell in this matrix is from one-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 suboptimal and nonobjective to detect all such clusters visually; they can be better classified using unsupervised algorithmic clustering techniques.

b. *DBSCAN*

To identify clusters within our droplet-size distribution measurements, we implement the DBSCAN algorithm. Many popular clustering algorithms rely on the user to specify the number of clusters as an input parameter. For our data, preassignment 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,

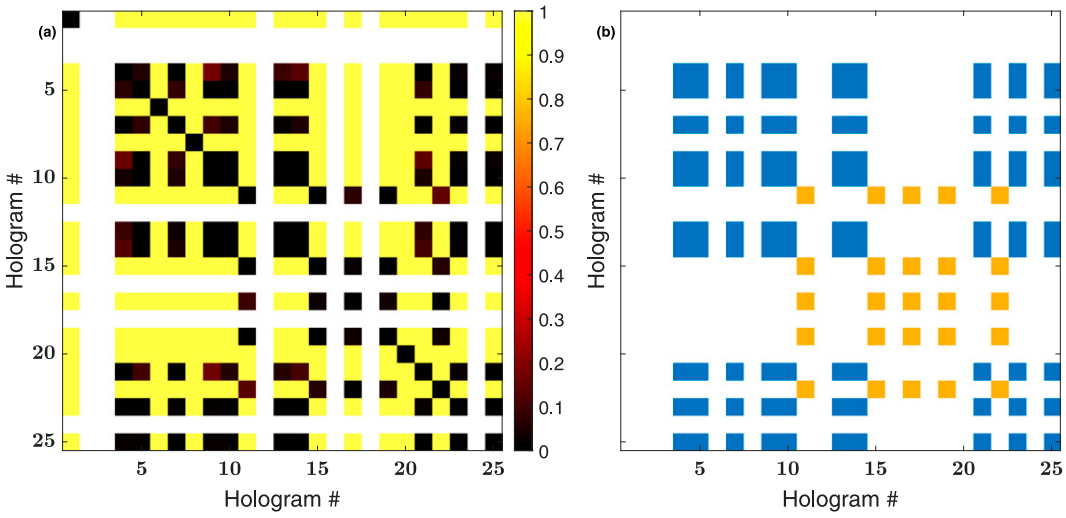

FIG. 2. (a) Illustration of a KS matrix from 25 synthetic holograms. The cell values range from 0 to 1. A value close to 0 (black) indicates that the size distributions are largely indistinguishable, and a value close to 1 (yellow) means they are clearly different. The holograms below the cutoff limit 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 colors. Here blue and yellow represent two clusters of sizes 10 and 5 holograms, respectively.

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 nonparametric; 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 (i.e., 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 fill 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 section 2d. Figure 2b 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% 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. 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% 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 of 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 dataset that meets the cutoff criteria as the primary hologram. It is then repeatedly subsampled 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 with 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 nonnumerical placeholder. 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 do not meet the cutoff have all n entries represented by a nonnumerical 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 nonnumerical 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 to 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% of the holograms in the flight transect. More details on implementation of the algorithm with different input parameters are included in the online supplemental material.

3. Using the algorithm with real data

a. Dataset: ACE-ENA

The dataset used in our study is derived from the HOLODEC deployment during the ACE-ENA campaign (Wang et al. 2022). 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

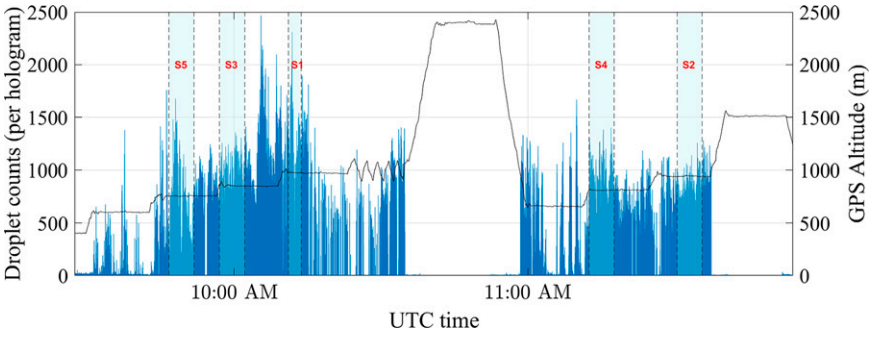

FIG. 3. Time series of the cloud droplets histogram counts (per hologram) (blue; left *y* axis) from the research flight on 18 Jul 2017 (IOP 1 RF18). The different flight segments (S1, S2, S3, S4, and S5) are shown as the shaded regions. Also shown is the measured onboard flight altitude for the G1 aircraft for the same flight (black line; right *y* axis).

board the G1 aircraft, operated as part of the Atmospheric Radiation Measurement (ARM) Aerial Facility, were used to make various measurements of the cloud and the boundary layer over two intensive operational periods (IOPs). The G1 flight moved at an approximate speed of 100 m s^{-1}. As HOLODEC has a hologram acquisition frequency of 3.3 Hz, this would mean that the holograms obtained are about 30 m apart. Therefore, a set of such holograms would capture the local variability across a large section of the cloud. More details about the instrument can be found in section 5 of the online supplemental material. In our analysis, we choose data from research flight on 18 July 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 and S2), two in the mid-cloud regions (S3 and S4), and one near the base of the cloud (S5). These segments and the corresponding droplet count histograms are shown in Fig. 3. They correspond to altitudes of approximately 950, 850, and 750 m, respectively. These segments contain lengths along the crosswind (S1, S4, and S5) and alongwind (S3) directions, along with a mix of along- and crosswind (S2) directions. Detailed information about the different segments is 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 seven 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 S1 is shown in Fig. 4. The algorithm identifies 7 clusters for the chosen parameters. Of these 7, 3 clusters have over 35 holograms, while the other smaller clusters have around 10 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 Fig. 5. Only a single cluster is found for segment S2, as seen in Fig. 6, indicating that most of the holograms in this segment are similar to each other. In both these cases, the PDF 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 and S5, respectively, the results of which can be found in more detail in section 2 of the supplemental material. 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

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

Segment	No. of holograms	Mean alt (m)	Std alt (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

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

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

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 4c, 5c, and 6c illustrate this clearly. These can be compared with 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

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

where d is the diameter of the droplets. The scale parameter θ has dimensions of length. The shape parameter k is nondimensional 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

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

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 Figs. 4d, 5d, and 6d. 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 Figs. 4d, 5d, and 6d. There is a distinction between the gamma parameters for the different clusters, which might be difficult to infer from a scatterplot 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 in diameter. This means that the mode diameter for our data will be larger and hence can help to 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 about 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, and 80 μm and 0.6, 0.5, and 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, because 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 of the holograms in the transect are assigned either as 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

