



Supplemental Material

© Copyright 2022 American Meteorological Society (AMS)

For permission to reuse any portion of this work, please contact permissions@ametsoc.org. Any use of material in this work that is determined to be “fair use” under Section 107 of the U.S. Copyright Act (17 USC §107) or that satisfies the conditions specified in Section 108 of the U.S. Copyright Act (17 USC §108) does not require AMS’s permission. Republication, systematic reproduction, posting in electronic form, such as on a website or in a searchable database, or other uses of this material, except as exempted by the above statement, requires written permission or a license from AMS. All AMS journals and monograph publications are registered with the Copyright Clearance Center (<https://www.copyright.com>). Additional details are provided in the AMS Copyright Policy statement, available on the AMS website (<https://www.ametsoc.org/PUBSCopyrightPolicy>).

**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.

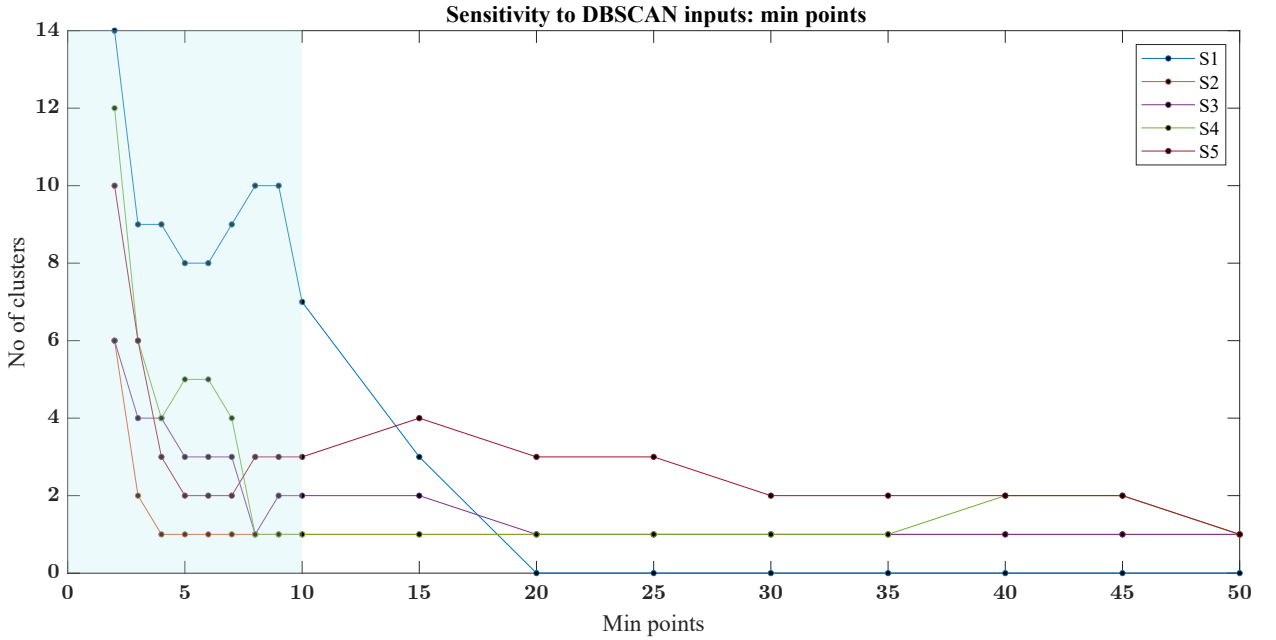


FIG. S 1. Sensitivity of the DBSCAN results to the input parameter: “min points”. The number of clusters for the different cloud segments are shown. Here the value of “epsilon” is fixed to be 0.1.

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.

