1 # Hydrometeor classification of quasi-vertical profiles of polarimetric

2 # radar measurements using a top-down iterative hierarchical

3 # clustering method

4 Maryna Lukach[1, 2], David Dufton[1, 2], Jonathan Crosier[3,4], Joshua M. Hampton[1,2], Lindsay Bennett[1, 2] and
5 Ryan R. Neely III[1, 2]

6 [1]National Centre for Atmospheric Sciences, Leeds, United Kingdom

7 [2] School of Earth and Environment, University of Leeds, Leeds, United Kingdom

8 [3]National Centre for Atmospheric Sciences, University of Manchester, United Kingdom

9 [4]Department of Earth and Environmental Sciences, University of Manchester, Manchester, United Kingdom

11 *Correspondence to*: Maryna Lukach (maryna.lukach@ncas.ac.uk)

12 **Abstract.**
13 Correct, timely and meaningful interpretation of polarimetric weather radar observations requires an accurate understanding
14 of hydrometeors and their associated microphysical processes along with well-developed techniques that automatize their
15 recognition in both the spatial and temporal dimensions of the data. This study presents a novel technique for identifying
16 different types of hydrometeors from Quasi-Vertical Profiles (QVP). In this new technique, the hydrometeor types are
17 identified as clusters belonging to a hierarchical structure. The number of different hydrometeor types in the data is not
18 predefined and the method obtains the optimal number of clusters through a recursive process. The optimal clustering is then
19 used to label the original data. Initial results using observations from the NCAS X-band dual-polarization Doppler weather
20 radar (NXPol) show that the technique provides stable and consistent results. Comparison with available airborne in situ
21 measurements also indicates the value of this novel method for providing a physical delineation of radar observations.
22 Although this demonstration uses NXPol data, the technique is generally applicable to similar multivariate data from other
23 radar observations.

1 Introduction

The task of radar-based hydrometeor classification (HC) can be broadly defined as the recognition of different hydrometeor types in the atmosphere as represented by the various observed moments collected by weather radar. In general, HC is able to label radar signatures observed at any one time with physical properties and, over a period of time, the evolution of these labels can provide insight into the underlying atmospheric processes. As such HC has many impactful applications: HC simplifies the detection of the melting layer (Baldini and Gorgucci, 2006), HC is necessary for obtaining accurate estimates of precipitation quantities (Giangrande and Ryzhkov, 2007) and HC provides critical information for improving modelling of physical processes in the atmosphere (Vivekanandan et al., 1999).

Radar-based HC requires an extensive and accurate (i.e. expert) knowledge of the physical properties of both multivariate polarimetric observations and the hydrometeor particles themselves (Hall et al., 1984). Achieving an accurate and precise radar-based HC is difficult due to the deficiencies (such as low spatial-temporal resolution) and inaccuracies (such as attenuation) that are inevitable in all radar measurements. The process of HC is made even more difficult when this analysis needs to be performed during the operational processing of the radar observations where there is a lack of time for expert assessment. Therefore, automatization of spatial and temporal analysis of multivariate polarimetric data is an important task for which an advanced and well-tested technique should be developed and utilized.

The development of radar-based HC started in the 1980s and 1990s with the works of Hall et al. (1984), Hendry and Antar (1984), Aydin et al. (1986), Straka and Zrnić (1993) and Straka (1996). Further refinement and development of automatic HC algorithms included the application of fuzzy-logic, machine-learning techniques (such as the identification of clusters representing data-wise similarities) and neural networks (Straka et al., 2000; Liu and Chandrasekar, 2000; Wen et al.,2015; Grazioli et al., 2015; Besic et al., 2016; Wang et al., 2017; and Ribaud et al., 2019).

Modern radar-based HC methods (Straka et al, 1996; Liu and Chandrasekar, 2000; Al-Sakka et al., 2013, Grazioli et al., 2015; Besic et al., 2016; and Wang et al., 2017) are based on the multivariate data of polarimetric Doppler radar observations. This includes (but is not limited to): the logarithmic reflectivity factor at horizontal polarization Z_H, differential reflectivity Z_{DR}, the copolar correlation coefficient ρ_{HV}, differential phase Φ_{DP}, and specific differential phase K_{DP} (for definitions see Bringi and Chandrasekar, 2001) as well as associated derived variables (e.g. standard deviation). Additionally, temperature and other meteorological data, retrieved from radiosondes or NWP models, are often utilized (Grazioli et al., 2015; Wen et al., 2015).

In most existing radar-based HC methods, the multivariate input data are analysed per measurement voxel and determined classes are assigned to the hydrometeor types only according to their characteristics. Such an approach neglects intra-class relationships and the temporal evolution of the identified classes. This valuable information can also be used in the labelling of the hydrometeor types and the identification of corresponding microphysical processes. Additionally, almost all methods within the existing literature are based on theoretical assumptions on the scattering properties of observed particles and/or are only applicable for a defined (i.e. previously recognized) number and type of classes. Both of these aspects of existing HC methods are limitations. For example, theoretical assumptions about the scattering properties of ice-phase hydrometeors are

57 very uncertain due to unknown size distributions, varying dielectric properties, fall orientation, and their diverse and often
58 complex geometry (Johnson et al., 2012). A pre-defined number of classes or hydrometeor types is subjective and creates
59 artificial boundaries for algorithms and, thus, subtle differences in undefined sub-classes are masked, which inhibits
60 identification of the underlying microphysical processes.

61 Thus, in this study we take a different approach and ask the following question: can a data-driven HC approach provide an
62 optimal number of classes from the observations? We define the optimal number as the lowest number of classes representing
63 all pronounced dissimilarities in the input data. Once the optimal set of classes is identified, the burden of analysis in this
64 approach is to relate the identified clusters of radar signatures to possible physical properties of hydrometeors. Thus, this
65 approach does not impose a predefined physical view on the observations but provides a framework for a more efficient
66 physical interpretation of the properties of the resulting clusters of observed multivariate data in which subtle differences and
67 intra-cluster relations are easier to identify. In this sense, this approach inverts the procedure of existing methods. Additionally,
68 we ask whether such an approach can be used to provide information on the temporal evolution of the identified hydrometeors
69 and reveal relationships between the identified classes. Such information is key for identifying the processes that lead to high
70 impact weather (i.e. flooding) and improving the physical parametrizations in NWP.

71 The existing data-driven unsupervised (Grazioli et al., 2015; Ribaud et al., 2019 and Tiira and Moisseev, 2019) and semi-
72 supervised approaches (Besic et al., 2016; Wang et al., 2017 and Roberto et al., 2019) only partially provide an answer to the
73 first question (Grazioli et al., 2015) and do not consider the temporal evolution or dependencies between the identified classes.
74 The approach described here performs an unsupervised clustering of quasi-vertical profiles. QVPs were first used in Kumjian
75 (2012) and Ryzhkov et al. (2016) as a way of constructing a substitute for a vertical profile from a scan conducted at constant
76 elevation, which is a typical mode of scanning for radars used in operational networks. Calculation of the QVPs requires
77 horizontal homogeneity of the observed atmospheric processes. The height-vs-time format of QVPs represents the general
78 structure of the storm or its evolution.

79 The QVP input is used in this work to build a hierarchical structure based on the identified clusters and deliver an optimal
80 number of clusters based solely on internal properties of the multivariate polarimetric data. The optimal clustering is then used
81 to label the hydrometeor classes and to analyse the temporal evolution of the labelled microphysical processes.

82 The paper is organised as follows: in Section 2 we introduce the clustering methods we employ, Section 3 contains a description
83 of the polarimetric radar data and their processing, Section 4 describes the iterative clustering approach, leading to the
84 development of the hierarchical structure, Section 5 is devoted to the characterization of the clusters and their labelling, and
85 Section 6 concludes with a summary, discussion, and thoughts on further perspectives.

2 Background of employed methods

87 The proposed hierarchical clustering algorithm identifies the optimal number of groups of data points (clusters) in a recursive
88 loop and organizes the clusters in a hierarchical structure (undirected weighted graph). The two main steps of this approach

are the cluster identification and the optimality check. The cluster identification is achieved after performing dimensionality reduction by principal component analysis followed by spectral clustering. The optimality check uses validity indexes to identify the final set of clusters, which best classifies the provided set of data. The description of the dimensionality reduction and clustering methods with background information about the validity indexes employed can be found in this section directly after a short introduction to the hierarchical clustering.

2.1 Hierarchical clustering

Hierarchical clustering is a type of clustering technique that splits or combines the data through an iterative process. Unlike "flat" clustering techniques, hierarchical clustering is not performed in one stage. Rather, it repeats the clustering process iteratively and keeps the information about each iteration of clustering in a hierarchical structure. In general, for a given set of multivariate data points, a hierarchical clustering algorithm, depending on top-down or bottom-up direction, either partitions (divides) or merges (agglomerates) the data into groups (a set of clusters) where data points assigned to the same cluster show similarity in multivariate values (depending on the context, it could be, for example, having a small distance to each other if the points are in Euclidean space). The direction of the process (top-down or bottom-up) may be chosen and depends on the number of individual points in the multivariate dataset and the needs of the underlying problem.

The top-down method begins with all available data points organized in one cluster and splits this cluster into subclusters until a certain criterion is reached or only solely singleton clusters of individual points remain in the set. The bottom-up method, on the other hand, begins with all points assigned to individual clusters and at each step, merges the most similar pairs of clusters into one until all the subclusters are agglomerated into a single cluster.

The optimal number of clusters in both approaches can be identified using a termination criterion. The hierarchical structure in the bottom-up approach needs to be completely finished before the optimal number of clusters can be identified otherwise the upper part of the tree will remain unknown. The top-down approach allows for the iterative process to be stopped at any point whilst preserving the upper part of the hierarchical structure. Another advantage of the top-down approach is the possibility to have more than two subclusters belonging to one parenting cluster. This allows an optimal number of subclusters for each parenting cluster, representing the data-driven inheritances in the resulting hierarchical structure. Although this advantage is often not used, and the bottom-up methods are preferred (Grazioli, 2015 and Rimbaud et al., 2019), the method presented here fully exploits it for the identification of an optimal number of subclusters in each iteration.

2.2 Eigenvectors and Principal Components

Principal component analysis (PCA) is a statistical technique mostly utilized in exploratory analysis of multivariate data. It extracts the most important information from the multivariate dataset generating a simplified view of the original data by dimensionality reduction.

To reduce a dataset's dimensionality a set of new orthogonal, non-correlated variables called principal components is calculated as linear combinations of the original variables. The first component is selected having the largest possible variance,

21 so it best represents the diversity of the given data. The second component is generated under the assumption of orthogonality

22 to the first component whilst also having the largest possible variance. This process is continued until the number of principal

23 components is equal to the number of original variables (d). These components are exactly the eigenvectors of the correlation

24 matrix and are employed as a basis for a new coordinate system (Wold, 1976; Abdi and Williams, 2010). The first q calculated

25 coordinates having satisfactory representativeness (e.g. 85 %) can be used to preserve the most important characteristics of the

26 original data. These q principal components can replace the initial d variables ($q < d$), and the original data set, consisting of

27 N measurements on d variables, is reduced to a data set consisting of N measurements on q principal components.

28 **2.3 Clustering method**

29 One of the most popular clustering methods is the k-means algorithm (Steinhouse, 1956). Through its simple interpretability,

30 it is often used either as a single method or as a part of more computationally expensive clustering methods (e.g. Gaussian

31 mixtures or spectral clustering). As a single method, it has difficulties with non-convex clusters and is known to perform poorly

32 if the input variables are correlated (von Luxburg, 2007). As a basis of a more complex method (e.g. spectral clustering) it

33 allows a solution of non-linear cluster shapes to be found (any low-dimensional manifolds of high-dimensional spaces).

34 The input data herein are multi-dimensional and were found to have non-convex cluster shapes, therefore the spectral clustering

35 method was applied (Shi & Malik, 2000; Ng et al., 2002; von Luxburg, 2007). It works by approximating the problem of

36 partitioning the nodes in a weighted graph as an eigenvalue problem of eigenvectors described above and by applying the k-

37 means algorithm to this representation in order to obtain the clusters. This work implements the Ng et al. (2001) approach and

38 analyses the eigenvectors of the normalized graph Laplacian.

39

40 Spectral clustering has several appealing advantages. First, embedding the data in the eigenvector space of a weighted graph

41 optimizes a natural cost function by minimizing the pairwise distances between similar data points and such an optimal

42 embedding is analytically deducible. Secondly, as it was shown in von Luxburg (2007), the spectral clustering variants arise

43 as relaxations of graph balanced-cut problems. Finally, spectral clustering was shown to be more accurate than other clustering

44 algorithms such as k-means (von Luxburg, 2007).

45 **2.4 Determining optimal number of clusters**

46 Clustering algorithms can be roughly divided into two groups based on whether the number of clusters to be found is

47 predetermined or undetermined. Spectral clustering is a rather flexible technique in the sense that it can be used with a

48 relaxation (i.e. when the number of clusters to be found is provided) or without it (when the number of clusters is determined

49 by the multiplicity of the eigenvalue 0). As the approach chosen here is not interested in a flat partitioning of the data, rather

50 we want to determine hierarchical structures, the determination of the optimal number of clusters is important. To identify this

51 optimal number of clusters two evaluation scores are used in our method: the Wemmert-Gançarski (WG) index (Hämäläinen

52 et al., 2017) and Bayesian Information Criterion (BIC) index (Pelleg and Moore, 2000; Hancock, 2017). While the exact use

53 of the indexes is described in Sect. 4, the WG and BIC indexes can be defined as follows:

54 Let the data set $X = \{x_i \in \mathbb{R}^d : i = 1, ..., N\}$ have clustering $C_K = \{c_k : k = 1, ..., K\}, (K < N)$, where n_k – number of

55 samples/points in the cluster c_k and I_{c_k} – indexes of the points in X belonging to the cluster c_k.