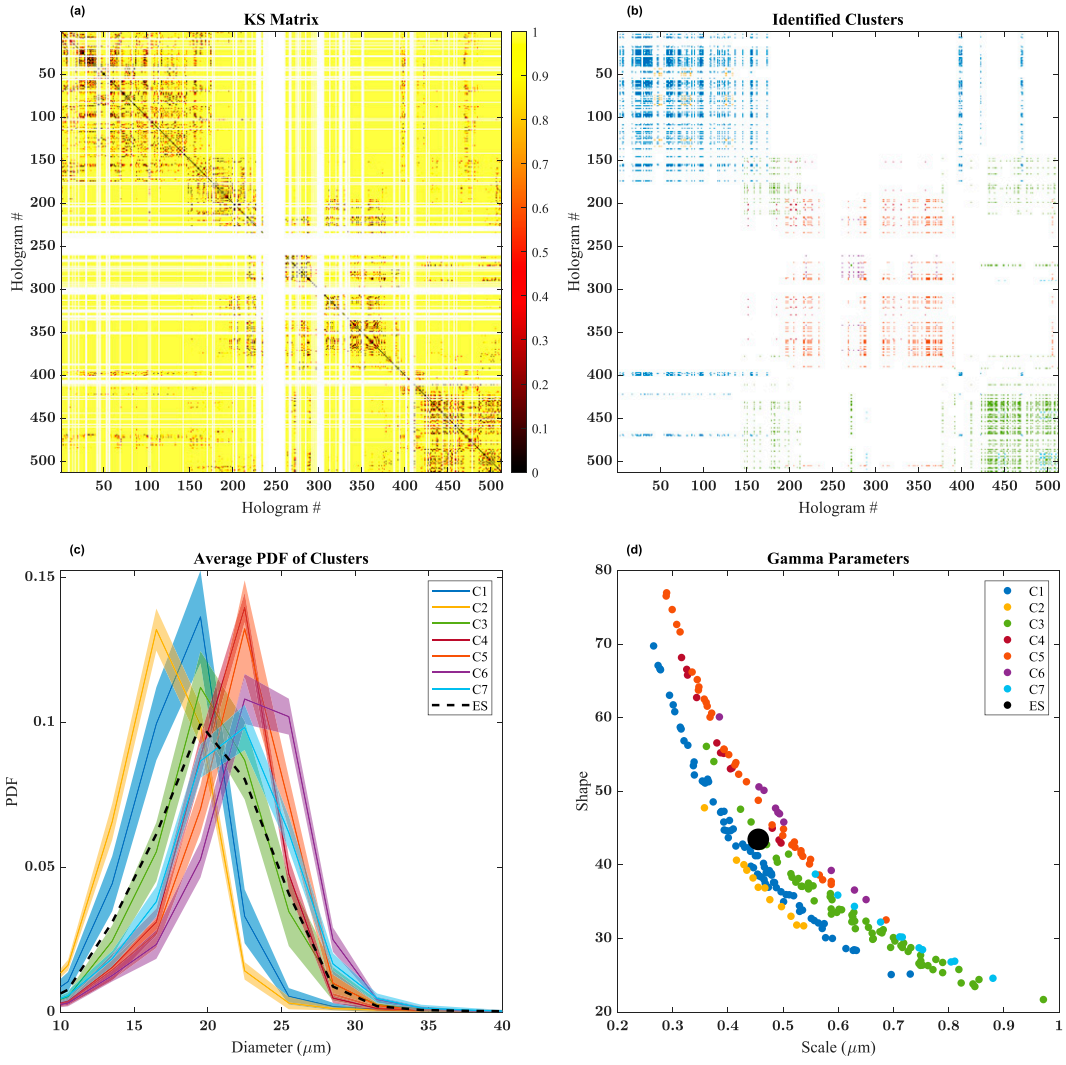

FIG. 4. Results from segment S1: (a) KS matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colors. (c) Average PDFs of the different clusters. The shaded portion represents 1 standard deviation. The dashed black line shows the PDF for the entire segment of holograms above the cutoff. (d) The fitted shape and scale parameters of the modified gamma distribution for the holograms in different clusters. The large black dot gives the shape and scale parameter for the entire segment.

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 dataset 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 μm 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 dataset. 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. 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

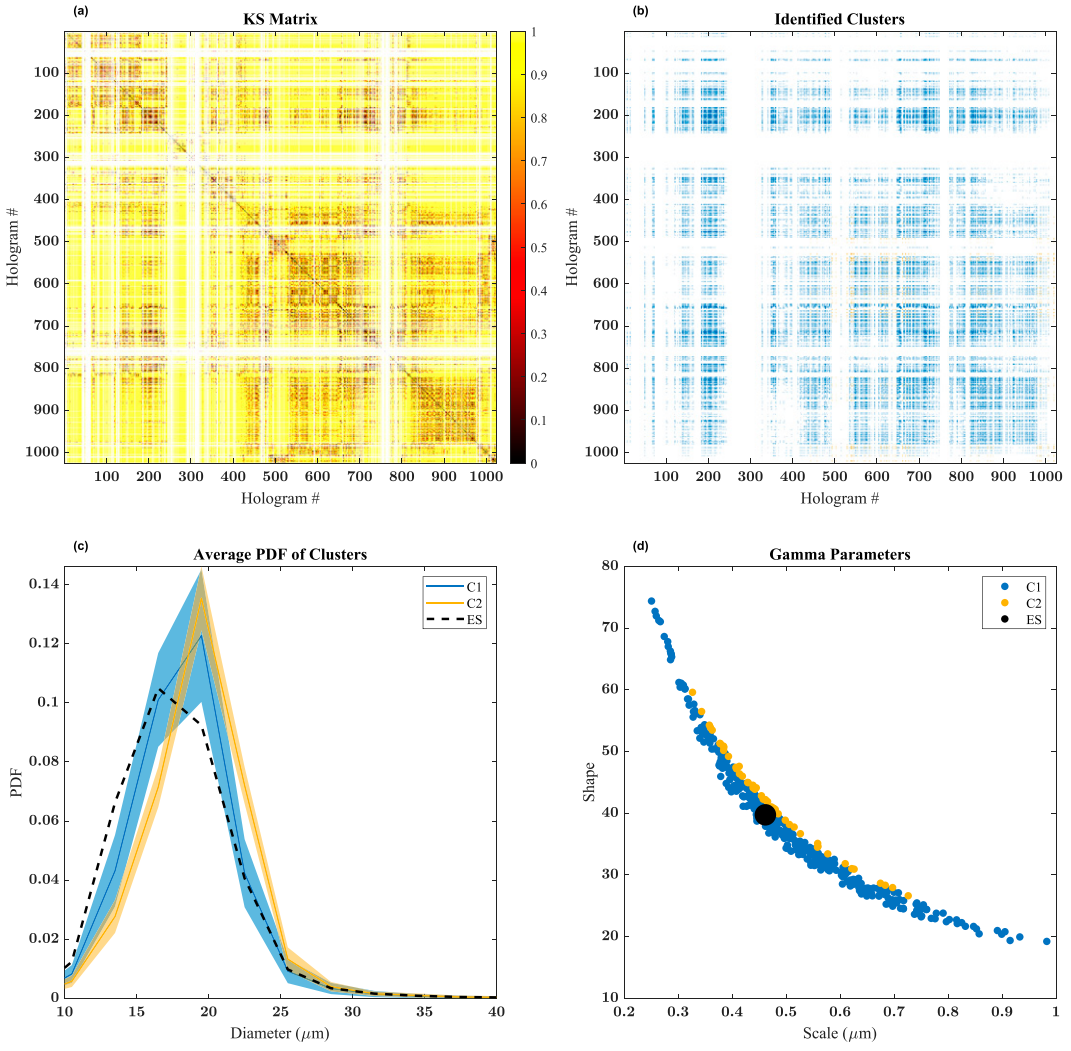

FIG. 5. As in Fig. 4, but for segment S4.

the same mean shape and scale 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 (the latter of which is presented in section 3 of the online supplemental material). 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 be seen for the segment SD2 given in Fig. 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 labeled 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

Our algorithm distinguishes statistically similar distributions in discrete data samples by combining hypothesis testing with machine learning clustering algorithms. We now proceed with a discussion of its sensitivity to input parameters and its