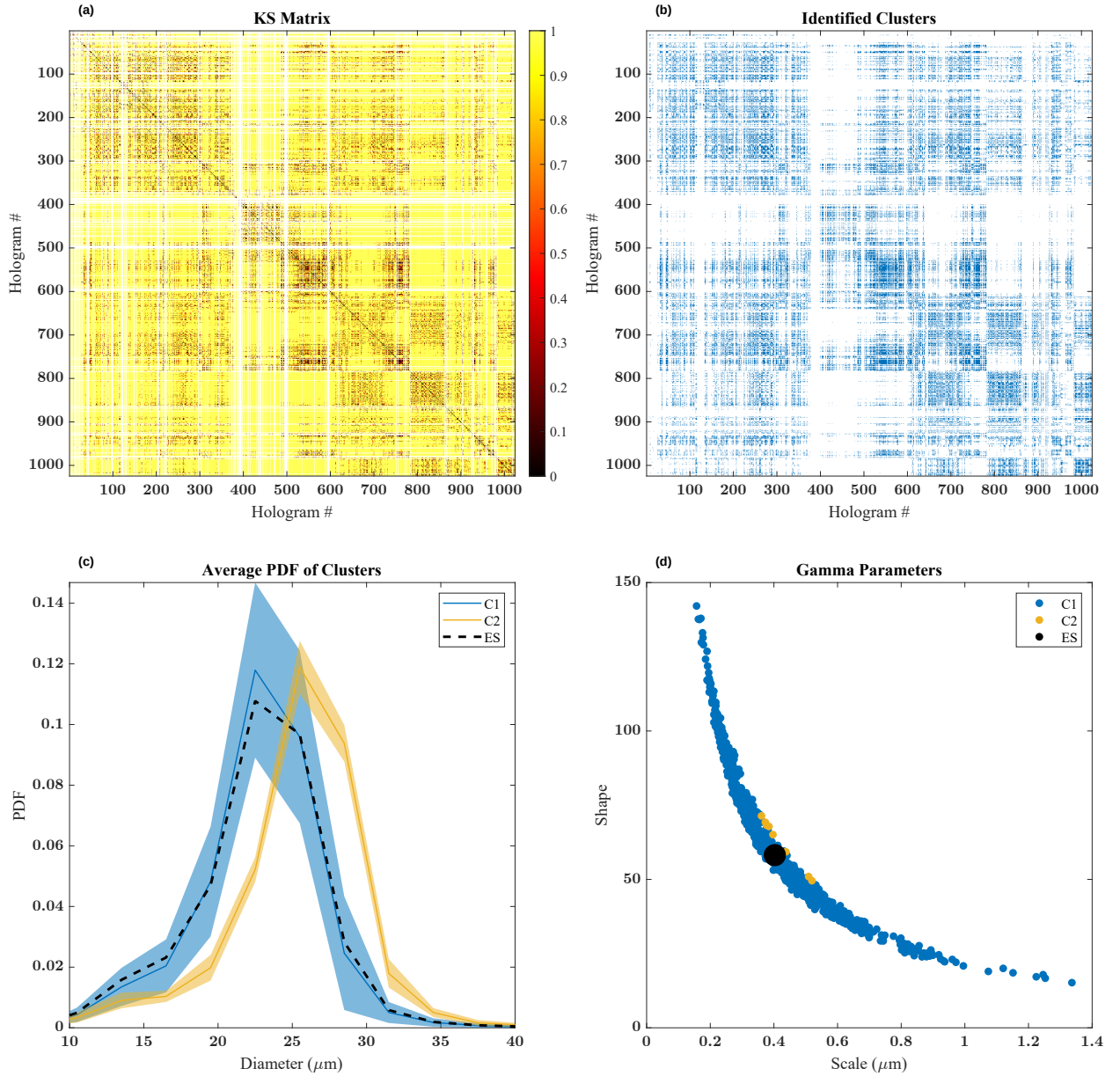


FIG. S 2. Results from Segment S3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire 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.