56   1)   For WG index: Let $R(x_i)$ represent the mean of relative distances between the points belonging to the cluster c_k and

57   the centre of its barycentric weight g_k. The $R(x_i)$ value is calculated for each point x_i

$$R(x_i) = \frac{\|x_i - g_k\|}{min_{k \neq k'} \|x_i - g_{k'}\|},$$

59 after that the WG–index

$$WG_X = \frac{1}{N} \sum_{k=1}^{K} max \left\{ 0, n_k - \sum_{i \in I_{c_k}} R(x_i) \right\}, \quad (1)$$

62   is calculated representing the WG-index for the set X of points partitioned into K clusters (Desgraupes, 2017).

63   2)   For the BIC index:

64   Let us model each cluster c_k as a multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$, where μ_k can be estimated as the

65   sample mean vector and $\Sigma_k = \frac{1}{d(N-K)} \sum_{k=1}^{K} \sum_{i \in I_{c_k}} \|x_i - g_k\|$, can be estimated as the sample covariance matrix.

66   Hancock (2017) showed that the optimal clustering is presented by maximum

$$BIC(C_K) = \sum_{k=1}^{K} (n_k \log \frac{n_k}{N} - \frac{d\, n_k}{2} \log 2\pi \Sigma_k - \frac{n_k - 1}{2} d) - \frac{1}{2} K(d+1) \log N. \quad (2)$$

3 Data and processing

69 This section presents a description of the polarimetric radar data used by the hierarchical algorithm in this study, and some

70 details of the data pre-processing that is applied before the QVP calculation-processing. Note that the method presented here

71 is generally applicable with similar multivariate data from other sources. In addition, in situ observations from the Facility for

72 Airborne Atmospheric Measurements (FAAM) are presented in this section. These data will be used for assigning the labels

73 of the hydrometeor classes to the detected clusters or cluster groups in the radar observations.

3.1 X-band radar observations

75 The polarimetric data employed to demonstrate the method developed in this study were collected by the NXPol radar whilst

76 it was located at Chilbolton Atmospheric Observatory (CAO), part of the UK's National Centre for Atmospheric Science's

77 Atmospheric Measurement and Observation Facility (AMOF), in southern England (Lat. 51.145° N, Long. 1.438° W) from

78 November 2016 to May 2018 (Fig. 1). The NXPol is a modified mobile Meteor 50DX (Leonardo Germany GmbH) X-band,

79 dual-polarization, Doppler weather radar. The radar is a magnetron-based system and operates at a nominal frequency of 9.375

80 GHz ($\lambda \sim 3.2$ cm). The detailed characteristics of the NXPol radar can be found in Neely et al. (2018). From the observations

81 made in 2017–2018 we selected eight dates with the longest precipitation events occurring above the 30 km circle area around

82 the radar presented on Fig. 1. The exact dates and the total number of volume scans per date can be found in Table 1.

83 **3.2 Polarimetric variables and temperature data**

84 Here we chose to use the polarimetric variables Z_H [dBZ], Z_{DR} [dB], ρ_{HV} [–], and K_{DP} [° km^{-1}] as well as temperature T [°

85 C] to demonstrate the described clustering technique. The four polarimetric variables were selected as a subset of all the

86 possible variables as they provide complementary information about the observed hydrometeor properties. Temperature was

87 added to the set of input variables following the reasoning of similar studies in which either height relative to 0° C–isotherm

88 (Grazioli et al., 2015) or the index representing ice- or liquid-phase of observed precipitation (Besic et al., 2016) was included.

89 The full input vector used in this study can be represented as:

90
$$x = [Z_H, Z_{DR}, \rho_{HV}, K_{DP}, T], \quad (3).$$
91

92 Note that this does not preclude the use of differing sets of variables in future studies. The input data used here were pre-

93 processed before being utilized in the clustering algorithm: all range bins that were located at distances less than 400 m from

94 the radar were removed from all input variables data to reduce the influence of side lobe noise.

95

96 The temperature data were taken from the Met Office Unified Model (UM) and interpolated onto the polar grid of the radar's

97 observations (Hampton et al., 2019). Past assessments of the accuracy of these temperature values suggest that the gridded

98 temperatures are within 1° C of co-incident profiles measured by radiosondes except in the case of strong inversions or frontal

99 boundaries (Hampton et al. 2020). An example of one day's observations (2017-05-17) represented in the height-vs-time

00 format of the QVPs of four polarimetric variables with the temperature presented as isotherms is found in Fig. 2. Similarly,

01 other dates from our list of cases (Table 1) are in Fig. B1.

02 **3.3 QVPs and thresholding**

03 QVPs of the input variables are obtained as the azimuthal average of the data from a standard plan position indicator (PPI)

04 scan at 20° antenna elevation angle (Ryzhkov et al., 2016). The 20° PPI is the highest of ten PPIs of the volume scanning

05 strategy used by NXPol which starts the scanning from 0.5° elevation angle. The use of 20° PPI minimizes the effects of radar

06 beam broadening and horizontal inhomogeneity. The beam broadening effect becomes dominant at higher altitudes when

07 observed by low elevation scan as was shown in Ryzhkov et al. (2016). The radar beam of 1° opening at 20° elevation is about

08 100 m at 2-km height, 240 m at 5-km height, and reaches almost 480 m at 10-km height. The resulting profiles have 197 voxels

09 in each QVP at the altitudes between 0 and 10 km above mean sea level (AMSL) and covering about 30 km range from the

10 radar. It was also shown in Ryzhkov et al. (2016) that the decrease of Z_{DR} from the oblate spheroidal hydrometeors at 20°

11 elevation is within the common measurement error of Z_{DR}(0.1–0.2 dB).

12 An advantage of QVPs is that they reduce statistical errors within the input dataset while the height-vs-time format of QVPs

13 naturally represents the temporal dynamics of microphysical processes observed in the radar data. To ensure the observations

14 are representative of large-scale meteorological features that may be averaged together, the QVP voxels are used in the analysis

15 only if more than 270 of the 360 azimuthal bins at the range in the PPI scan contain valid data.

3.4 In situ Observations

17 For the labelling of the clusters, in situ observations can be used to assess any of the clusters within the hierarchical structure

18 produced by the clustering algorithm. This allows for flexibility in the granularity used to examine the observations. In this

19 study, the in situ FAAM observations are used to demonstrate the labelling and assessment of the final set of clusters. The

20 FAAM is a publicly funded research facility that, as part of the National Centre for Atmospheric Science (NCAS), supports

21 atmospheric research in the United Kingdom by providing a large instrumented atmospheric research aircraft (ARA) and the

22 associated services. The ARA is a modified British Aerospace 146-301 aircraft. Further details of the FAAM aircraft

23 instrument systems are available at https://www.faam.ac.uk/the-aircraft/instrumentation/. In situ data for this study comes from

24 FAAM flights C013, C076, C081, and C082 (FAAM, 2017; FAAM, 2018a; FAAM, 2018b; FAAM, 2018c). The dates of the

25 flights with corresponding flight numbers can be found in Table 1.

26 For the cases examined, FAAM was equipped with two Cloud Imaging Probes (CIP), which are manufactured by Droplet

27 Measurement Technologies and described in Baumgardner et al., 2001. The CIP's are mounted underneath the aircraft wings

28 and provide 2-bit grayscale images of cloud particles as they pass through the instrument sample volumes. Each CIP houses a

29 64 elements photodiode detector, with one CIP having an effective pixel size of 15μm (referred to as CIP15), and the other

30 having an effective pixel size of 100μm (referred to as CIP100). Therefore, the CIP's provide images of particles in size ranges

31 of 7.5 μm to 952.5 μm for CIP15 and 50 μm to 6350 μm for CIP100. All probes have 'Korolev' anti-shatter tips – the width

32 of which is 70 mm for the CIP100 and 40 mm for the CIP15.

33 Particle size distributions are calculated based on CIP data where particle size is defined as being the maximum recorded length

34 in either the axis of the detector array (X) or along the direction of motion (Y). All particles with inter-arrival times $< 10^{-6}$ s

35 are rejected as indicative of shattering as in Field et. al. 2006. Particle concentrations are calculated from the sample volume

36 using the centre-in approach (Heymsfield and Parrish, 1978) is used to calculate the size of partially imaged particles. It should

37 be noted that there are significant uncertainties associated with the derived properties from the CIP's, which are of the order

38 20% for number-based properties (Baumgardener et al., 2017). In our analysis we are not concerned with absolute

39 concentrations from the CIPs. Instead, we are using the CIP data in a qualitative manner to provide a general framework for

40 comparison with the HC results obtained from the radar observations.

41 As we base our clustering on the QVP of radar observations, which at each range from the radar are averaged over all available

42 azimuths, a direct comparison to aircraft observations taken at an exact position and timestamp would not make a representative

43 comparison. Thus, for the comparison, we have selected the 20-second intervals from the CIP15 and CIP100 data that

44 correspond to the spatial domain and times of the individual of the 20° PPI scans that are used to create the QVPs. Over these

45 20 s intervals the mean number concentrations per particle size bin are calculated (Fig.10). Fig. 9 presents examples of particle
46 imagery from the CIPs, which typically represents less than 1s of total 20s data and shows derived properties of the particles
47 over the entire 20s sampling period when the airplane observed the atmosphere over the QVP domain.

48 In order to provide insight into the nature of the particle imagery, we have separated the particle concentrations in Figure 10
49 into three categories, two of which are based on an analysis of the particle shapes, and another category for partially imaged
50 particles. "Round" particles are those which have a circularity between 0.9 and 1.2, and particles with a larger circularity are
51 labelled as "Irregular" - this gives a rough separation into particles which are likely to be liquid water vs ice (Crosier et al
52 2011). "Edge" particles are those which are only partially images, as indicated by pixels at the extreme edge of the array being
53 triggered. For the particles which are considered round, the particle size and subsequent concentration is corrected for out-of-
54 focus effects (Korolev 2007). This out-of-focus correction is not applied to the "irregular" or "edge" images, as there is no
55 evidence to show this is an appropriate correction to make. When interpreting the CIP data, it should be noted that out-of-focus
56 effects can alter the measured shape of the particle, which can result in "irregular" shaped ice particles appearing "round",

4 Clustering of QVPs

58 Here the clustering steps are described in general. A corresponding overview of the approach is provided in Fig. 3. The
59 proposed approach uses QVP voxels and temperature data that has been interpolated to the same volume. This data forms the
60 points of a 5-dimensional space ($d = 5$). The PCA (Sect. 2.2) reduces the number of dimensions to q. The q-dimensional data
61 are partitioned into K clusters, where K iteratively increases ($K = 2, 3, ...$) until the optimal number of clusters is reached,
62 according to the WG index Eq. (1) in Sect. 2.4. With each of the clusters achieved in this level ("Outer Loop" in Fig. 3), the
63 process is recursively repeated starting with the PCA calculation and continuing until the optimal partitioning of the sub-
64 clusters is reached ("Inner Loop" in Fig. 3). The total partitioning is confirmed with the BIC index Eq. (2) in Sect. 2.4. When
65 the BIC's local maximum is reached the partitioning is considered to be optimal. A detailed description of each of these steps
66 can be found in the following subsections.

4.1 Start of hierarchical clustering

68 The hierarchical clustering starts with data standardization and dimensionality reduction of the original 5-variable input data
69 X Eq. (3) into a q-dimensional dataset of principal components (Sect. 2.2). To start the loop, all N pixels of the input data are
70 used. In later loops, only subsets of the original 5-variable data (I_{c_k}) belonging to active cluster c_k are processed in the Inner
71 Loop (Fig. 3). The first q principal components with the largest variance, having in total at least 85 % representativity of the
72 original dataset, are selected in this step.

73 The representativity threshold of 85 % was chosen arbitrarily as it reduces the initial 5-variable input space up to 3-dimensions
74 ($q \leqq 3$) in most cases, which effectively simplifies the clustering problem and does not influence the overall outcome. The
75 threshold can be reduced to further reduce the dimensionality, but it was found that this negatively influences the clustering

76 accuracy. A higher threshold will retain the high dimensionality of the original dataset but will slow down the clustering

77 process without gaining further information from the dataset.

4.2 Iterative process to find the optimal number of clusters (Inner Loop)

79 At the start of the hierarchical clustering, we begin directly with the first call of the Inner Loop (Fig. 3). The iterative process

80 in the Inner Loop commences with all N QVP pixels represented by the first q principal components. The spectral clustering

81 processes these input data starting with the number of clusters $K = 2$. The number of clusters increases ($K = 2, 3, …$) with

82 each cycle of spectral clustering within the Inner Loop, and at the end of each iteration the WG index Eq. (1) is calculated for

83 the achieved clustering C_K. At the moment the local maximum is achieved in the WG index values, the clustering in which it

84 was reached (C_K') is accepted as the main cluster set of the current level of the hierarchical tree and these clusters become the

85 set of active clusters (A). Set A will be used in the Outer Loop of the implemented hierarchical algorithm. The active clusters

86 detected in the first level of the hierarchical structure by spectral clustering for the data on 2017-05-17 are shown in Fig. 4.