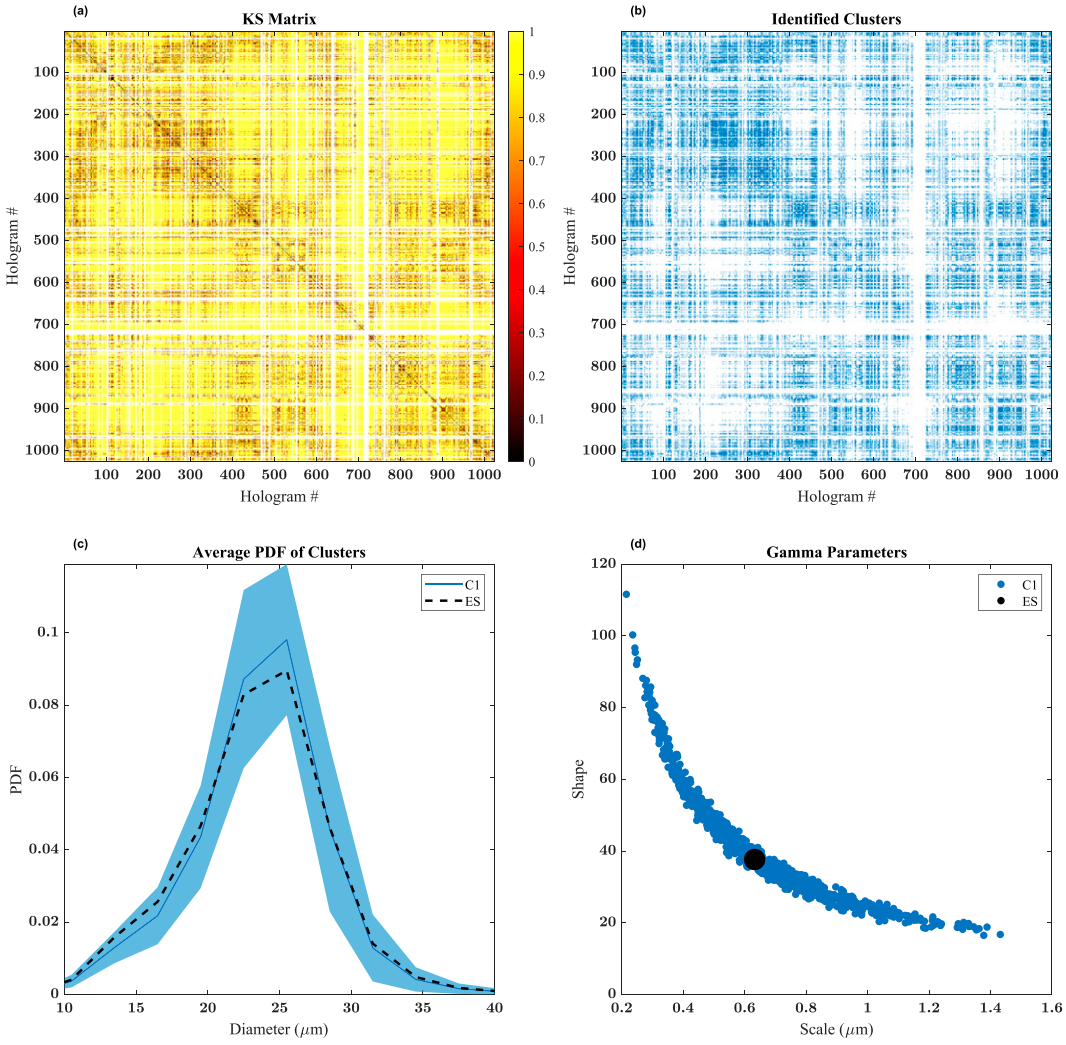

FIG. 6. As in Fig. 4, but for segment S2.

extension to other machine learning methods for cluster identification. We will then provide one example to illustrate the kinds of microphysical insights that can be gained from these clusters. However, we emphasize that the focus of this paper is not on exploring the physics behind the existence of clusters. The intriguing questions about physical processes related to characteristic droplet size distributions, including their size and distribution in space, will be looked at in more detail in subsequent work.

a. Generalizing the algorithm

The method employs standard statistical and machine learning tools. The two-sample KS 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 KS hypothesis test. It is one of a family of algorithms that have the advantage of not having a specified number of clusters or the requirement that the clusters are spatially connected. Despite our proposed approach having strengths of being "distribution agnostic" and not determining the number of clusters a priori, there are still choices of parameters that influence the outcome of the data processing. In particular, the critical value associated with rejecting the null hypothesis in the KS test (referred to as the alpha value and set to 0.05 for the results shown thus far), the cutoff value for subsampled holograms (set to 70% of the mean drop number in each hologram), the DBSCAN closeness and min-points parameters (set to 0.1 and 10 above, respectively), require some attention. Here (and in section 1 of the online supplemental material) we present some consideration of the impacts of the parameters used in our analysis; in general, the choices of these parameters may influence the number of clusters and number of holograms associated with each to some degree, but the qualitative capabilities of this hypothesis testing–data clustering approach remain largely unaltered.

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

	Predefined cluster properties			DBSCAN clusters: fraction from …		
Segment	Cluster elements	Noise holograms	No. of clusters identified	Original cluster	Noise holograms	Other clusters
SD1	664, 310, and 34	16	3	0.993	0.007	0
				1	0	0
				0.903	0.097	0
SD2	518, 274, and 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, and 47	357	3	0.776	0.224	0
				1	0	0
				0.366	0.644	0

Perhaps the most critical parameter in the processing structure is the choice of alpha in the 2-sample KS test. Formally, alpha specifies the significance level for each binary KS test; small values of alpha are more lenient in regard to concluding two distributions are likely from the same parent population. As such, if alpha is chosen smaller than our utilized value of 0.05, we find more dark-colored regions within the KS matrices, as seen, for example, in Fig. 2. Decreasing alpha does a better job of ensuring that statistically similar distributions are evaluated as coming from the same parent distribution, but it accomplishes this at the expense of potentially erroneously judging some dissimilar distributions to have come from the same parent distribution. As in any statistical testing problem, it is up to the user to decide whether "false alarms" or "missed detections" are more detrimental. We can be confident, in any case, that numerically experimenting with alternative values of alpha was found to have an impact on the clustering observed that is consistent with this understanding.

When lowering the alpha value, the number of dark cells in the KS matrix increases and consequently DBSCAN identifies fewer total clusters. This typically happens when the presence of noise clusters helps to bridge the gap between two or more collections of holograms that, when using a larger alpha value, were isolated as separated clusters due to the more stringent requirements of the KS test. This examination of the effects of changing alpha highlights that one limitation of DBSCAN is a lack of robustness in retaining cluster information in the presence of considerable background noise. This issue is resolved in more advanced clustering algorithms like OPTICS (Ankerst et al. 1999; Ester 2013). OPTICS uses information about the density of points in space to eliminate these noise outliers. More discussion on this is included in section 6 of the online supplemental material.

Because of the nature of the noise holograms, changing the cutoff for the KS tests also influences the results in nontrivial ways. A smaller threshold means more holograms are introduced for the KS comparisons, leading to smaller new clusters in some cases and joining the clusters in others. A larger threshold, on the other hand, eliminates more holograms from the KS tests and thus is not desirable. However, in numerous tests we have observed that information about the major clusters is mostly retained in all these cases. The exact nature and prevalence of the noise holograms and its effect on the final results is to be explored more in the future.

b. *Microphysical interpretation of the identified clusters*

Most flights of the ACE-ENA campaign consisted of vertical transects spanning the length of the boundary layer and the lower free troposphere. The flight studied here had an L-shaped horizontal path that performed "crosswind" and "along wind" sampling (Wang et al. 2022). The segments that we chose had crosswind (S1, S4, and S5), along-wind (S3), and mixed along- and crosswind (S2) components sampled at different altitudes. The strategy employed here was to choose homogeneous looking segments with little emphasis on their spatial locations. They are, however, at a specific altitude above the sea level. Unique clusters are identified for the different flight segments. Of particular interest is the nature of the major 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. No significant dependence is found for the clustering with altitude or the sampling strategy 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. Segment S2 was sampled along the crosswind and along-wind directions and was 2 times the length of segment S1. Note that S1 is sampled at a higher altitude relative to S2 and therefore might be closer to the cloud-top, providing clues to the larger spatial variability in its size distributions. The cloud's midregions (S3 and 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 clusters. From our preliminary analysis, we do not see evidence for a strong dependence of the cluster shapes and