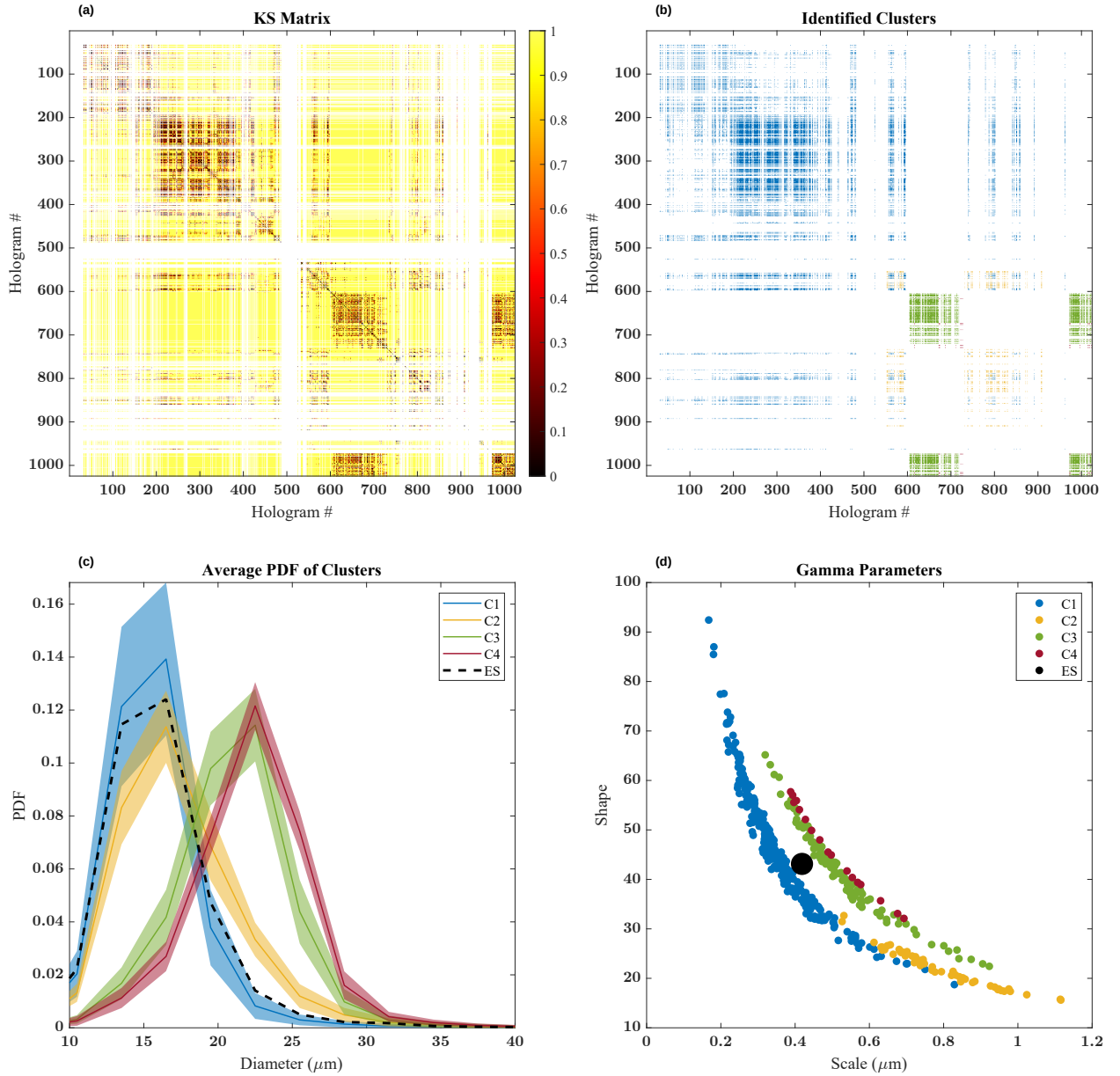


FIG. S 3. Results from Segment S5 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows the PDF for the entire 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.