4.3 Optimal number of clusters for the total dataset (Outer Loop)

88 In the Outer Loop of the hierarchical algorithm, the BIC index Eq. (2) is calculated for the active clusters produced by the

89 Inner Loop (Fig. 3). If the BIC index is calculated for the first time (i.e. the start of the algorithm run, $j = 1$) or the BIC index

90 values do not show any local maximum, the algorithm continues by calling the Inner Loop for each individual cluster from the

91 set of active clusters (A_j) formed by the calls of the Inner Loop described above.

92 For the first level of hierarchical clustering all C_K' are immediately accepted as active $A_1 = C_K'$. In the Outer Loop, after

93 calculating the BIC, the original 5-dimensional data belonging to each cluster $c_k' \in C_K'$ are sent to the Inner Loop and clustering

94 (C_K'') achieved by the Inner Loop is used to replace the cluster c_k' in the set of active clusters A_1. If the BIC index calculated on

95 this 'suggested set' shows that the clusters introduced to the A_1 increase the value of BIC, the suggested replacement is accepted

96 and the set of active clusters is updated as $A_2 = C_K'' \cap C_K'/c_k'$. The Outer Loop then continues with the next cluster from the

97 original set A_1. When the BIC value does not increase with the 'suggested set' the set of active clusters does not change, and

98 the algorithm continues with the next cluster from the set A_1.

4.4 Next recursion or finalization of results

00 The final set of active clusters is reached when the value of the BIC index does not increase with any further suggested split

01 in the current active set of clusters. At this level of detailization, the optimal clustering for the provided input data has been

02 reached. For the QVP dataset described in Sect. 3, a final set of 13 active clusters is reached (see Fig. 5 and Fig. 6). The

03 relations between these final clusters (f_cl1, ..., f_cl13) and the 3 parent clusters from the first Inner Loop run (Fig. 4, panel

04 (a)) are shown in Fig. 5.

5 Labelling

Once the optimal number of clusters is determined and the hierarchical clustering structure is built, the clusters can be characterized by their centroids and labelled with appropriate hydrometeor classes using the available verification data. The clusters for which direct verification data are not available may still be labelled with an appropriate hydrometeor class based on the scattering characteristics described by the original polarimetric radar variables and considering their position in the hierarchical tree and height-vs-time QVP representation. Labelling the obtained clusters can be performed for the different levels of granularity depending on the user's needs and interests. Note the purpose of the labelling shown here is to demonstrate the ability and validity of the technique rather than performing a rigorous study of the underlying microphysics observed. The latter will be reserved for follow-up studies utilizing this technique in a focused manner.

5.1 Level by level clusters check

From the visual verification of the first level parent clusters in Fig. 4 panel (a) and panel (b) we can deduce that there are two child clusters representing the upper/elevated (ice dominated root.cl2) and the lower (water dominated root.cl1) parts. The second level clusters from the second loop (panels (c) and (d)) show a well-identified "bright band", belonging to the melting layer (ML), (root.cl2.cl2) and a main solid-phase cluster (root.cl2.cl1), both belonging to the cluster representing the ice-phase dominated part (root.cl2) of the QVP (panel (b)). The three child clusters of the parent root.cl1 cluster (panels (c) and (d)) are the two rain-type clusters (root.cl1.cl2 and root.cl1.cl3) below the "bright band" and cluster root.cl1.cl1 with most points located above the "bright band". In further loops, the main ice-phase cluster (root.cl2.cl1) is split into nine child clusters (Fig. 5) and examples of their positioning in time-vs-height format of QVP can be observed in Fig. 6 or Fig. B1.

5.2 Characteristics of the clusters

The 13 final clusters can be characterized by their centroids (Fig. 7) or their relevant statistics (Fig. 8 and Table A1). The centroid characteristics in Fig. 7 are plotted as spider plots where each of the five variables is represented by an azimuthal axis. The filled pentagons in each subplot represent the cluster's centroid in the five-variable space based on all the data available in this study. Each vertex of the pentagon shows the centroid's value in one of the five variables. The non-solid lines in the subplots of Fig. 7 represent the centroids of the same cluster but based solely on the data from one of the eight cases (Table 1).

Figure 7 confirms a distinction made at the first and the second cycles of the Outer Loop (Fig. 4) between three types of clusters: liquid-phase clusters (f_cl1, f_cl2 and f_cl3), having lower K_{DP} and warmer T values, ice-phase clusters (f_cl4, f_cl5, f_cl6, f_cl7, f_cl8, f_cl9, f_cl10, f_cl11), and f_cl12 all with more pronounced ρ_{HV} values; and a very different looking f_cl13, having warmer T and rather low ρ_{HV} values.

The largest differences between the centroids selected from the total dataset and the centroids corresponding to the eight considered cases occur in the temperature (T) values, especially for the clusters f_cl2, f_cl3, f_cl7, f_8, f_cl10 and f_cl12.

36   These variations can be explained by the origin of the temperature data, which are estimates from the NWP model and not

37   always correctly represent the real situation.

38   The next variable with a rather large variation in several clusters is K_{DP}. Clusters f_cl1, f_cl7, f_cl11, f_cl12 and f_cl13 have

39   the highest variation in K_{DP} values between the centroids calculated on different cases (Fig. 7). As K_{DP} can be influenced by

40   the amount of ice/water particles in the atmosphere, it might be that the clusters have variations in the number of particles.

41   This hypothesis can only be verified with FAAM observations of the same cluster on different dates. Unfortunately, such

42   verification is not possible for all clusters and more in situ observations are required.

43   From the variations of centroid values in the five input variables in Fig. 7, we can also see that the main liquid-phase cluster

44   (f_cl2) has rather different characteristics in different cases. Case to case it shows large variations in Z_H, K_{DP}, ρ_{HV}, T and the

45   highest mean temperature value (6.8° C) among all other clusters. As observed in the histogram of percentages of the cluster

46   points in Fig. 8 panel (f), f_cl2 is rather big (13.7 % of the total number of points), but does not have the highest percentages

47   of points in all eight analyzed cases, only in the 2017-05-17 case (9.3 % compared to ≤ 1.5 % in other cases) (Fig. 7, histogram

48   per case). Combining all these aspects together we deduce that f_cl2 includes rain of varying intensities and different drop

49   sizes.

50   The other rain cluster f_cl3 has less variability in centroid values and has ~20 dBZ smaller Z_H values than f_cl2 probably due

51   to the smaller drop sizes in this cluster. f_cl3 has the smallest mean Z_H (7.18 dBZ) and mean K_{DP} values (-0.097 ° km^{-1}) of

52   all "water" clusters (f_cl1, f_cl2 and f_cl3). This cluster is often observed at the beginning and at the end of the storm in height-

53   vs-time format representation of the optimal clusters (Fig. B1) and is labelled as "light rain".

54   Almost all centroids in Fig. 7 have no or very limited variation in ρ_{HV} or Z_{DR} values except for f_cl13. This cluster has no

55   variation in temperature (T), having all centroids at 0° C. As such, f_cl13 corresponds to the "bright band" of the ML.

56   According to the box and whiskers plots in Fig. 8 the "bright band" cluster f_cl13 has the highest mean Z_H (24 dBZ), mean

57   Z_{DR} (0.99 dB), and the lowest mean ρ_{HV} (0.93) compared to the other optimal clusters and it is mostly located near 0° C. These

58   characteristics immediately indicate that f_cl13 can be labelled as the "bright band" cluster.

59   The "bright band" cluster (f_cl13 in Fig. 6) is observed to have some sagging areas: between 10:00 (UTC) and 12:20 (UTC),

60   around 16:00 (UTC) and near 18:00 (UTC). Note that f_cl1 is observed above the "bright band" cluster f_cl13 exactly at these

61   time intervals (Fig. 6). This sagging "bright band" signature is often observed where aggregation and riming processes are

62   occurring directly above the ML (Kumjian et al., 2016 and Ryzhkov & Zrnic, 2019). This suggests that f_cl1 can be associated

63   with the processes of aggregation and riming and labelled accordingly.

64   Looking at the percentage of points belonging to each cluster in the optimal clustering set (Fig. 8, panel (f)), we see that clusters

65   f_cl7 and f_cl12 have less than 2 % of points and most probably represent some sporadic and/or special conditions. These

66   clusters also have been separated early from the other "low ice" (f_cl4, f_cl5, f_cl6) and "elevated ice" (f_cl8, f_cl9, f_cl10,

67   f_cl11) clusters and are located near the top of the hierarchical tree (Fig. 5). Both clusters f_cl7 and f_cl12 have smaller

68   absolute mean Z_{DR} values (-0.062 dB and 0.097 dB correspondingly) than the other "ice" clusters (Fig. 8, panel (b).). f_cl7 is

69   also characterised by the highest mean K_{DP} value (0.44 ° km^{-1}) among all of the clusters (Fig. 8, panel (d)). The combination

70   of rather high Z_H (17 dBZ) and high K_{DP} indicates potentially a cluster with high particle number concentration of small ice

71   crystals mixed with a small amount of bigger aggregates. f_cl12 has the lowest mean Z_H value (-3 dBZ) from all optimal

72   clusters. Combining these low mean Z_H with low mean Z_{DR} (0.097 dB) we can assume that f_cl12 can be labelled as low

73   particle number concentration of very small ice crystals.

74   f_cl11 belongs to the "elevated ice" clusters and in most cases (see Appendix, Fig. B1) appears as a column in the height-vs-

75   time representation (around 07:00 UTC in Fig. 6 or in Fig. B1 panels (a) 07:00–08:00 UTC; (c) 05:00 UTC and 07:30 UTC;

76   (g) 12:00 UTC; (h) 12:00–12:30 UTC) filling all the altitudes from the top of the cloud to the ML. This cluster has low mean

77   Z_H value (3 dBZ), one of the highest mean Z_{DR} values (0.92 dB) and a close to zero mean K_{DP} value (-0.009 ° km^{-1}). f_cl11

78   can be labelled as the pristine ice crystals class, as they typically have high aspect ratios (≫1) and tend to fall preferentially

79   with their major axis aligned horizontally (Keat and Westbrook, 2017).

80   Clusters f_cl8, f_cl9 and f_cl10 belong to the "elevated ice" branch of the hierarchical tree (Fig. 5). Among these clusters,

81   f_cl9 has the most different characteristics compared to f_cl8 and f_cl10 clusters. f_cl9 has higher mean Z_H (11 dBZ) and Z_{DR}

82   (0.59 dB) in combination with a lower ρ_{HV} (0.97). Areas with similar characteristics in the height-vs-time representation of the

83   QVPs were labelled as dendritic growth layer (DGL) by Ryzhkov et al. (2016) and Trömel et al. (2019). f_cl8 and f_cl10 have

84   rather similar characteristics to each other (Fig. 7 and Table A1). The small difference between these two clusters is in a higher

85   mean Z_H (8.6 dBZ) and K_{DP} (0.14 ° km^{-1}) for f_cl8 compared to 5.8 dBZ and 0.028 ° km^{-1} for f_cl10. Both clusters are the

86   main "elevated ice" clusters. f_cl10 has a warmer mean temperature (-8.2° C compared to -16.7° C for f_cl8) and most of the

87   time is located near f_cl8 at the beginning or at the end of the observed event (Fig. B1).

88   The main "low ice" clusters are f_cl4, f_cl5 and f_cl6. Clusters f_cl5 and f_cl6 are often observed together with f_cl6 located

89   above f_cl5. The f_cl4 has several appearances in height-vs-time formats of events (see Appendix, Fig. B1, e.g. 09:00 UTC

90   panel (a); 11:00 UTC panel (b); 09:00 and 17:00-18:00 UTC panel (e)), mostly above f_cl1 reaching higher altitudes in the

91   data.

92   Of course, not all clusters can be labelled with absolute confidence and in situ observations can help to verify these initial

93   suppositions.

5.3 Clusters versus in situ observations

95   For the verification of the preliminary labelling made in Sect. 5.2, data from the CIP15 and CIP100 on board the FAAM is

96   utilised. FAAM aircraft flights were performed on four out of the eight days (Table 1) of radar observations: 2017-05-17,

97   2018-01-24, 2018-01-24 and 2018-02-14. The flight altitudes and the timestamps when the aircraft was inside the QVP domain

98   can be observed on the height-vs-time representations of the optimal clusters in Fig. B1.

99   Out of the four available flights, there are 23 periods of 20-second intervals which result in a total of 460 seconds of flight

00   time when the aircraft was inside the QVP domain and a cluster can be assigned to the corresponding height. Of these 23

01 periods there are observations corresponding to 9 unique clusters (f_cl1, f_cl2, f_cl3, f_cl4, f_cl5, f_cl6, f_cl8, f_cl10, f_cl12).

02 This samples 70 % of the final clusters. From these time series, we present examples of CIP15 and CIP100 images for each

03 cluster (Fig. 9) and mean particle size distributions of the data observed during the 20-second interval (Fig. 10).