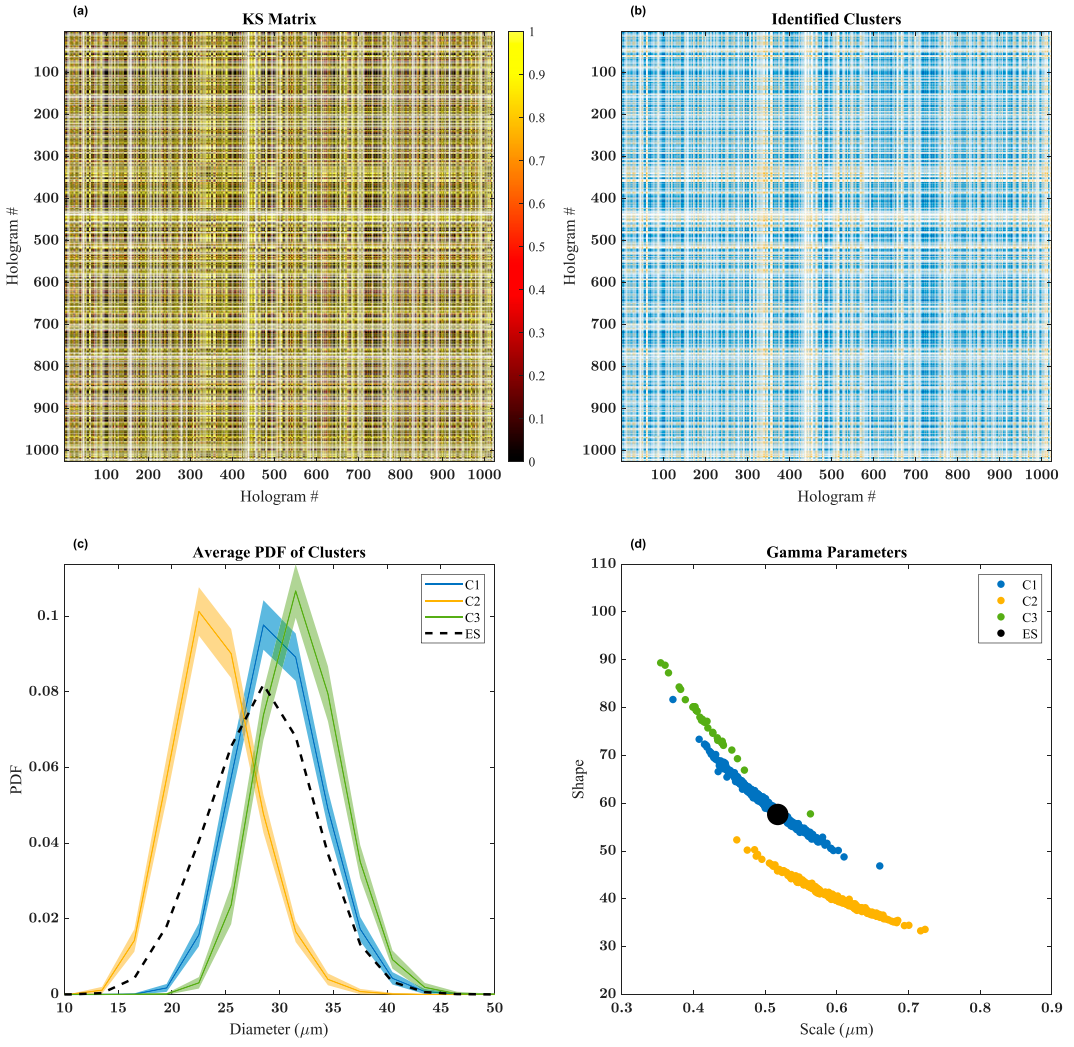

FIG. 7. As in Fig. 4, but for synthetic data SD1.

numbers on sampling along the crosswind and along wind directions.

More investigation is required to understand the physics underlying the existence of characteristic size distributions. 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). Additionally, the noise holograms in all these cases represent distributions with gamma parameters that are outliers from the clusters, which may have significant implications for the cloud microphysics. The range of PDFs for the noise holograms for these segments can be found in section 4 of the online supplemental material.

Once extended to more flights, the presence of these similar droplet size distributions may help in improving the understanding of the cloud processes. What physical processes influence these distribution shapes for the individual clusters are to be explored more in the future. The correlations of different microphysical and dynamical quantities with the identified clusters are of particular interest. As a first example of insight that can be gained, we consider the "location" of the identified cluster distributions in a microphysical phase space similar to a mixing diagram. Figure 9 shows liquid water content versus droplet number concentration for segment S5, with each hologram displayed as a single data point. As expected, there is a general trend for liquid water content to increase with number concentration. The data also seem to lie in various patterns that are roughly linear in shape, and rather strikingly, the linear features correspond to the identified clusters of similar distribution shape. We interpret this behavior as suggestive of inhomogeneous mixing, where the shape of the distribution is essentially fixed, and dilution results solely in a change of droplet number concentration (i.e., total evaporation of a subset of droplets, with the

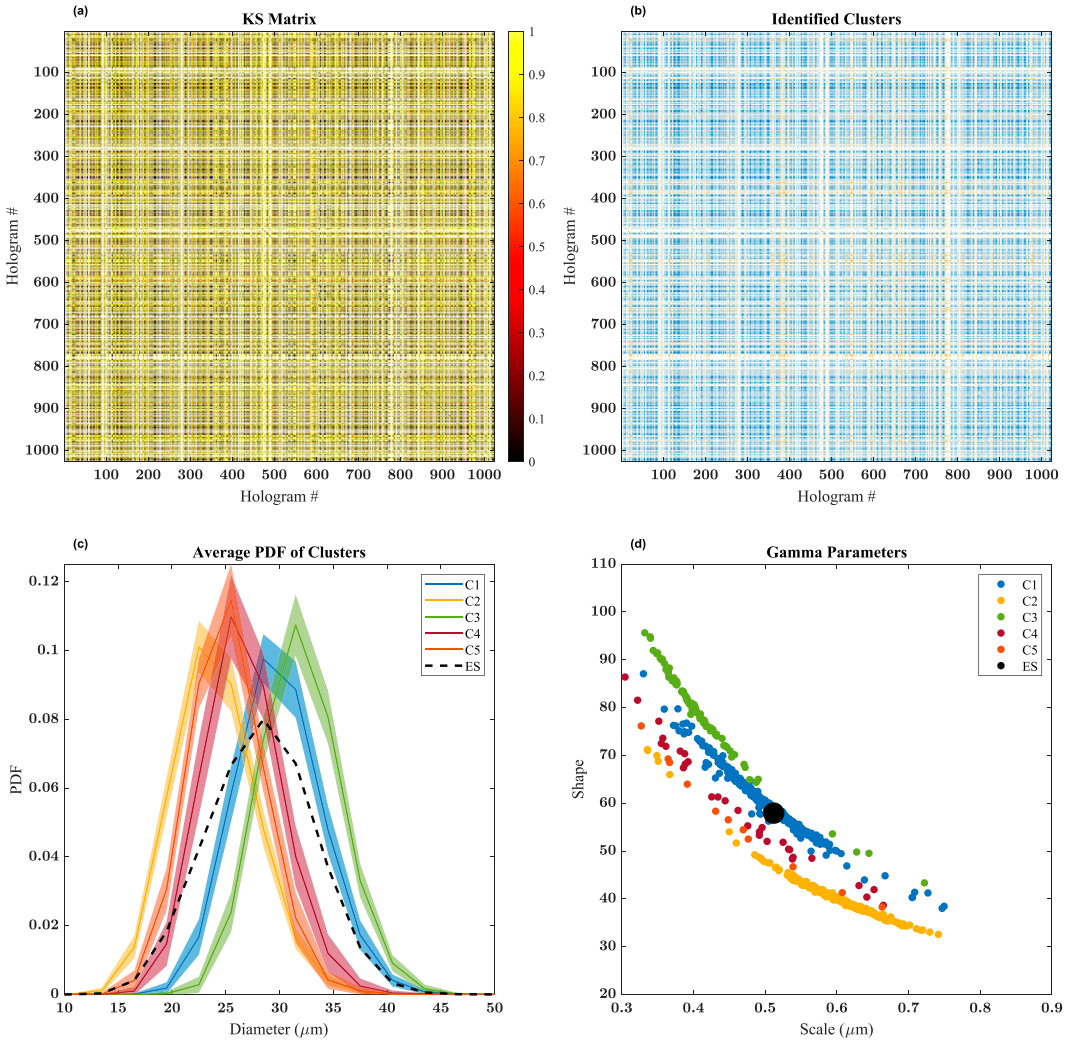

FIG. 8. As in Fig. 4, but for synthetic data SD2.

rest not changing in size appreciably). The slope of a best-fit line through each cluster is related to the mean-volume diameter of the characteristic distribution identified by the KS test combined with the data clustering algorithm. We have chosen this segment to illustrate the point because the linear features are visually evident. In other flight segments, however, the clusters overlap significantly in this microphysical space and therefore the clustering algorithm is necessary for their identification. This intriguing physics will be further explored in subsequent research involving more research flights from the ACE-ENA dataset.

c. *Concluding remarks*

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 18 July 2017. These distributions perhaps could be thought of as analogous to basis sets of a coordinate system in linear algebra.

The reliability and robustness of the algorithm are also verified using a synthetic dataset that mimics multiple elements of our field data. Synthetic datasets 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.

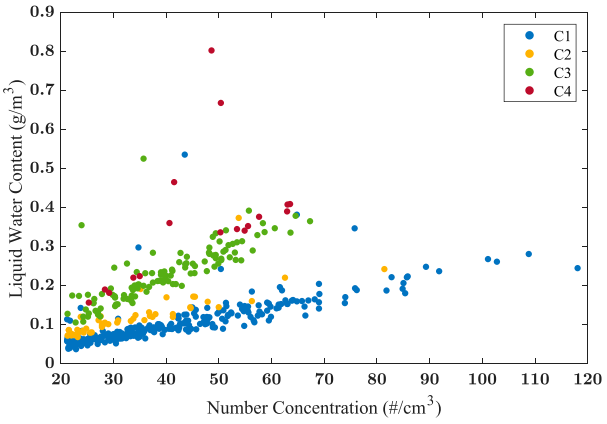

FIG. 9. Cloud liquid water content (g m^{-3}) vs cloud droplet number concentration (No. cm^{-3}) for flight segment S5. Each point corresponds to a single hologram. The color of the points corresponds to the characteristic size distributions identified by the clustering algorithm.

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 drop size distributions. Similarly, the Doppler spectrum from 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 advantages 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.