35 **3. Results of synthetic holograms-Segment SD3**

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

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

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

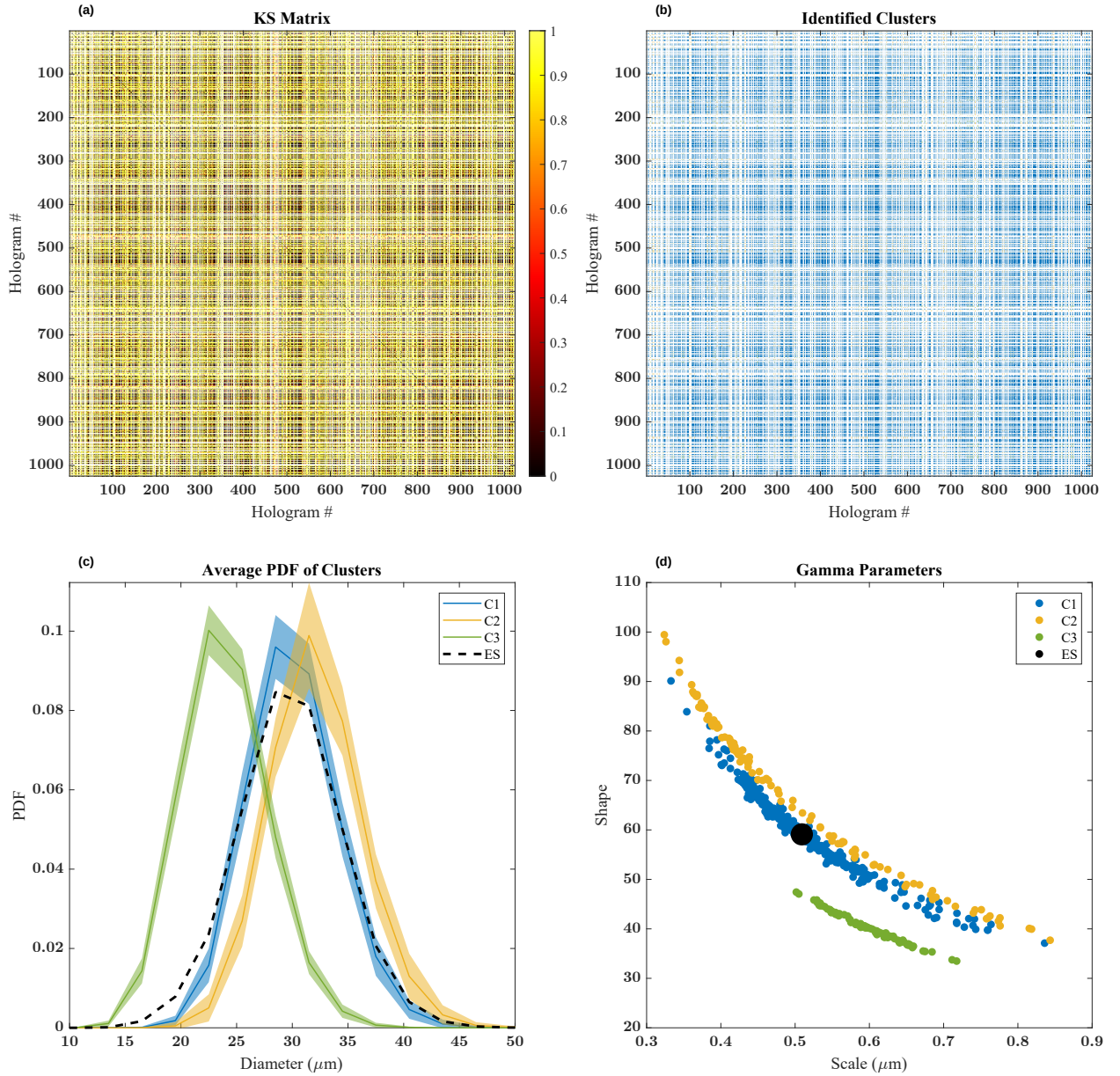
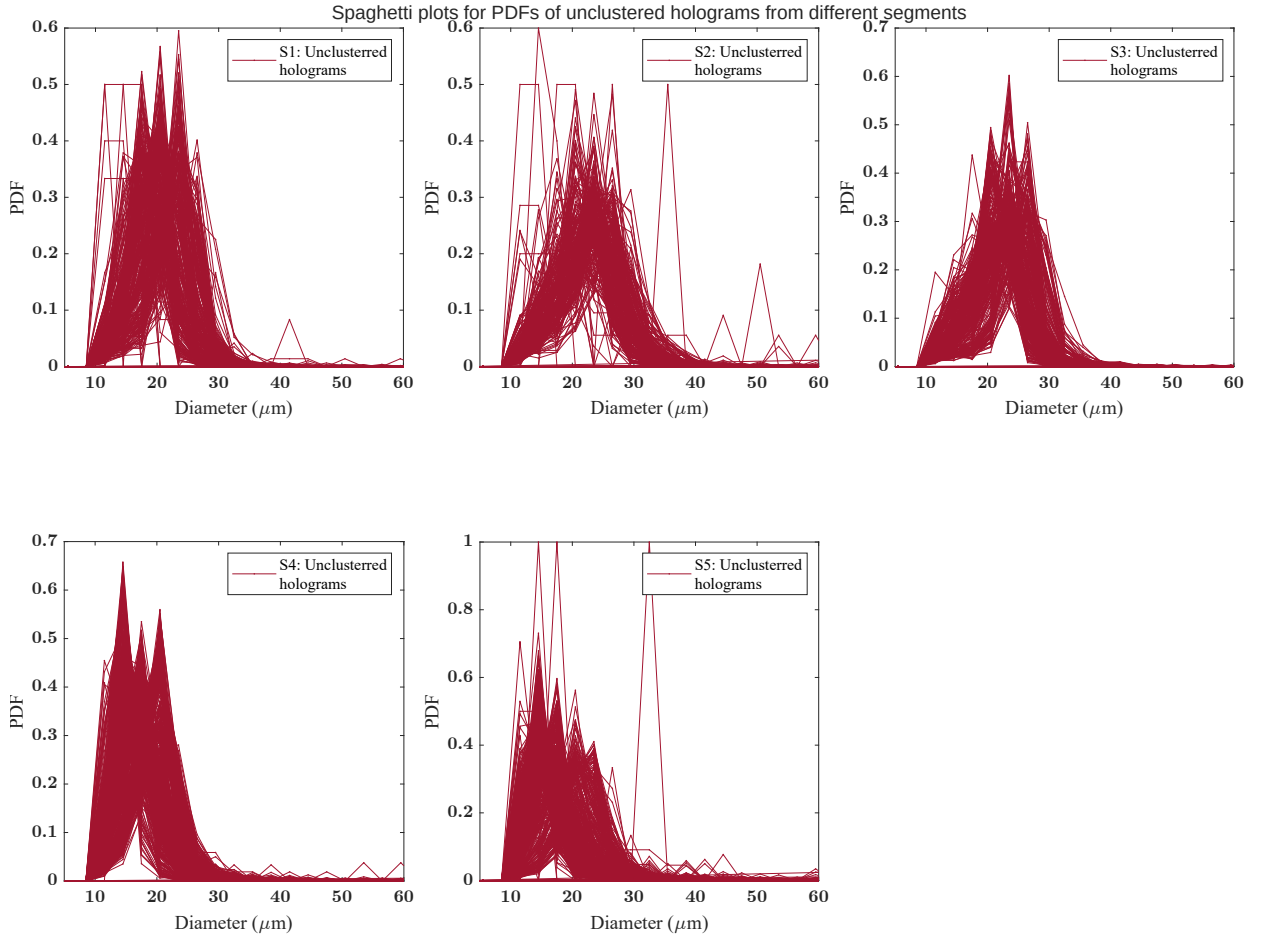