04 The liquid-phase clusters f_cl1, f_cl2 and f_cl3 correspond to in situ data which contains relatively high concentrations (>1

05 000 m^{-3}) of small particles (mostly < 200 μm in size) which appear round (Fig. 9 and Fig.10, panels (a)–(c)). This strongly

06 supports the idea of large amounts of liquid water being present in the cloud, which supports the labelling of f_cl1-f_cl3 as

07 being influenced by liquid water hydrometeors. When looking in detail at CIP100 in situ observations, and to some extent

08 CIP15 observations at sizes > 200 μm, we can see some significant difference between f_cl1-f_cl3, which we will now discuss.

09 f_cl1, observed above the "bright band", was previously assigned to be the result of aggregation/riming. In this region, the

10 CIP100 (Fig. 9, lower part of panel (a)) shows particle imagery and particle size distributions segregated by shape which show

11 the presence of large ice particles (Fig. 10, panel (a)), again confirming the previous cluster labelling. The CIP100 data shows

12 the presence of irregularly shaped particles, ranging in size from ~1-4 mm, with concentrations in each bin of the order 1 m^{-3}.

13 This suggests a mode of snow particles is present at the same time as the previously mentioned liquid droplet mode. Many

14 small water droplets in the CIP15 observations (Fig. 9 and Fig.10, CIP15 of panel (a)) could indicate either the presence of

15 warm cloud processes or small ice crystals melting first at the upper part of the ML. The second interpretation is supported by

16 the imagery from the CIP100 which suggests melting has not started to occur on the larger particles. In this case, the larger

17 aggregate snowflakes fall to lower altitudes before they start to melt and form the clear "bright band" in the QVP.

18 f_cl2 was characterized by the strong variation in Z_H, K_{DP}, ρ_{HV}, and T of the cluster's centroids in different cases. Both CIP15

19 and CIP100 have small round shape particles in the corresponding images (Fig. 9, panel (b)). The mean concentrations per

20 particle size distributions (Fig. 10, panel (b)) show the prevalence of particles recognised by shape as water droplets. The

21 droplets of < 2 mm size have the occurrences of the order from 10 m^{-3} to 90 000 m^{-3} with higher orders corresponding to

22 particle sizes < 200 μm. Summing up previous analysis and in situ observations we can assign f_cl2 to a "liquid" cluster, which

23 includes rain of varying intensities and different drop sizes.

24 f_cl3 also has predominantly small round shape particles (mean size μ = 128 μm) in the CIP15 panels (Fig. 9 and Fig.10,

25 upper part of panel (c)). The CIP100 data (Fig. 9, lower part of panel (c)) were not processed due to technical issues with the

26 probe so water/ice concentrations based on this data are unfortunately not available. The high concentrations (1000-5000 m^{-3})

27 of small size (< 200 μm) particles are assigned to water (Fig.10, panel (c)). Concentrations of the larger particles (> 200 μm

28 and < 800 μm) are very low (< 100 m^{-3}). Considering these observations, the cluster's characteristics and the fact that f_cl3

29 appears mostly in the beginning or at the end of the events (Fig. B1) we can assume that the cluster either represents a very

30 light rain/drizzle or indicates a partially filled QVP domain in the original data.

31 The CIP15 images for f_cl4 show a mix of larger irregularly shaped particles (aggregates of snowflakes) and relatively few

32 tiny ice crystals (Fig. 9, panel (d)) having very small irregular shapes. The mix of small (< 1 mm) particles recognised as water

33 and ice has low total concentrations (500-800 m^{-3}; Fig.10, upper part of panel (d)). Particles of larger sizes (> 1 mm and < 3.0

34   mm) have concentrations from 100 m^{-3} to 800 m^{-3} (Fig.10, lower part of panel (d)) and were recognised as ice due to their

35   irregular shapes. The small number of "round" particles recognized as liquid are likely an artefact of the data processing due

36   to out-of-focus imaging of the numerous ice particles which are present - such artefacts appear when particles are observed at

37   the edges of the depth of field (O'Shea et al, 2019). Accordingly, this cluster can be assigned to the mix of pristine ice and

38   some formed aggregates, all having low concentrations.

39   f_cl5 and f_cl6 show very similar images in CIP15, but the CIP100 images illustrate the difference between these two clusters

40   (Fig. 9 panels (e) and (f)). The shape analysis of CIP100, suggests both clusters include a low concentration (40-50 m^{-3}) of

41   small round shape particles (< 1 mm size; the lower part of panels (e) and (f), Fig. 10). Similar to f_cl4, the number of liquid

42   particles are likely an artefact of the data processing (O'Shea et al, 2019). The main difference between these clusters is shown

43   in the part of the data recognized as ice. The mean size of the particles of f_cl5 is about 1.7 mm and has occurrences of order

44   100 m^{-3} (Fig. 10, panel (e)). While the mean size of particles in f_cl6 is a bit smaller - 1.4 mm and particles around that size

45   have a higher occurrence of order 400 m^{-3} (Fig. 10, panel (f)). This difference between f_cl5 and f_cl6 resembles the

46   aggregation processes when dendritic crystals of higher concentrations formed at higher altitudes (f_cl6) start to clump together

47   during their fall and form aggregates (f_cl5) with a lower concentration of particles.

48   Clusters f_cl8 and f_cl10, belonging to the "elevated ice" branch of the hierarchical tree (Fig. 5), are also represented by the

49   very similar images of CIP15 and CIP100 observations (Fig. 9, panels (g)–(h)) with the difference in the particles'

50   concentration (Fig. 10, panels (g)–(h)). f_cl8 has a higher concentration of small size (<1 mm) particles (up to 1000 m^{-3})

51   recognised as spherical than f_cl10 (up to 300 m−3). The bigger particles captured in CIP100 corresponding to f_cl8 are of

52   bigger size (from 1 mm and up to 3 mm) and also have a higher concentration (< 500 m^{-3}) compared to particles of f_cl10

53   having as maximum size particles of about 2 mm with concentrations < 200 m^{-3}. The example of CIP100 images suggests that

54   these particles are dendritic in nature (Fig. 9, panel (h)). Again, similar to the artefacts discussed when looking at the CIP

55   observations for f_cl4, f_cl5 and f_cl6, there is likely an erroneous classification of small size particles (O'Shea et al, 2019).

56   Thus, according to in situ data, f_cl8 can be assigned to mix of pristine ice and bigger aggregates and f_cl10 to a low

57   concentration of pristine and smaller aggregates mix.

58   The last in the list of clusters verifiable with the FAAM data is f_cl12 (Fig. 9, panel (i)). The lower panel (i) of Figure 10

59   shows occurrences of order 10 000 m^{-3} for very small particles up to 150 μm and occurrences of order 1000 m^{-3} for the

60   droplets of sizes between 150 μm and 250 μm with almost no occurrences of bigger particles. Unfortunately, concentrations

61   from CIP100 data are not available for this cluster due to technical issues in the CIP100 probe. f_cl12 contains the fewest

62   number of data points in QVP analysis (Fig 8. panel (f)) and there are no other in situ observations related to this cluster. Based

63   on available data, we could assume that the cluster has a high concentration of tiny water droplets (potentially drizzle). On the

64   other hand, similar to the data for other classes, the CIP analysis may have misclassified these as liquid due to their small round

65   appearance in the observations when in reality the observations could represent high concentrations of small ice particles.

66   Likewise, the mean temperature of this cluster is close to 0° C, so no definitive label may be given based on the observations.

67   Thus, the physical interpretation of this cluster is ambiguous though the cluster is separate within the multivariate space.

68   The rest of the clusters need to be assigned by means of human interpretation according to the cluster characteristics or

69   deduction from the interactions and temporal evolution of already assigned clusters. The summary of the assigned clusters can

70   be found in Table A1 of the Appendix A. Application of in situ observations for the assessment of QVP-based clusters has its

71   limits as not all optimal clusters were captured by the FAAM flights and this process requires a comparison of data from

72   essentially one-point measurement to the cluster based on the mean QVP domain values. An appropriate validation process

73   would utilise columnar vertical profiles (CVP) as described in Murphy et al. (2018) with the thorough collocation of the aircraft

74   observations. Utilising CVPs within the presented technique is a part of the planned work for the future.

6 Summary and Conclusions

76   This paper presents a novel technique of hydrometeor classification from QVPs. In this technique, the hydrometeor types are

77   identified from an optimal number of hierarchical clusters, obtained through a recursive process. This recursive process

78   includes an initial dimensionality reduction by principal component analysis followed by spectral clustering. Spectral

79   clustering performed in the PCA space allows us to identify clusters that would have a non-convex form in the original

80   multivariate input space. This property of the algorithm makes it unique and advantageous in comparison to other classification

81   methods, which separate classes by hyperplanes.

82   The final set of clusters is identified with an optimality check using validity indexes. This represents the first attempt, in

83   clustering of weather radar data, to identify the number of clusters based solely on the embedded data characteristics. This

84   data-driven technique produces an optimal number of clusters and keeps the hierarchical structure built in the clustering

85   process. The final set of clusters may be labelled based on their positioning in the hierarchical structure, the characteristics of

86   their centroids and co-incident in situ observations. Depending on the user's needs and interests, the labelling can also be

87   performed for different levels of granularity. In the example shown in this study, we utilise observations collected during

88   several FAAM flights to demonstrate the advantages this technique has in the labelling process. In this case, based on the data

89   available, 70 % of the clusters were labelled using the coincident CIP observations. The other 30 % of the clusters, which were

90   not sampled during the FAAM flights of this study, were labelled based on the cluster characteristics, their positioning in the

91   hierarchical structure and considering interactions with clusters in a height-vs-time format of original QVP data.

92   Thus, in this study, we find that a data-driven HC approach is capable of providing an optimal number of classes from the

93   observations. Moreover, the embedded flexibility in the extent of granularity is the main advantage of the technique. Each

94   branch of the hierarchical structure can be cut out at any level and the parenting cluster characteristics can be used for labelling

95   and identifying more general processes in the atmosphere, while the lower level clusters can provide information about more

96   specific properties and features of the observations.

The centroids of the clusters represent characteristics of the points belonging to a cluster in the multivariate input space (in this case, the polarimetric radar variables and temperature). The identification of these centroids allows the clusters to be tracked in time and altitude as the centroids are calculated based on QVPs from single scans. An analysis of the time series of the radar volume scans is possible and would allow the clusters to be tracked in time and 3D space. Though unexplored in this study, the application of the presented approach in this way could be used to provide information on the temporal evolution of the identified hydrometeors and reveal relationships between the identified classes.

Note that the final set of clusters is optimal only for the provided input dataset, which gives the user an opportunity to select the input dataset depending on their needs. Thus, for the clustering to reflect ice properties and processes, the appropriate input data climatology should be used. For identification of specific features in the data (e.g. birds or insects) a subset of cases potentially including these features should be selected for the analysis. A long-term dataset can be used to create a set of climatologically representative clusters that could be used to study general processes and inform the development of an operational HC scheme.

In this paper, the technique was used for classification of QVPs of long-lasting precipitation events, but the same algorithm can be applied to various needs (e.g. identification of birds, insects' or clustering of volume scans of radar data). In parallel with the application of hierarchical clustering technique to other radar observations a thorough validation of the clusters using CVPs following the FAAM is planned.

Acknowledgments

The authors would like to thank NCAS for support of all scientific activities and, in particular, the FAAM team for collecting the in situ observations. The authors are grateful to Chris Reeds and Graeme Nott for their help with CIP data processing. This work was supported by the Natural Environment Research Council [grant numbers NE/P012426/1, NE/S001298/1] and used JASMIN, the UK collaborative data analysis facility.

References

Abdi, H., Williams, L.J.: Principal component analysis. WIREs Com- put Stat; 2(4): 433-59, 2010.

Aydin, K., Seliga, T.A., Balaji, V.: Remote sensing of hail with a dual linear polarization radar. J. Appl. Meteorol. 25, 1475–1484, 1986.

Baldini, L., Gorgucci, E.: Identification of the Melting Layer through Dual-Polarization Radar Measurements at Vertical Incidence. J. Atmos. Oceanic Technol., 23, 829–839, https://doi.org/10.1175/JTECH1884.1, 2006.

24 Baumgardner, D., Jonsson, H., Dawson, W., O'Connor, D., Newton, R.: The cloud, aerosol and precipitation spectrometer: a
25 new instrument for cloud investigations, Atm. Res., 59–60, 251-264, ISSN 0169-8095, https://doi.org/10.1016/S0169-
26 8095(01)00119-3, 2001.

27 Baumgardner, D., Abel, S.J., Axisa, D., Cotton, R., Crosier, J., Field, P., Gurganus, C., Heymsfield, A., Korolev, A., Krämer,
28 M., Lawson, P., McFarquhar, G., Ulanowski, Z., and Um, J.: Cloud Ice Properties: In Situ Measurement Challenges.
29 Meteorological Monographs, 58, 9.1–9.23, https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0011.1, 2017.

30 Besic, N., Figueras I Ventura, J., Grazioli, J., Gabella, M., Germann, U., and Berne, A.: Hydrometeor classification through
31 statistical clustering of polarimetric radar measurements: a semi-supervised approach, Atmospheric Measurement Techniques,
32 9, 4425–4445, 2016.

33 Desgraupes, B.: ClusterCrit: Clustering Indices. R Package Version 1.2.3., 2013. Available online: https://cran.r-
34 project.org/web/packages/clusterCrit/ (accessed on 6 September 2017).