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 are available online for downloading (https://www.arm.gov/research/campaigns/aaf2017ace-en).

REFERENCES

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander, 1999: Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, **28**, 49–60, https://doi.org/10.1145/304181.304187.

Barlow, R. J., 1993: *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. John Wiley and Sons, 222 pp.

Beals, M. A., 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**, 87–90, https://doi.org/10.1126/science.aab0751.

Ester, M., 2013: Density-based clustering. *Data Clustering: Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds., Chapman and Hall/CRC, 111–126.

——, H.-P. Kriegel, J. Sander, and X. Xu, 1996: A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Second Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, Association for Computing Machinery, 226–231.

Fugal, J. P., and R. A. Shaw, 2009: Cloud particle size distributions measured with an airborne digital in-line holographic instrument. *Atmos. Meas. Tech.*, **2**, 259–271, https://doi.org/10.5194/amt-2-259-2009.

——, ——, E. W. Saw, and A. V. Sergeyev, 2004: Airborne digital holographic system for cloud particle measurements. *Appl. Opt.*, **43**, 5987–5995, https://doi.org/10.1364/AO.43.005987.

——, T. J. Schulz, and R. A. Shaw, 2009: Practical methods for automated reconstruction and characterization of particles in digital in-line holograms. *Meas. Sci. Technol.*, **20**, 075501, https://doi.org/10.1088/0957-0233/20/7/075501.

Gan, G., C. Ma, and J. Wu, 2020: *Data Clustering: Theory, Algorithms, and Applications*. 2nd ed. SIAM, 430 pp.

Glienke, S., A. B. Kostinski, R. A. Shaw, M. L. Larsen, J. P. Fugal, O. Schlenczek, and S. Borrmann, 2020: Holographic observations of centimeter-scale nonuniformities within marine stratocumulus clouds. *J. Atmos. Sci.*, **77**, 499–512, https://doi.org/10.1175/JAS-D-19-0164.1.

Hahn, C. J., and S. G. Warren, 2007: A gridded climatology of clouds over land (1971–96) and ocean (1954–97) from surface observations worldwide. Oak Ridge National Laboratory Carbon Dioxide Information Analysis Center Doc. ORNL/CDIAC-153 NDP-026E, 71 pp., https://atmos.washington.edu/~sgw/PAPERS/2007_ndp026e.pdf.

Hartmann, D. L., M. E. Ockert-Bell, and M. L. Michelsen, 1992: The effect of cloud type on earth's energy balance: Global analysis. *J. Climate*, **5**, 1281–1304, https://doi.org/10.1175/1520-0442(1992)005<1281:TEOCTO>2.0.CO;2.

Igel, A. L., and S. C. van den Heever, 2017: The importance of the shape of cloud droplet size distributions in shallow cumulus clouds. Part II: Bulk microphysics simulations. *J. Atmos. Sci.*, **74**, 259–273, https://doi.org/10.1175/JAS-D-15-0383.1.

Jaffrain, J., and A. Berne, 2011: Experimental quantification of the sampling uncertainty associated with measurements from PARSIVEL disdrometers. *J. Hydrometeor.*, **12**, 352–370, https://doi.org/10.1175/2010JHM1244.1.

Jameson, A. R., and A. B. Kostinski, 2000: Fluctuation properties of precipitation. Part VI: Observations of hyperfine clustering and drop size distribution structures in three-dimensional rain. *J. Atmos. Sci.*, **57**, 373–388, https://doi.org/10.1175/1520-0469(2000)057<0373:FPOPPV>2.0.CO;2.

——, M. L. Larsen, and A. B. Kostinski, 2015a: Disdrometer network observations of finescale spatial-temporal clustering in rain. *J. Atmos. Sci.*, **72**, 1648–1666, https://doi.org/10.1175/JAS-D-14-0136.1.

——, ——, and ——, 2015b: On the variability of drop size distributions over areas. *J. Atmos. Sci.*, **72**, 1386–1397, https://doi.org/10.1175/JAS-D-14-0258.1.

——, ——, and ——, 2018: On the detection of statistical heterogeneity in rain measurements. *J. Atmos. Oceanic Technol.*, **35**, 1399–1413, https://doi.org/10.1175/JTECH-D-17-0161.1.

Kendall, M., and A. Stuart, 1979: *The Advanced Theory of Statistics*. Vol. 2. Charles Griffin, 723 pp.

Larsen, M. L., and K. A. O'Dell, 2016: Sampling variability effects in drop-resolving disdrometer observations. *J. Geophys. Res. Atmos.*, **121**, 112777–112791, https://doi.org/10.1002/2016JD025491.

——, A. B. Kostinski, and A. Tokay, 2005: Observations and analysis of uncorrelated rain. *J. Atmos. Sci.*, **62**, 4071–4083, https://doi.org/10.1175/JAS3583.1.

——, R. A. Shaw, A. B. Kostinski, and S. Glienke, 2018: Fine-scale droplet clustering in atmospheric clouds: 3D radial distribution function from airborne digital holography. *Phys. Rev. Lett.*, **121**, 204501, http://doi.org/10.1103/PhysRevLett.121.204501.

Miles, N. L., J. Verlinde, and E. E. Clothiaux, 2000: Cloud droplet size distributions in low-level stratiform clouds. *J. Atmos. Sci.*, **57**, 295–311, https://doi.org/10.1175/1520-0469(2000)057<0295:CDSDIL>2.0.CO;2.

Shaw, M. A., 2016: Testing lidar-radar derived drop sizes against in situ measurements. M.S. thesis, Dept. of Physics, Michigan Technological University, 41 pp.

Slingo, A., 1990: Sensitivity of the Earth's radiation budget to changes in low clouds. *Nature*, **343**, 49–51, https://doi.org/10.1038/343049a0.

Spuler, S. M., and J. Fugal, 2011: Design of an in-line, digital holographic imaging system for airborne measurement of clouds. *Appl. Opt.*, **50**, 1405–1412, https://doi.org/10.1364/AO.50.001405.

Stephens, G. L., 2005: Cloud feedbacks in the climate system: Acritical review. *J. Climate*, **18**, 237–273, https://doi.org/10.1175/JCLI-3243.1.

Straka, J. M., 2009: *Cloud and Precipitation Microphysics: Principles and Parameterizations*. Cambridge University Press, 408 pp.

Wang, J., and Coauthors, 2022: Aerosol and Cloud Experiments in the Eastern North Atlantic (ACE-ENA). *Bull. Amer. Meteor. Soc.*, **103**, E619–E641, https://doi.org/10.1175/BAMS-D-19-0220.1.

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. *Atmos. Chem. Phys.*, **19**, 1077–1096, https://doi.org/10.5194/acp-19-1077-2019.

Supplemental Material

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 S 1. This demonstrates that even for smaller "min points", the algorithm identifies finite number of characteristic distributions.

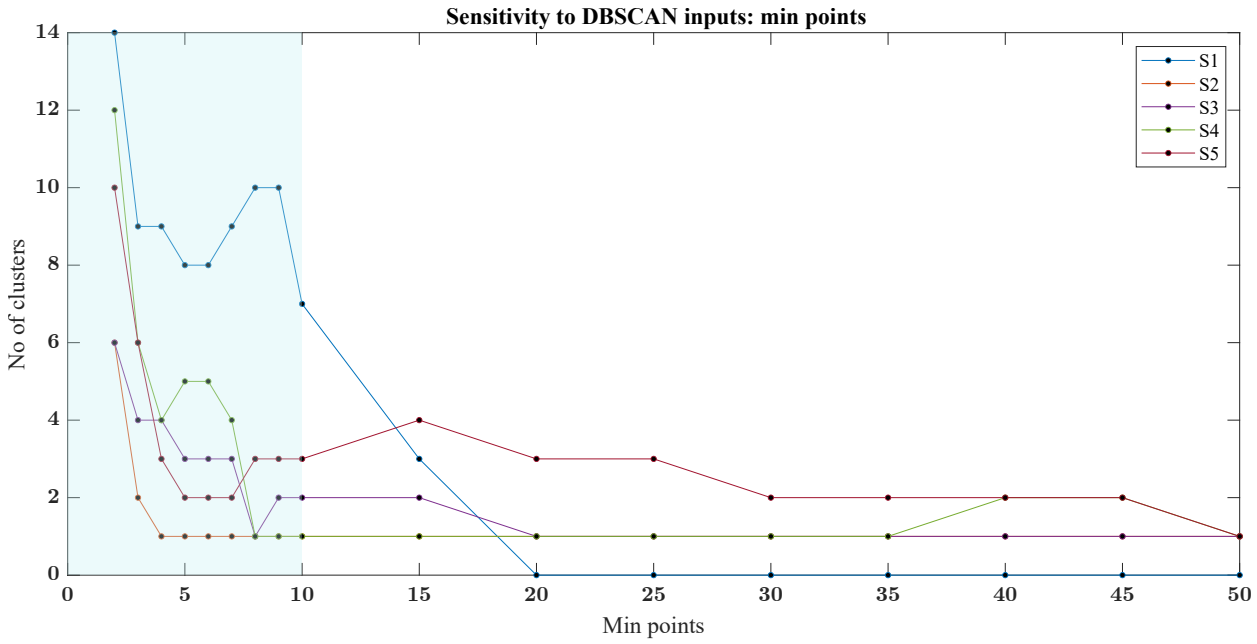

FIG. S 1. Sensitivity of the DBSCAN results to the input parameter: "min points". The number of clusters for the different cloud segments are shown. Here the value of "epsilon" is fixed to be 0.1.