FIG. S 4. Results for the synthetic hologram set: SD3 (a) KS Matrix for the segment. (b) Clusters identified by the algorithm. The clusters are depicted by different colours. (c) Average PDFs of the different clusters. The shaded portion represents one standard deviation. The dashed black line shows 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

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



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

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

62 Holographic Detector for Clouds (HOLODEC) employs digital in-line holography to detect all
63 particles within a sample volume. HOLODEC gives the shape, size and three dimensional position
64 of all particles within a size range of about 6-2000 μm . The instrument operates at a frequency of
65 3.3 Hz and has a sample volume of approximately 19 cm^3 . A laser illuminates a region between
66 the arms of the instrument, constituting the sample volume. The particles in this volume scatter
67 light, along with the main beam, into the detector CCD camera. The electric field reaching the
68 camera is from the interference of the reference beam and the scattered elements by the individual
69 particles. Because of this, phase information is preserved and this information can be used to
70 back propagate the intensity field to obtain the accurate shape and position of every particle in the
71 sample volume. This process is done computationally and is resource intensive. Once the potential
72 particles are reconstructed digitally, they are then processed using a series of automated and manual
73 techniques to remove the noise and obtain the final data product. More information on HOLODEC,
74 its specifications (including photos), data samples, and a history of field deployments can be found
75 on the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL)
76 web page for HOLODEC (<https://www.eol.ucar.edu/instruments/holographic-detector-clouds>).