35 FAAM, 2017: Facility for Airborne Atmospheric Measurements; Natural Environment Research Council; Met Office (2017):
36 FAAM C013 Instrument Test flight: Airborne atmospheric measurements from core and non-core instrument suites on board
37 the   BAE-146   aircraft.   Centre   for   Environmental   Data   Analysis.
38 http://catalogue.ceda.ac.uk/uuid/b1c81845ad294f8399f9b564da59c68c, 2017.

39   FAAM, 2018a: Facility for Airborne Atmospheric Measurements; Natural Environment Research Council; Met Office (2018):
40 FAAM C076 PICASSO flight: Airborne atmospheric measurements from core and non-core instrument suites on board the
41 BAE-146   aircraft.   Centre   for   Environmental   Data   Analysis.
42 http://catalogue.ceda.ac.uk/uuid/dbbf8c12f3ad48309829735f7f2e6359, 2018.

43 FAAM, 2018b: Facility for Airborne Atmospheric Measurements; Natural Environment Research Council; Met Office (2018):
44 FAAM C081 PICASSO flight: Airborne atmospheric measurements from core and non-core instrument suites on board the
45 BAE-146   aircraft.   Centre   for   Environmental   Data   Analysis.
46 http://catalogue.ceda.ac.uk/uuid/64c9279112bb4e0cadb6adaebf1141eb, 2018.

47 FAAM, 2018c: Facility for Airborne Atmospheric Measurements; Natural Environment Research Council; Met Office (2018):
48 FAAM C082 PICASSO flight: Airborne atmospheric measurements from core instrument suite on board the BAE-146 aircraft.
49 Centre for Environmental Data Analysis. http://catalogue.ceda.ac.uk/uuid/25b2a346c6f24032ba746b9dc852ff75, 2018.

50 Field, P.R., Heymsfield, A.J. and Bansemer, A.: Shattering and Particle Interarrival Times Measured by Optical Array Probes
51 in Ice Clouds. J. Atmos. Oceanic Technol., 23, 1357–1371, https://doi.org/10.1175/JTECH1922.1, 2006

52 Grazioli, J., Tuia, D., and Berne, A.: Hydrometeor classification from polarimetric radar measurements: a clustering approach,
53 Atmos. Meas. Tech., 8, 149–170, doi:10.5194/amt-8-149-2015, 2015.

54 Jianbo Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine
55 intelligence, 22(8):888–905, 2000.

Johnson, B.T., Petty, G.W., Skofronick-Jackson, G.: Microwave Properties of Ice-Phase Hydrometeors for Radar and Radiometers: Sensitivity to Model Assumptions. J. Appl. Meteor. Climatol., 51, 2152–2171, https://doi.org/10.1175/JAMC-D-11-0138.1, 2012.

Hall, M.P. M., Goddard, J.W., Cherry, S.M.: Identification of hydrometeors and other targets by dual-polarization radar. Radio Sci., 19, 132-140, 1984.

Hämäläinen, J., Jauhiainen, S., Kärkkäinen, T.: Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. Algorithms. 10. 105. 10.3390/a10030105, 2017.

Hancock, B.: Notes on Bayesian Information Criterion Calculation for X- Means Clustering, 2017.

Hendry, A., Antar, Y.M.M.: Precipitation particle identification with centimeter wavelength dual-polarization radars. Radio Sci., 19, 115–122, 1984.

Heymsfield, A.J. and Parrish, J.L.: A Computational Technique for Increasing the Effective Sampling Volume of the PMS Two-Dimensional Particle Size Spectrometer. J. Appl. Meteor., 17, 1566–1572, https://doi.org/10.1175/1520-0450(1978)017<1566:ACTFIT>2.0.CO;2, 1978.

Keat, W. J., & Westbrook, C. D.: Revealing layers of pristine oriented crystals embedded within deep ice clouds using differential reflectivity and the copolar correlation coefficient. Journal of Geophysical Research: Atmospheres, 122. https://doi.org/10.1002/2017JD026754, 2017.

Korolev, A. V.: Reconstruction of the sizes of spherical particles from their shadow images. Part I: Theoretical considerations. J. Atmos. Oceanic Technol., 24, 376–389, doi:https://doi.org/10.1175/JTECH198, 2007.

Kumjian, M. R., Mishra, S., Giangrande, S. E., Toto, T., Ryzhkov, A. V., and Bansemer, A.: Polarimetric radar and aircraft observations of saggy bright bands during MC3E, J. Geophys. Res. Atmos., 121, 3584– 3607, doi:10.1002/2015JD024446, 2016.

Liu, H., Chandrasekar, V.: Classification of hydrometeors based on polarimetric measurements: Development of fuzzy logic and neuronfuzzy systems, and in situ verification. J. Atmos. Oceanic Technol., 17, 140-164, 2000.

Neely III, R. R. and Bennett, L. and Blyth, A. and Collier, C. and Dufton, D. and Groves, J. and Walker, D. and Walden, C. and Bradford, J. and Brooks, B. and Addison, F. I. and Nicol, J. and Pickering, B.: The NCAS mobile dual-polarisation Doppler X-band weather radar (NXPol). Atmos. Meas. Tech., vol. 11, 12, 6481-6494, doi: 10.5194/amt-11-6481-2018, 2018.

Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14, pp. 849–856. MIT Press, Cambridge, 2002.

O'Shea, S. J. and Crosier, J. and Dorsey, J. and Schledewitz, W. and Crawford, I. and Borrmann, S. and Cotton, R. and Bansemer, A.: Revisiting particle sizing using grayscale optical array probes: evaluation using laboratory experiments and synthetic data. Atmos. Meas. Tech.,vol. 12, 6, 3067--3079, doi:10.5194/amt-12-3067-2019, 2019.

Pelleg, D. and Moore, A.W.: X-means: extending K-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conference on Machine Learning (ICML). Stanford, USA, pp. 727–734, 2000.

89 Ribaud, J.-F., Machado, L. A. T., & Biscaro, T.: X-band dual-polarization radar-based hydrometeor classification for Brazilian
90 tropical precipitation systems. Atmos. Meas. Tech., 12(2), 811–837. https://doi.org/10.5194/amt-12-811-2019, 2019.

91 Roberto, N., Baldini, L., Adirosi, E., Facheris, L., Cuccoli, F., Lupidi, A., Garzelli, A. A.: Support Vector Machine
92 Hydrometeor Classification Algorithm for Dual-Polarization Radar. Atmos., 8, 134, 2017.

93 Ryzhkov, A., P. Zhang, H. Reeves, M. Kumjian, T. Tschallener, S. Trömel, and C. Simmer: Quasi-vertical profiles—A new
94 way to look at polarimetric radar data. J. Atmos. Oceanic Technol., 33, 551–562, https://doi.org/10.1175/JTECH-D-15-
95 0020.1, 2016.

96 Ryzhkov, A. V., Zrnic D. S.: Radar Polarimetry for Weather Observation. Springer Atmospheric Sciences, Springer
97 International Publishing. https://doi.org/10.1007/978-3-030-05093-1, 2019.

98 Straka, J.M., Zrnić, D.S.: An algorithm to deduce hydrometeor types and contents from multiparameter radar data. Preprints,
99 26th Int. Conf. on Radar Meteorology, Norman, OK, Amer. Meteor. Soc., 513-516, 1993.

00 Straka, J. M.: Hydrometeor fields in a supercell storm as deduced from dual-polarization radar. Preprints, 18th Conf. on Severe
01 Local Storms, San Francisco, Amer. Meteor. Soc., 551-554, 1996.

02 Straka, J.M., Zrnić, D.S., Ryzhkov, A.V.: Bulk hydrometeor classification and quantification using polarimetric radar data:
03 Synthesis of relations. J. Appl. Meteor., 39, 1341-1372, 2000.

04 Tiira, J. and Moisseev, D. N.: Unsupervised classification of vertical profiles of dual polarization radar variables, Atmos. Meas.
05 Tech. Discuss., https://doi.org/10.5194/amt-2019-307, 2019.

06 Trömel, S., Ryzhkov, A.V., Hickman,B., Mühlbauer, K. and Simmer, C.: Polarimetric Radar Variables in the Layers of Melting
07 and Dendritic Growth at X Band—Implications for a Nowcasting Strategy in Stratiform Rain. J. Appl. Meteor. Climatol., 58,
08 2497–2522, https://doi.org/10.1175/JAMC-D-19-0056.1, 2019.

09 United States Geological Survey: GTOPO30. Retrieved from http://www1.gsi.go.jp/geowww/globalmap-
10 gsi/gtopo30/gtopo30.html, 2006.

11 Vivekanandan, J., Ellis, S. M., Oye, R., Zrnic, D. S., Ryzhkov, A. V., Straka, J.: Cloud microphysics retrieval using S-band
12 dual-polarization radar measurements. Bull. Amer. Meteor. Soc., 80, 381–388, doi:10.1175/1520-0477(1999)080,0381:
13 CMRUSB.2.0.CO;2, 1999.

14 von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing, 17(4):395–416, 2007.

15 Wang, H., Ran, Y., Deng, Y., and Wang, X.: Study on deep-learning-based identification of hydrometeors observed by dual
16 polarization Doppler weather radars. EURASIP Journal on Wireless Communications and Networking, 179, DOI
17 10.1186/s13638-017-0965-5, 2017.

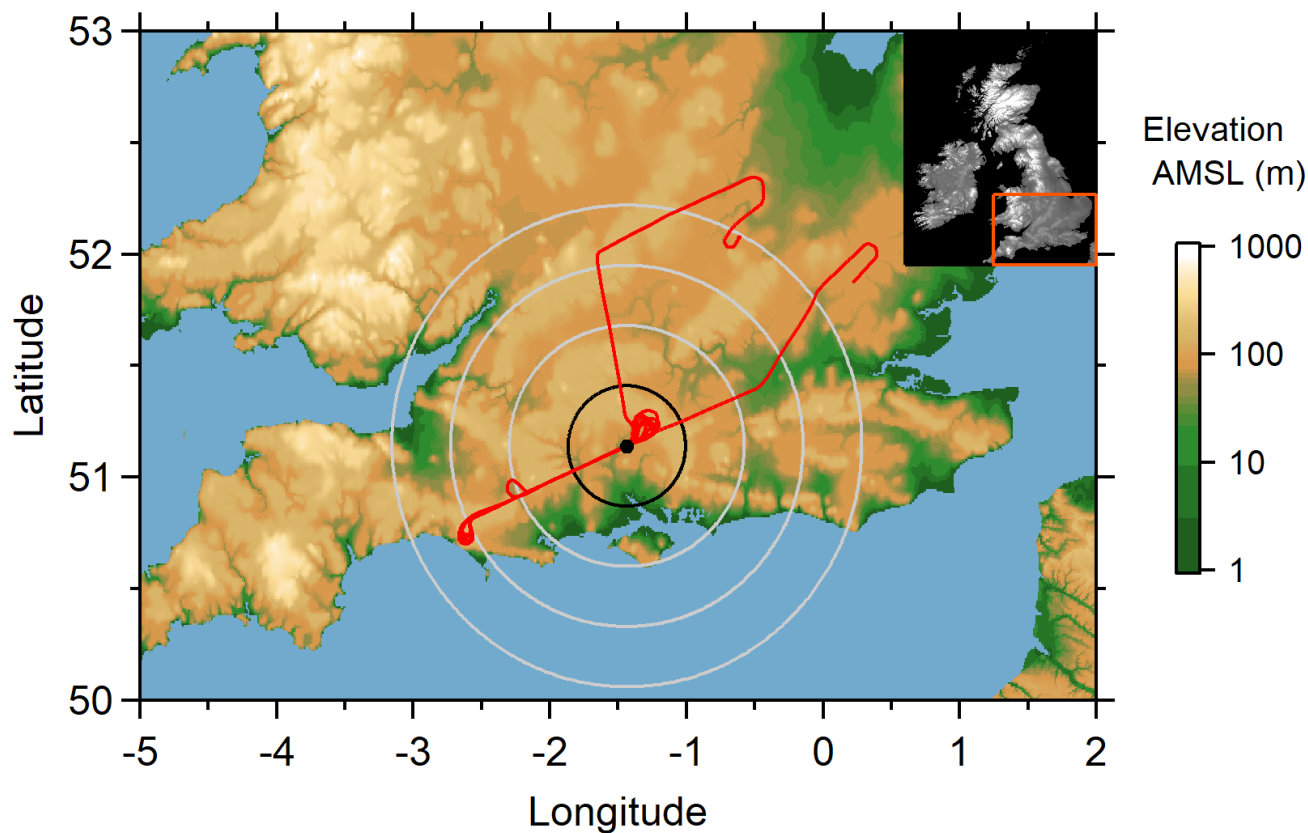



Figure 1: NXPol radar location at the Chilbolton Atmospheric Observatory. Circles with the centre at the radar position represent 30, 60, 90 and 120 km range. Credit: USGS (2006).