2. Results for segment S3 & S5

In the main paper, KS and DBSCAN results for S1, S2, and S4 are shown. Here, for completeness, we show the results from S3 and S5 as well. 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.

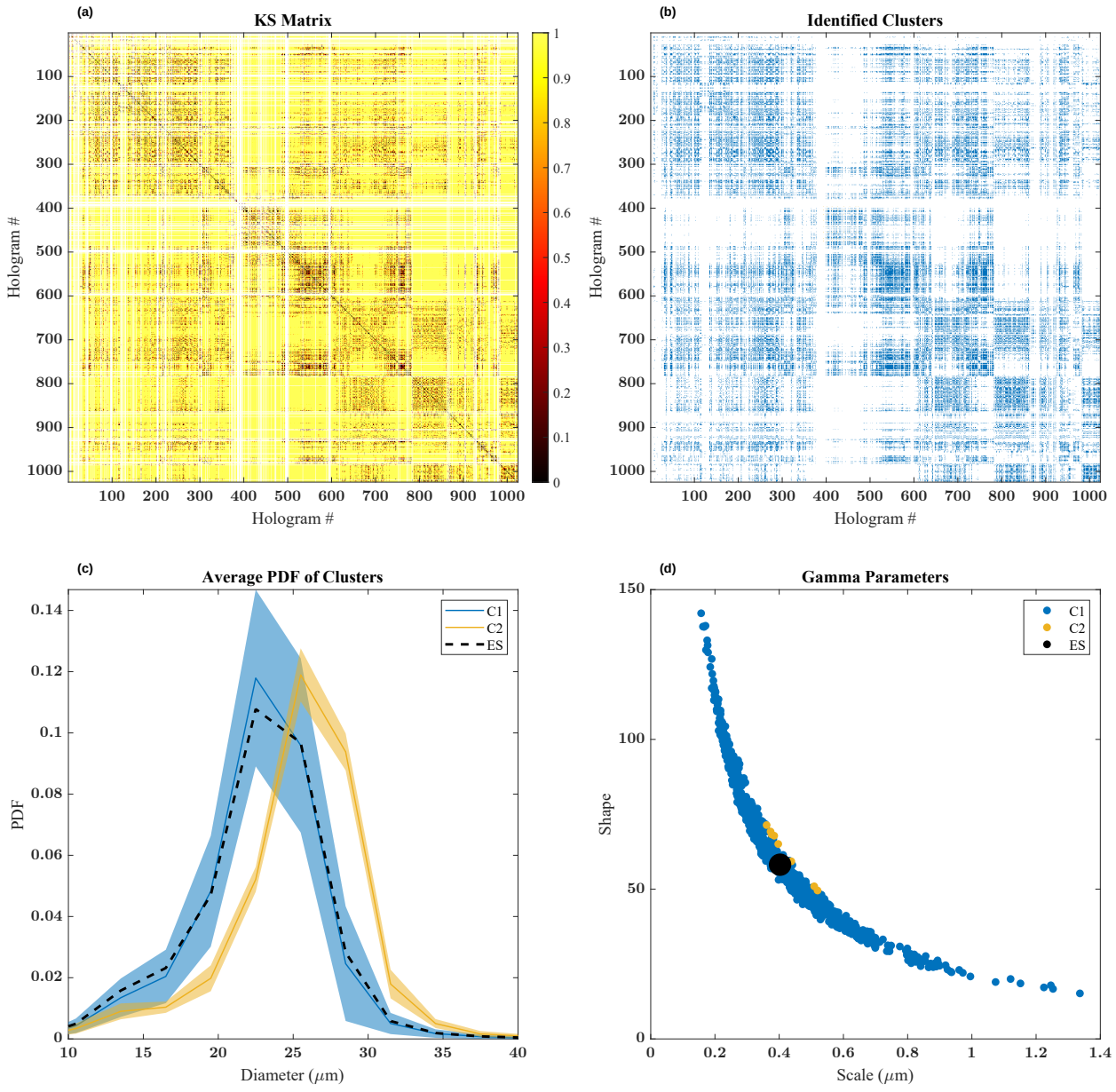

FIG. S 2. Results from Segment S3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire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

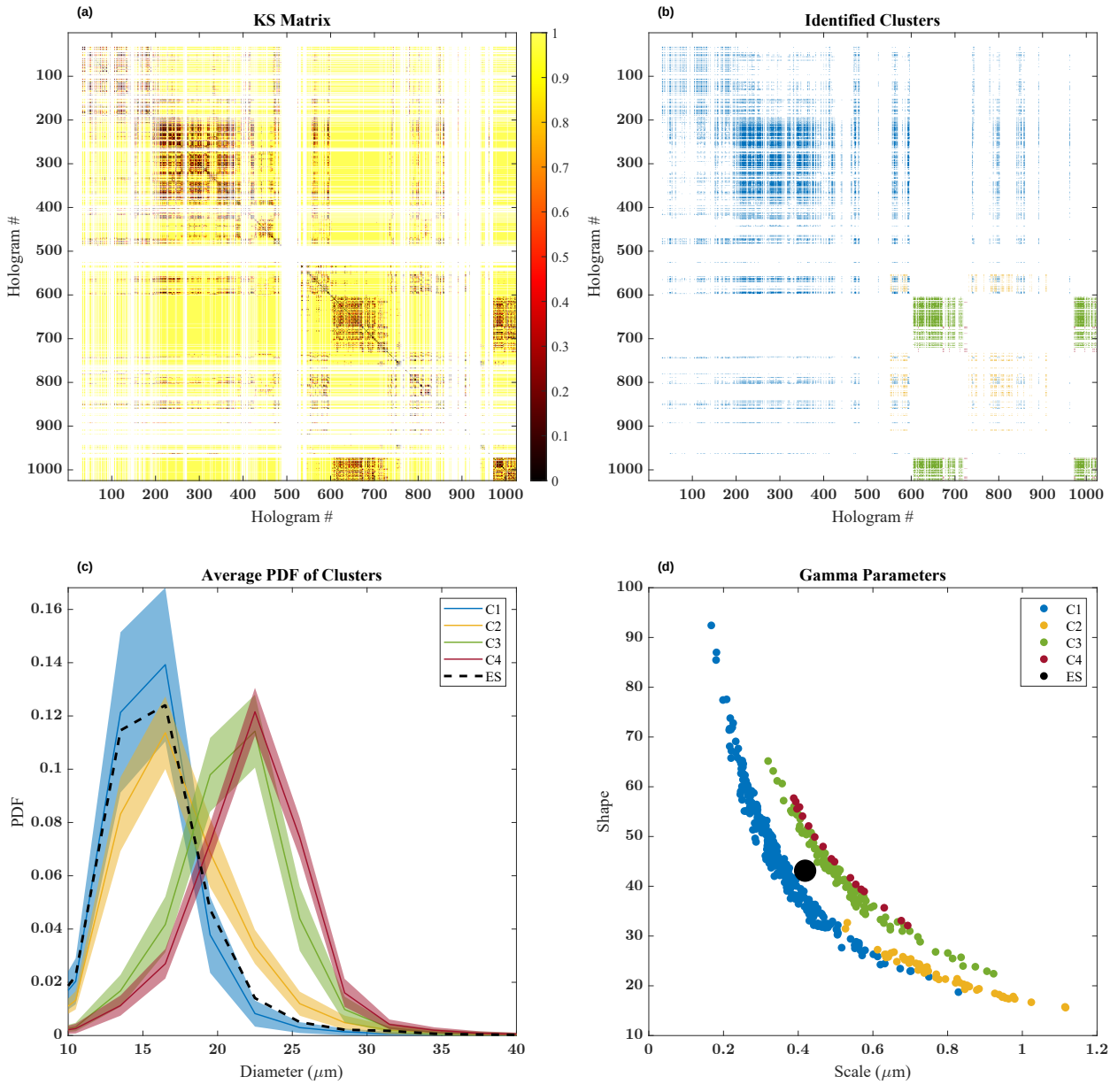

FIG. S 3. Results from Segment S5 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire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

3. Results of synthetic holograms-Segment SD3

In the main paper, results for synthetic datasets SD1 and SD2 are shown. For completeness, we show the results from SD3 as well.