77 **6. Results for different alpha values**

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

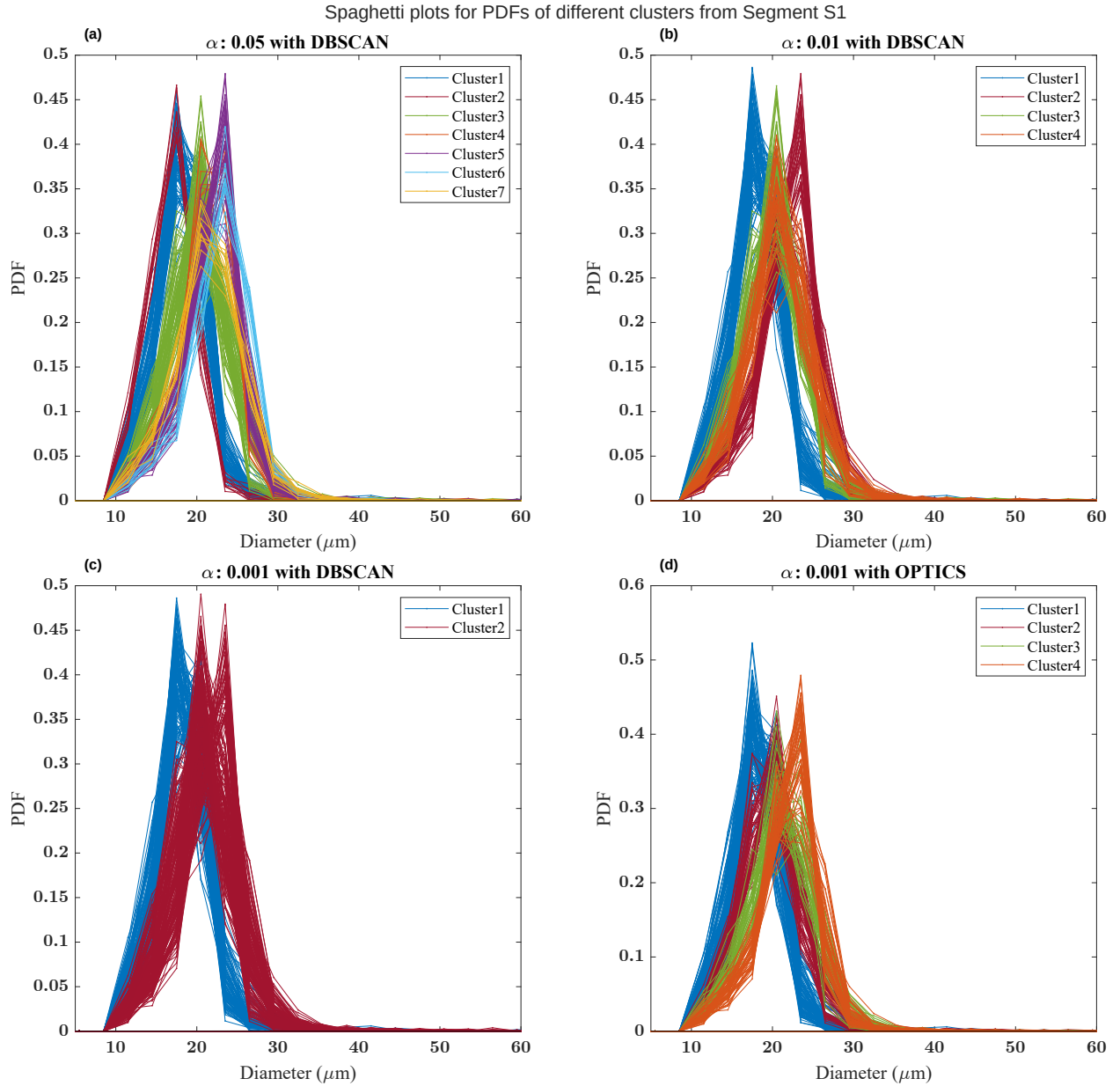
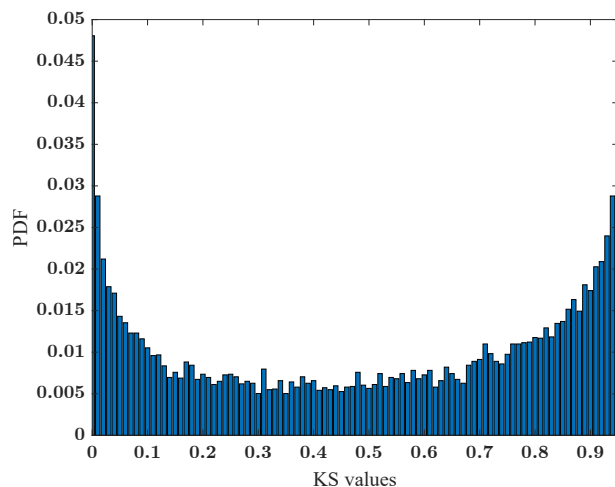


FIG. S 6. Spaghetti plots of the PDFs for the different clusters from the segment S1 corresponding to alpha values of 0.05, 0.01, 0.001 and 0.001. The sub-figures (a) to (c) are classified using DBSCAN and the sub-figure (d) is classified using OPTICS

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

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



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