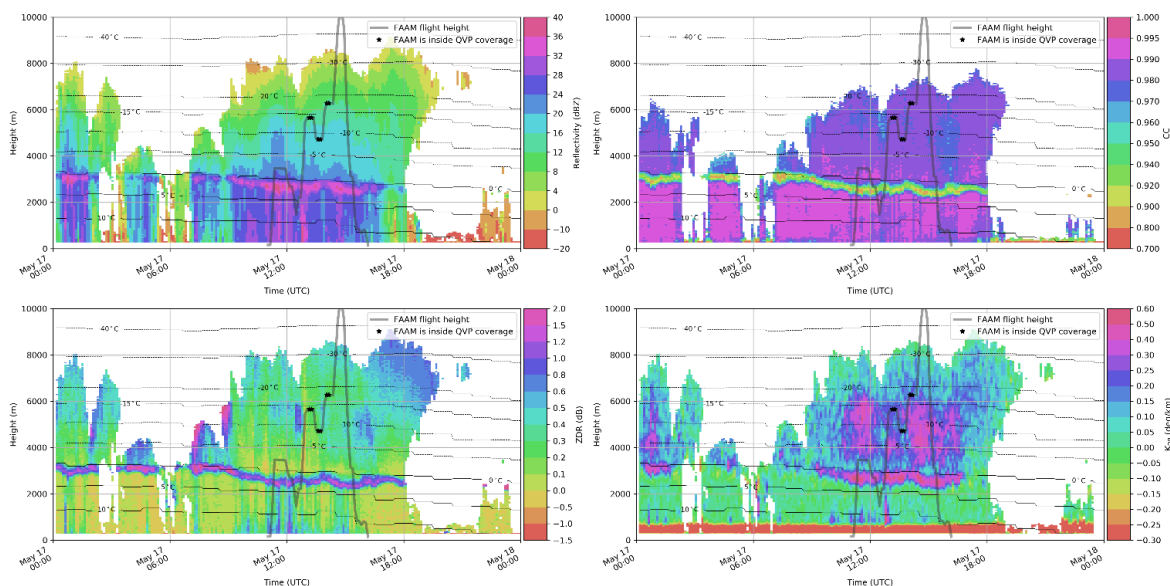



Figure 2: The time-vs-height QVPs of Z_H [dBZ], Z_{DR} [dB], ρ_{HV} [–], and K_{DP} [° km^{-1}] retrieved from the NXPol radar observations at Chilbolton on 2017-05-17. Overlaid by temperature isotherms T [° C].

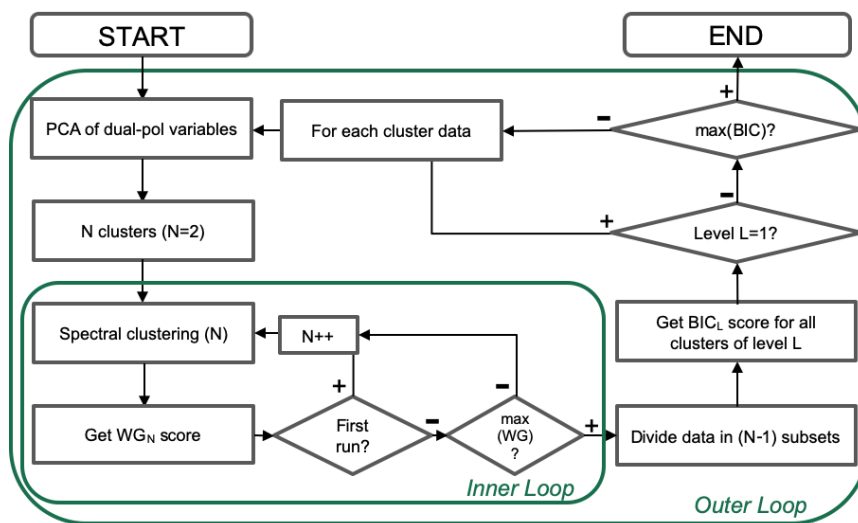



Figure 3: Flow chart of the implemented hierarchical top-down clustering algorithm.

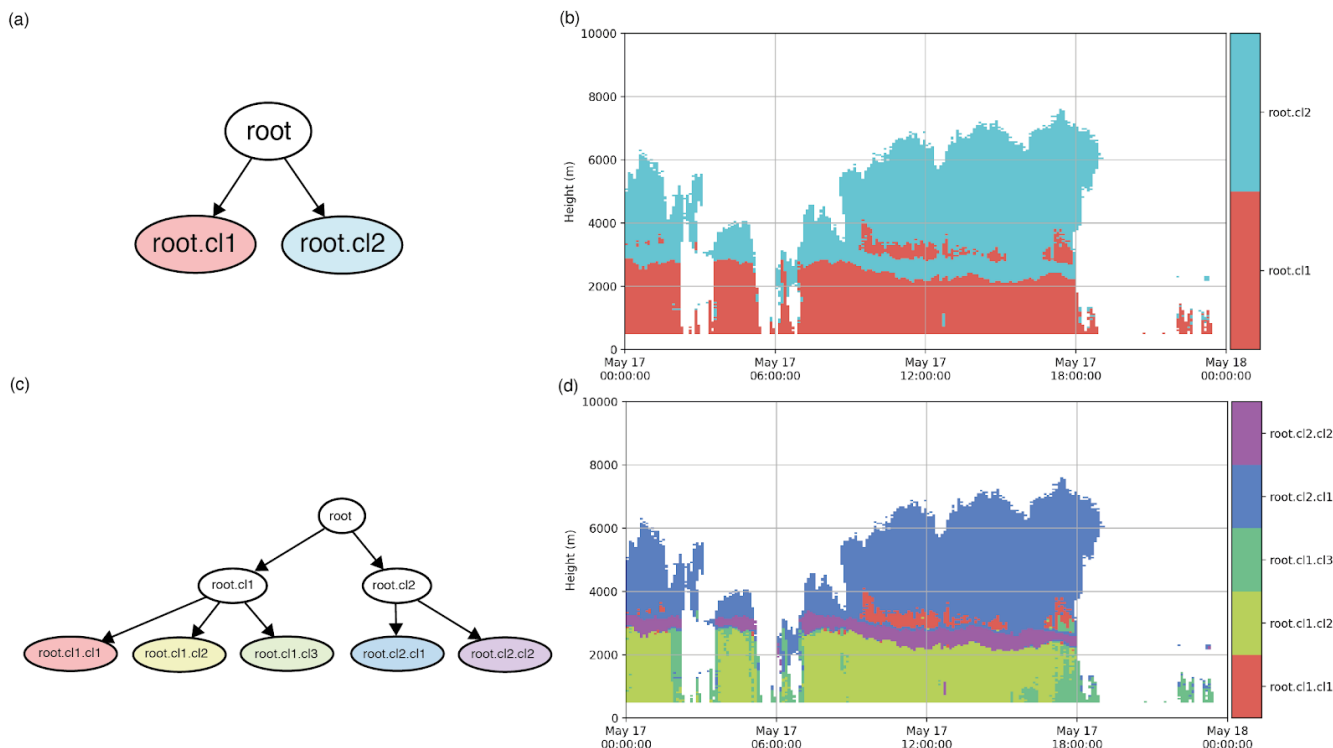



Figure 4: The active clusters at the end of the first (panels (a) and (b)) and second (panels (c) and (d)) cycle of the outer loop of the hierarchical clustering algorithm. Panels (a) and (c) plotted in the hierarchical (tree) structure and panels (b) and (d) plotted in time-vs-height format of the observations on the 17 of May 2017.

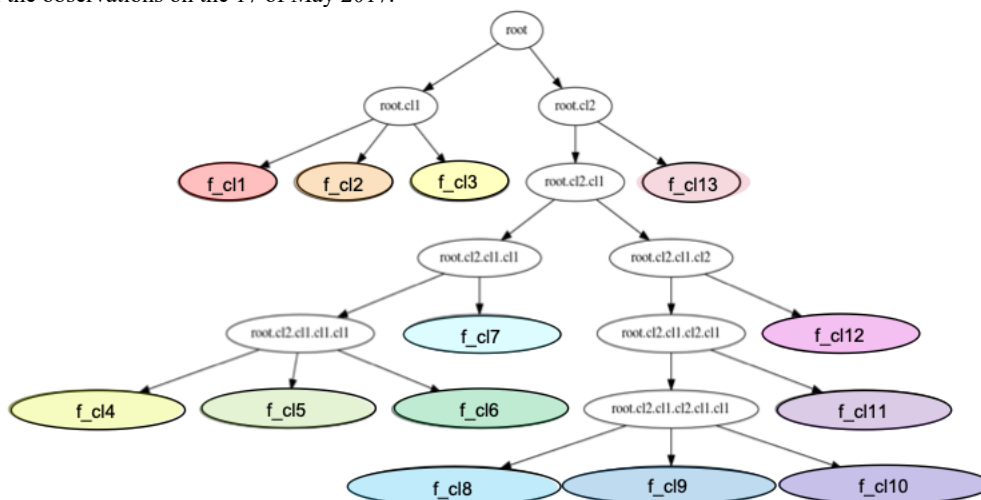



Figure 5: Final hierarchical structure of the optimal clustering found for the QVP input data described in Table 1. The final set of optimal clusters consists of coloured clusters f_cl1, f_cl2, …, f_cl13.

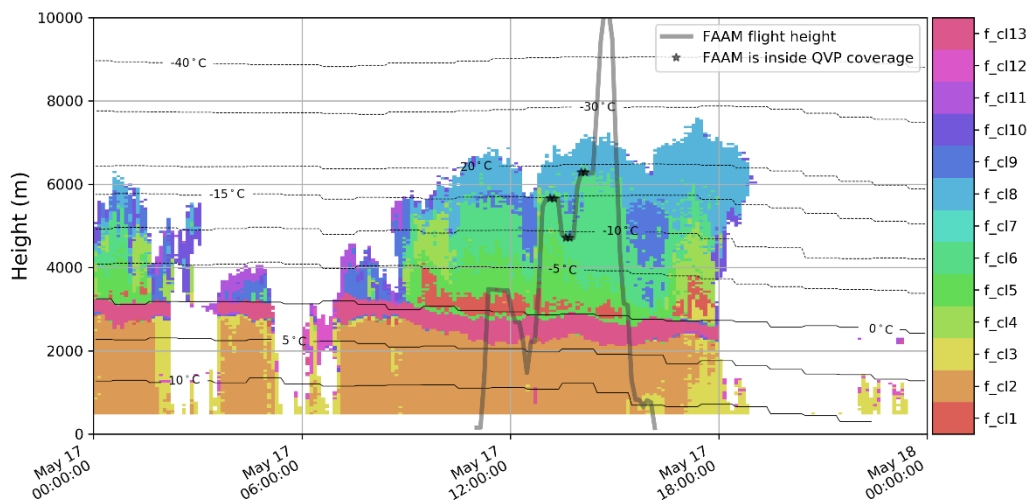



Figure 6: The time-vs-height format of the final optimal set of active clusters found by the top-down hierarchical clustering for the QVP input data described in Table 1 (The hierarchical structure behind the optimal clustering is found in Fig. 4). Example of clusters in the time-vs-height format of the 17 of May 2017 QVPs presented in Fig. 2.

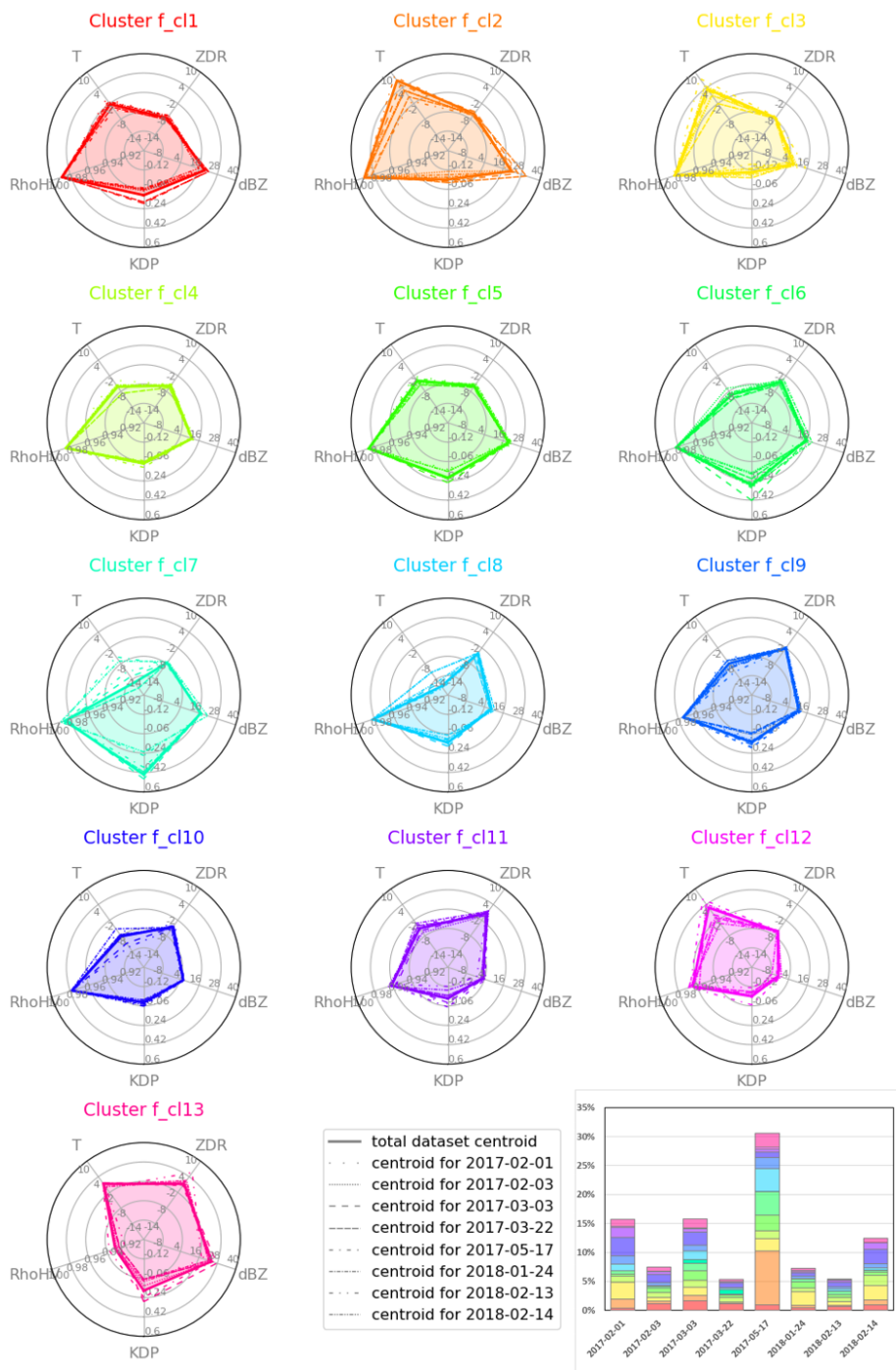



Figure 7: Characteristics of the optimal clustering centroids in four polarimetric variables and temperature. The scales of the variables: from -20 dBZ to 40 dBZ for Z_H, from -1.5 dB to 2.0 dB for Z_{DR}, from 0.9 to 1.0 in ρ_{HV}, from -0.3 ° km^{-1} to 0.6 ° km^{-1} for K_{DP}, and from -20° C to 10° C in temperature (T).