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.

Note that, despite *more* noise than in SD2, here we only see the three clusters actually explicitly created, whereas SD2 showed an additional two clusters. This hints at the complicated interplay between the prevalence of noise and sensitivity to KS similarity statistics illustrated in the main text and below that makes direct intercomparison between the KS alpha statistic and statistical certainty more nuanced than might be expected.

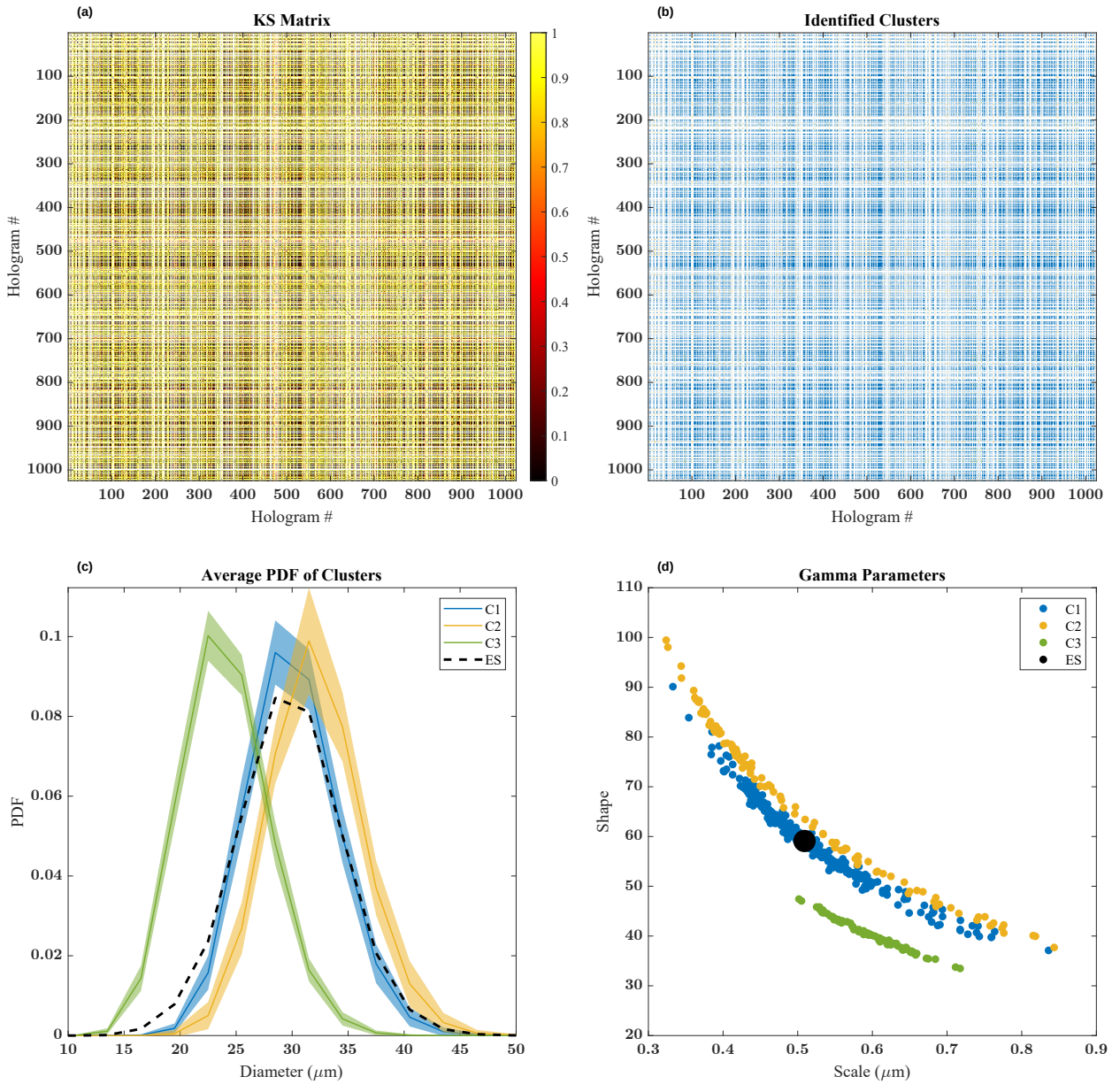

FIG. S 4. Results for the synthetic hologram set: SD3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows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

4. Noise Holograms from the different segments

Noise holograms are those that are not included in any identified cluster. These outliers are still of interest when looking at cloud processes. Figure S 5 shows the noise holograms for the different

segments. We have plotted all distributions so that the range of observations can be seen. The range of distribution shapes found within these noise holograms are quite varied. They still contain some features that appear to belong to the already detected clusters in the respective segments. In some cases there is bimodality that suggests the holograms may be mixtures of two characteristic distributions. In any case, they represent holograms that did not meet the algorithm's selection criteria to be assigned to the clusters.

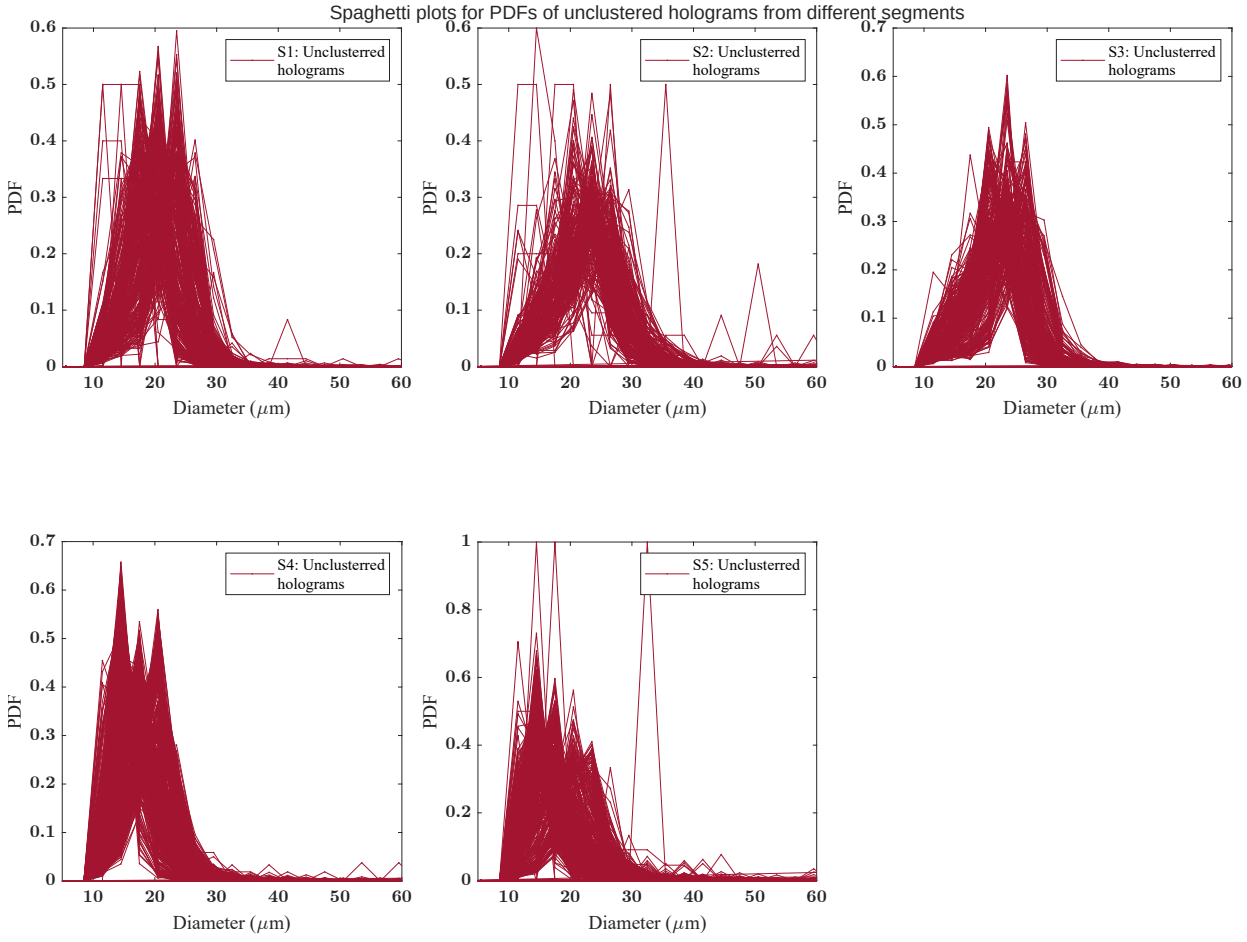

FIG. S 5. Spaghetti plots of the PDFs for the noise holograms, i.e., the holograms not belonging to any clusters from the segments S1-S5.