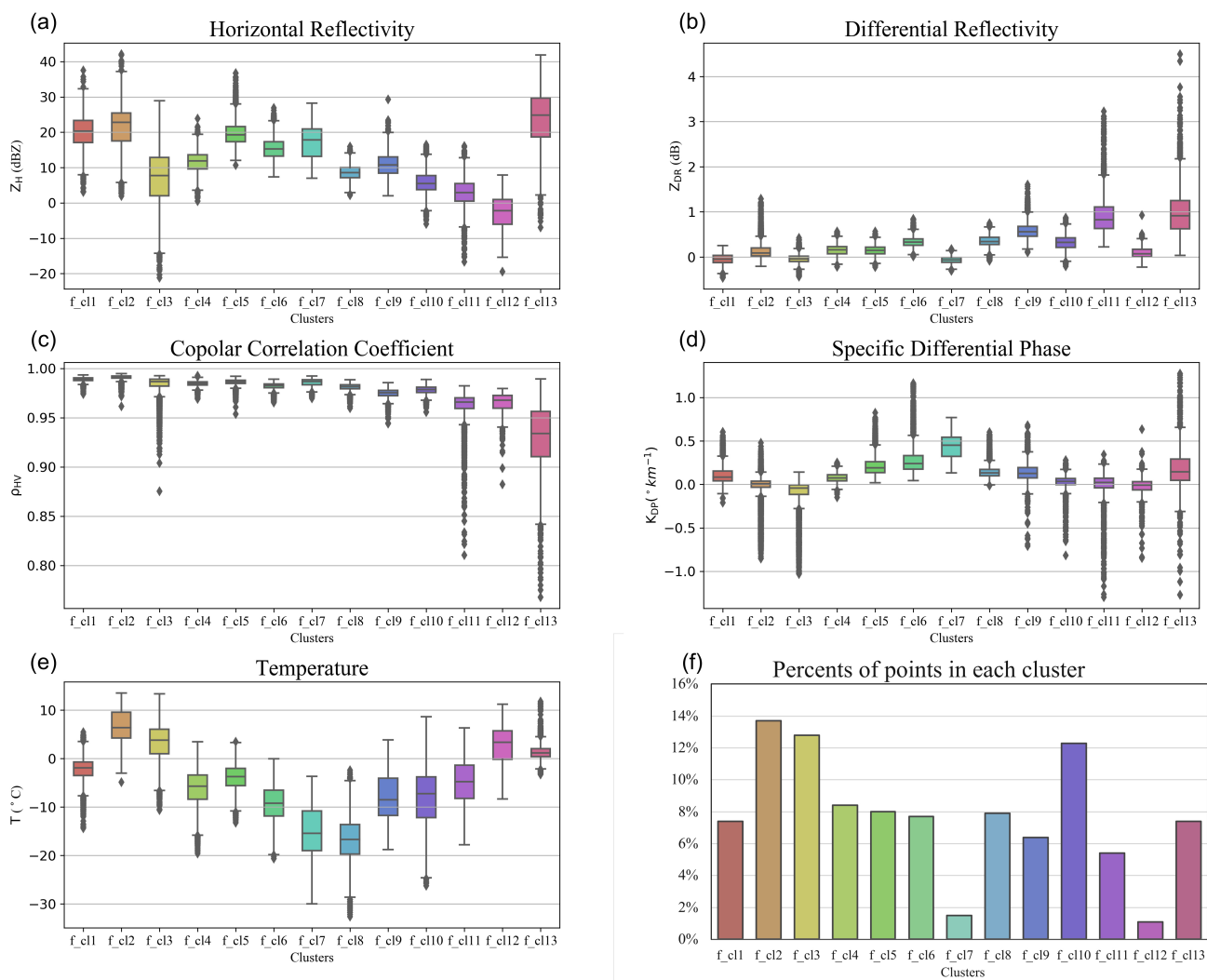



Figure 8: Characteristics of the optimal clustering centroids in four polarimetric variables ((a) – Z_H, (b) - Z_{DR}, (c) – ρ_{HV}, (d) – K_{DP}) and temperature (e). The percentage of points in each cluster is in the panel (f).

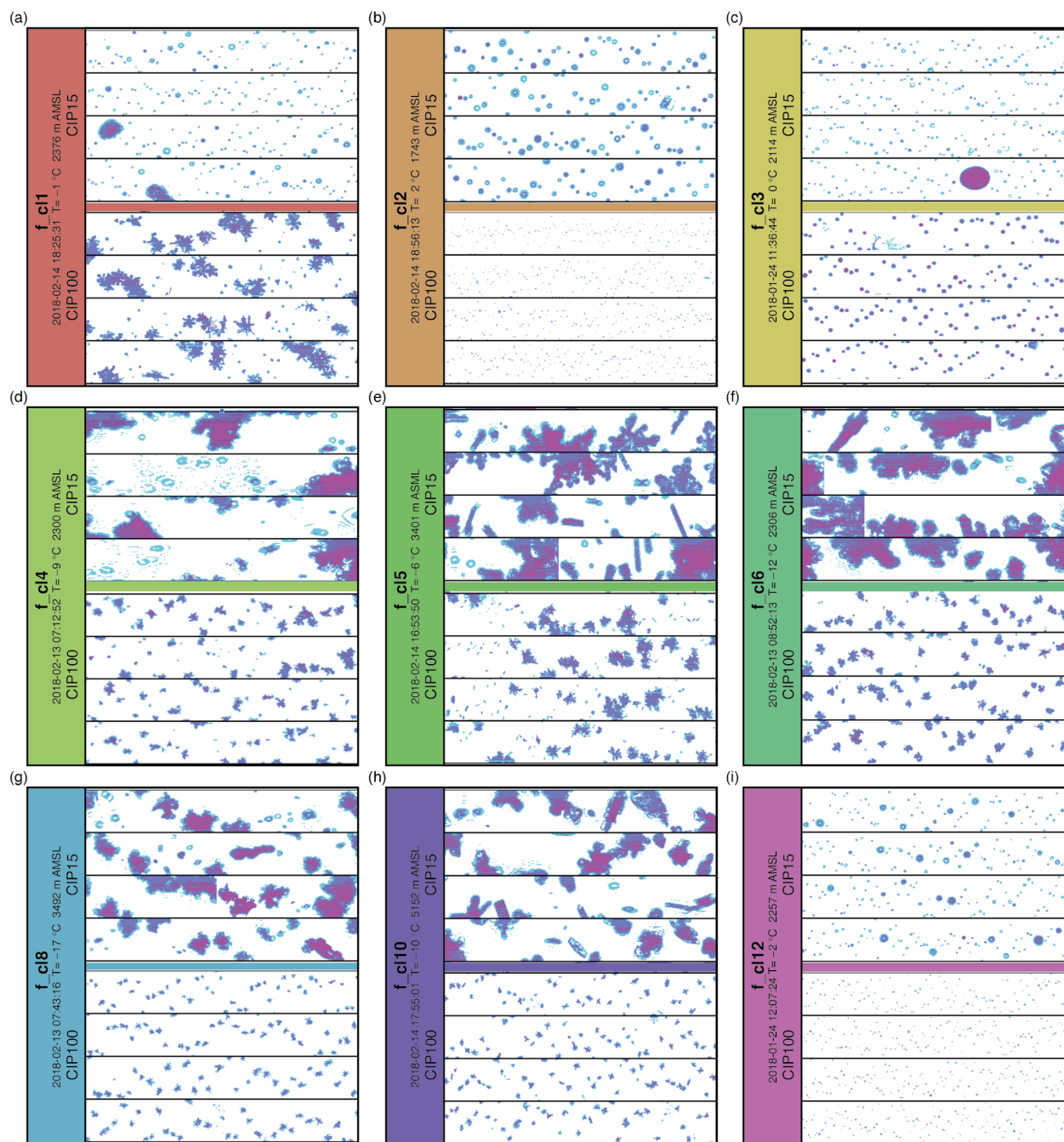
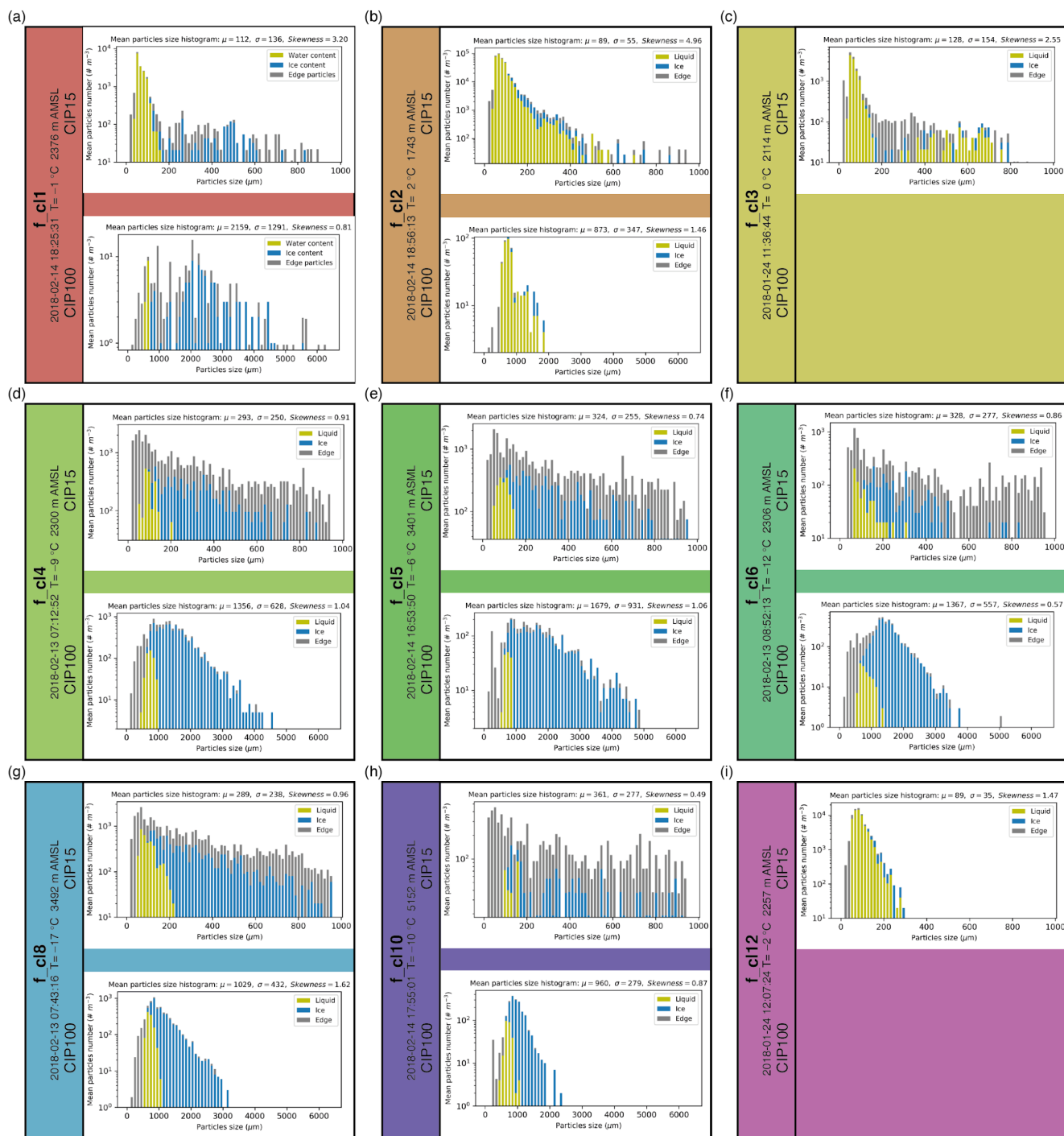



Figure 9: Examples of the images taken from the Cloud Image Probe CIP-15 (15μm, upper) and CIP-100 (100μm, lower) within 30 km range from the radar position: (a) for f_cl1 – on 18:25:31 UTC 14-02-2018; (b) for f_cl2 – on 18:56:13 UTC 14-02-2018; (c) f_cl3 – on 11:36:44 UTC 24-01-2018; (d) f_cl4 – on 18:17:53 UTC 14-02-2-18; (e) f_cl5 – on 16:53:50 UTC 14-02-2018; (f) f_cl6 – on 08:52:13 UTC 13-02-2018; (g) f_cl8 – on 07:43:16 UTC 13-02-2018; (h) f_cl10 – on 17:55:01 UTC 14-02-2018; (i) f_cl12 – on 12:07:24 UTC 24-01-2018.The image widths are 960 and 6400 μm, respectively. The temperature values are derived from the model data and the heights are derived from the location of the clusters in the QVPs.

57
58 **Figure 10:** Corresponding to Fig. 9 particle size distributions from the Cloud Image Probe CIP-15 (15μm, upper) and CIP-100 (100μm,
59 lower) within 30 km range from the radar position: (a) for f_cl1 – on 18:25:31 UTC 14-02-2018; (b) for f_cl2 – on 18:56:13 UTC 14-02-
60 2018; (c) f_cl3 – on 11:36:44 UTC 24-01-2018; (d) f_cl4 – on 18:17:53 UTC 14-02-2-18; (e) f_cl5 – on 16:53:50 UTC 14-02-2018; (f) f_cl6
61 – on 08:52:13 UTC 13-02-2018; (g) f_cl8 – on 07:43:16 UTC 13-02-2018; (h) f_cl10 – on 17:55:01 UTC 14-02-2018; (i) f_cl12 – on

62 12:07:24 UTC 24-01-2018.The image widths are 960 and 6400 μm, respectively. The temperature values are derived from the model data
63 and the heights are derived from the location of the clusters in the QVPs.
64

65 **Table 1: in situ data collection campaigns**

Date	FAAM flight number	Number of volume scans	Number of QVP voxels
20170201	-	243	46656
20170203	-	213	40896
20170303	-	213	40896
20170322	-	213	40896
20170517	C013	100	19226
20180124	C076	196	37632
20180213	C081	189	36288
20180214	C082	189	36288

66

67 **Appendix A: Polarimetric characteristics of the optimal clusters**

68 Table A1 provides the relevant statistics of each of the thirteen optimal clusters identified in this work from a database of X-
69 band radar data.

70 **Table A1.** Statistics describing the content of the thirteen optimal clusters identified in Sects. 4. For each polarimetric variable and for
71 each cluster, we provide the mean value, standard deviation σ, and [minimum, maximum] values.

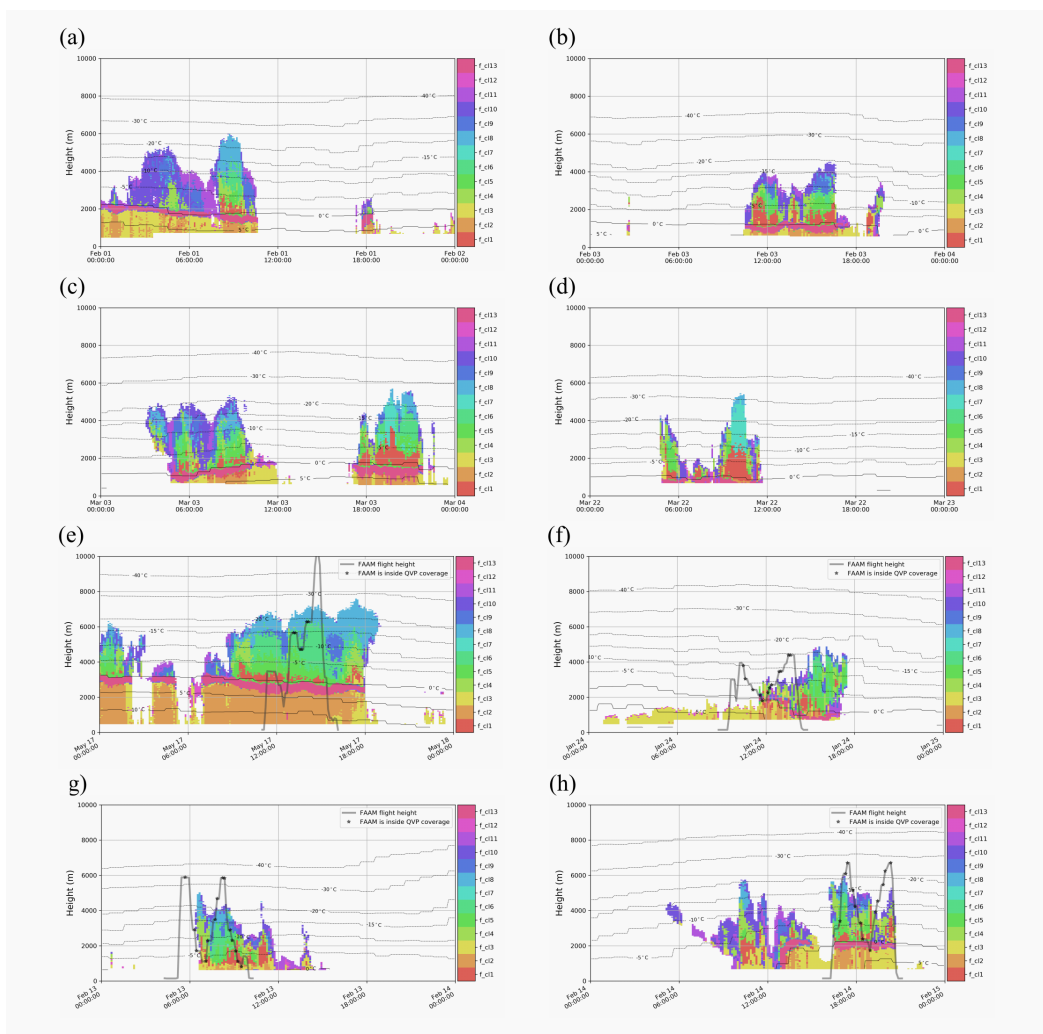
Variable	Cluster	Unit	MeanValue	σ	MinValue	25 %	50 %	75 %	MaxValue	Label
Z_H	f_cl1	dBZ	20	5	3	17	20	23	38	upper part of ML
Z_{DR}		dB	-0.04	0.11	-0.46	-0.12	-0.05	0.03	0.25	
ρ_{HV}		-	0.989	0.002	0.974	0.987	0.989	0.990	0.994	
K_{DP}		$^\circ km^{-1}$	0.1	0.1	-0.2	0.0	0.1	0.2	0.6	
T		$^\circ C$	-2	3	-14	-4	-2	-1	5	
Z_H	f_cl2	dBZ	22	6	2	18	23	25	42	rain
Z_{DR}		dB	0.13	0.17	-0.21	0.02	0.08	0.20	1.29	
ρ_{HV}		-	0.991	0.002	0.961	0.990	0.991	0.992	0.995	

K_{DP}		$°\,km^{-1}$	0.0	0.2	-0.9	0.0	0.0	0.0	0.5	
T		$°\,C$	7	3	-5	4	6	10	14	
Z_H	f_cl3	dBZ	7	8	-21	2	8	13	29	light rain/drizzle
Z_{DR}		dB	-0.04	0.09	-0.43	-0.10	-0.04	0.02	0.43	
ρ_{HV}		-	0.984	0.008	0.875	0.982	0.987	0.989	0.993	
K_{DP}		$°\,km^{-1}$	-0.1	0.2	-1.0	-0.1	0.0	0.0	0.1	
T		$°\,C$	4	4	-10	1	4	6	13	
Z_H	f_cl4	dBZ	12	3	1	10	12	14	24	low concentration pristine ice & aggregates
Z_{DR}		dB	0.15	0.11	-0.22	0.07	0.16	0.23	0.57	
ρ_{HV}		-	0.984	0.002	0.969	0.983	0.985	0.986	0.993	
K_{DP}		$°\,km^{-1}$	0.1	0.1	-0.15	0.04	0.07	0.11	0.25	
T		$°\,C$	-6	4	-20	-8	-6	-3	3	
Z_H	f_cl5	dBZ	20	3	11	17	19	22	37	low concentration larger aggregates
Z_{DR}		dB	0.15	0.11	-0.22	0.07	0.15	0.22	0.57	
ρ_{HV}		-	0.986	0.003	0.954	0.985	0.986	0.988	0.992	
K_{DP}		$°\,km^{-1}$	0.2	0.1	0.0	0.1	0.2	0.3	0.8	
T		$°\,C$	-4	3	-13	-6	-4	-2	3	
Z_H	f_cl6	dBZ	15	3	7	13	15	17	27	higher concentration dendritic crystals & low concentration aggregates
Z_{DR}		dB	0.33	0.11	0.01	0.26	0.33	0.40	0.84	
ρ_{HV}		-	0.982	0.003	0.965	0.980	0.983	0.984	0.989	
K_{DP}		$°\,km^{-1}$	0.3	0.2	0.0	0.2	0.2	0.3	1.2	
T		$°\,C$	-9	4	-21	-12	-9	-7	0	
Z_H		dBZ	17	4	7	13	18	21	28	

Z_{DR}	f_cl7	dB	-0.06	0.08	-0.03	-0.12	-0.06	-0.01	0.18	high concentration pristine ice & low concentration larger aggregates
ρ_{HV}		-	0.985	0.004	0.970	0.984	0.987	0.988	0.992	
K_{DP}		$^\circ\,\mathrm{km}^{-1}$	0.4	0.1	0.1	0.3	0.5	0.5	0.8	
T		$^\circ$ C	-15	6	-30	-20	-15	-10	-4	
Z_H	f_cl8	dBZ	9	2	2	7	9	10	15	high concentration pristine ice & low concentration of dendrites
Z_{DR}		dB	0.35	0.13	-0.08	0.28	0.34	0.44	0.75	
ρ_{HV}		-	0.981	0.003	0.960	0.979	0.982	0.984	0.989	
K_{DP}		$^\circ\,\mathrm{km}^{-1}$	0.1	0.1	0.0	0.1	0.1	0.2	0.6	
T		$^\circ$ C	-17	4	-33	-20	-17	-14	-2	
Z_H	f_cl9	dBZ	11	33	2	8	11	13	29	dendritic growth layer
Z_{DR}		dB	0.59	0.18	0.10	0.46	0.56	0.68	1.61	
ρ_{HV}		-	0.975	0.004	0.944	0.972	0.975	0.977	0.985	
K_{DP}		$^\circ\,\mathrm{km}^{-1}$	0.1	0.1	-0.7	0.1	0.1	0.2	0.7	
T		$^\circ$ C	-8	5	-19	-12	-8	-4	4	
Z_H	f_cl10	dBZ	6	3	-6	4	5	8	17	low concentration of pristine ice & dendrites
Z_{DR}		dB	0.32	0.16	-0.21	0.22	0.32	0.42	0.87	
ρ_{HV}		-	0.978	0.004	0.956	0.976	0.9979	0.981	0.989	
K_{DP}		$^\circ\,\mathrm{km}^{-1}$	0.0	0.0	-0.8	0.0	0.0	0.1	0.3	
T		$^\circ$ C	-8	5	-26	-12	-7	-4	9	
Z_H	f_cl11	dBZ	3	4	-17	1	3	5	16	pristine ice crystals
Z_{DR}		dB	0.92	0.43	0.23	0.63	0.83	1.11	3.23	
ρ_{HV}		-	0.961	0.017	0.810	0.959	0.966	0.970	0.982	
K_{DP}		$^\circ\,\mathrm{km}^{-1}$	0.0	0.2	-1.3	0.0	0.0	0.1	0.3	

T		° C	-5	4	-18	-8	-5	-1	6	
Z_H		dBZ	-3	5	-19	-6	-2	1	8	
Z_{DR}	f_cl12	dB	0.10	0.13	-0.23	0.01	0.07	0.17	0.93	Ambiguous small ice/drizzle
ρ_{HV}		-	0.097	0.012	0.88	0.960	0.968	0.973	0.980	
K_{DP}		° km^{-1}	0.0	0.1	-0.8	-0.1	0.0	0.0	0.6	
T		° C	3	4	-8	0	3	6	11	
Z_H		dBZ	24	8	-7	19	25	30	42	
Z_{DR}	f_cl13	dB	0.99	0.49	0.04	0.63	0.92	0.96	0.99	bright band of the ML
ρ_{HV}		-	0.931	0.032	0.768	0.910	0.934	0.956	0.989	
K_{DP}		° km^{-1}	0.2	0.2	-1.3	0.0	0.2	0.3	1.3	
T		° C	1	2	-3	0	1	2	11	

72
73

74 **Appendix B: The optimal clusters in eight events**

75

76 **Figure B1:** The time-vs-height format of the final optimal set of active clusters found by the top-down hierarchical clustering for the QVP
77 input data described in Table 1. The observations were made on (a) 2017-02-01, (b) 2017- 02-03, (c) 2017-03-03, (d) 2017-03-22, (e) 2017-
78 05-17, (f) 2018-01-24, (g) 2018-02-13, and (h) 2018-02-14. (The hierarchical structure behind the optimal clustering is found in Fig. 4).