5. More information on Holographic Detector for Clouds (HOLODEC)

Holographic Detector for Clouds (HOLODEC) employs digital in-line holography to detect all particles within a sample volume. HOLODEC gives the shape, size and three dimensional position of all particles within a size range of about 6-2000 μm. The instrument operates at a frequency of 3.3 Hz and has a sample volume of approximately 19 cm^3. A laser illuminates a region between the arms of the instrument, constituting the sample volume. The particles in this volume scatter light, along with the main beam, into the detector CCD camera. The electric field reaching the camera is from the interference of the reference beam and the scattered elements by the individual particles. Because of this, phase information is preserved and this information can be used to back propagate the intensity field to obtain the accurate shape and position of every particle in the sample volume. This process is done computationally and is resource intensive. Once the potential particles are reconstructed digitally, they are then processed using a series of automated and manual techniques to remove the noise and obtain the final data product. More information on HOLODEC, its specifications (including photos), data samples, and a history of field deployments can be found on the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) web page for HOLODEC (https://www.eol.ucar.edu/instruments/holographic-detector-clouds).

6. Results for different alpha values

Due to the nature of dependence of the clustering algorithm on the KS matrix, the final results are influenced by the significance level specified for the KS tests. Figure S 6 show the results for different alpha values for the segment S1. As smaller alpha values makes the KS test more lenient, the number of clusters reduces with decrease in alpha value. However, the larger clusters are still preserved for an alpha value of 0.01. For the extreme case of 0.001, the distinction between different clusters is lost. This may be due to the inability of DBSCAN to resolve the clusters in the presence of considerable "noise" holograms. This can be resolved using other advanced density clustering algorithms like OPTICS. Ordering points to identify the clustering structure (OPTICS) resolves this issue by classifying points of varying spatial densities to form different clusters. Here, the KS statistic is used instead of the spatial points. And therefore, OPTICS has only one input parameter (Min points). It performs better classification in the presence of considerable noise as illustrated Figure S 6(d).

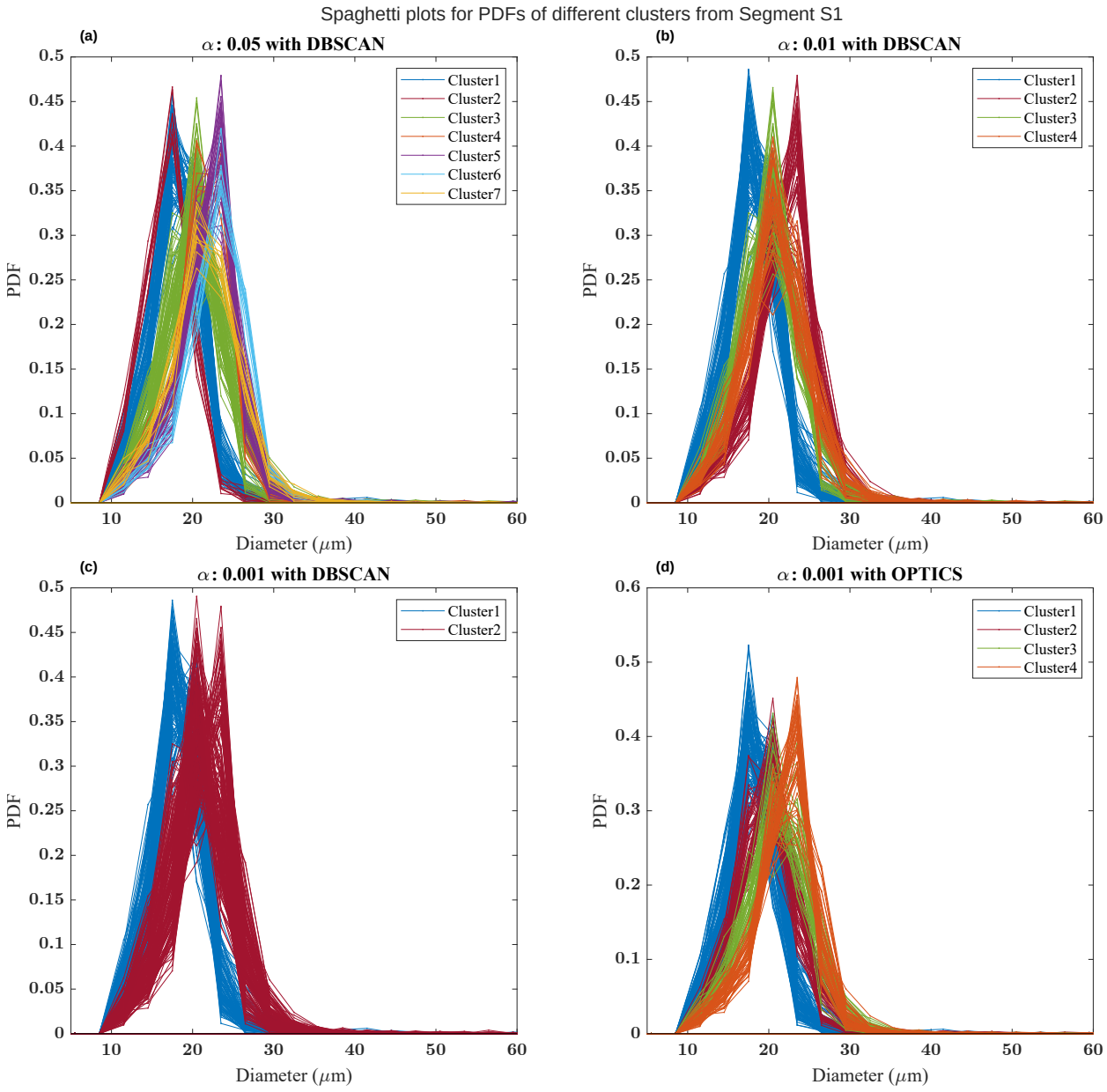

FIG. S 6. Spaghetti plots of the PDFs for the different clusters from the segment S1 corresponding to alpha values of 0.05, 0.01, 0.001 and 0.001. The sub-figures (a) to (c) are classified using DBSCAN and the sub-figure (d) is classified using OPTICS

7. PDF of the values for KS matrix from segment S1

The values of the KS test statistic in the KS matrix spans from 0 to 1 in steps of 0.001. To get an idea of how these values populate the KS matrix we plot the PDF of the values from the KS

matrix for segment S1 in Figure S 7. Here the values near 1 are not included in the calculation as they are the most numerous and masks the contribution of other values when plotted as a PDF. The distribution of values we see here comes from the stochastic element introduced by hologram resampling.

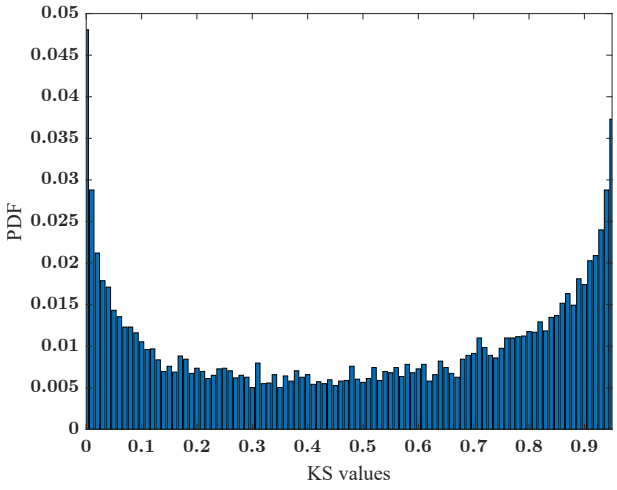

FIG. S 7. PDF of the values from KS matrix for segment S1. Only the values in the range 0-0.95 are shown because the bin at 1 overwhelms the rest of the distribution. The bin size is 0.01.