AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1029/2018JD029163

Key Points:
- A general and versatile method to automatically identify ice particle habit from airborne probe images is proposed
- The method is successfully tested on three different airborne imaging probes
- Good performance with classification accuracies >90% for each probe

Correspondence to:
A. Berne,
alexis.berne@epfl.ch

Citation:
Praz, C., Ding, S., McFarquhar, G. M., & Berne, A. (2018). A versatile method for ice particle habit classification using airborne imaging probe data. *Journal of Geophysical Research: Atmospheres*, *123*, 13,472–13,495. https://doi.org/10.1029/2018JD029163

Received 12 JUN 2018
Accepted 30 OCT 2018
Accepted article online 8 NOV 2018
Published online 6 DEC 2018

Author Contributions
Conceptualization: C. Praz, A. Berne
Data curation: C. Praz, S. Ding
Funding Acquisition: A. Berne
Methodology: C. Praz, A. Berne
Validation: C. Praz
Writing - Original Draft: S. Ding, G. M. McFarquhar, A. Berne
Formal Analysis: C. Praz
Investigation: C. Praz
Supervision: A. Berne
Visualization: C. Praz
Writing - review & editing: C. Praz

A Versatile Method for Ice Particle Habit Classification Using Airborne Imaging Probe Data

C. Praz[1], **S. Ding**[2], **G. M. McFarquhar**[2,3], **and A. Berne**[1]

[1] Environmental Remote Sensing Laboratory (LTE), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2] Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK, USA, [3] School of Meteorology, University of Oklahoma, Norman, OK, USA

Abstract

A versatile method to automatically classify ice particle habit from various airborne optical array probes is presented. The classification is achieved using a multinomial logistic regression model. For each airborne probe, the model determines the particle habit (among six classes) based on a large set of geometrical and textural descriptors extracted from the two-dimensional image of a particle. The technique is applied and evaluated using three probes with significantly different specifications: the high volume precipitation spectrometer, the two-dimensional stereo probe, and the cloud particle imager. Performance and robustness of the method are assessed using standard machine learning tools on the basis of thousands of images manually labeled for each of the considered probes. The three classifiers show good performance characterized by overall accuracies and Heidke skill scores above 90%. Depending on the application and user preferences, the classification scheme can be easily adapted. For a more precise output, intraclass subclassification can be achieved in a nested fashion, illustrated here with columnar crystals and aggregates. A comparative study of the classification output obtained with the three probes is presented for two aircraft flight periods selected when the three probes were operating together. Results are globally consistent in term of proportions of habit identified (once blurry and partial images have been automatically discarded). A perfect agreement is not expected as the three considered probes are sensitive to different particle size range.

Plain Language Summary

An automatic classification method to identify ice particle habit from images is proposed. The technique is applied and evaluated using three airborne probes mounted on research aircraft with significantly different specifications: the high volume precipitation spectrometer, the two-dimensional stereo probe, and the cloud particle imager. The method relies on thousand of images manually classified and advanced machine learning techniques to determine the snow crystal habit among six preset classes. High classification performance is achieved, with accuracies above 90% for each of the considered probes.

1. Introduction

Clouds play a central role in global climate studies through their direct influence on the Earth-atmosphere radiative budget and water cycle. In mixed-phase and ice clouds, both models and remote sensing retrievals suffer from large uncertainties, partially due to the rich variety of the size and shape that particles can adopt depending on local environmental conditions and particle growth history.

In remote sensing studies, the mass and geometry of distributions of ice crystals cannot be neglected as they directly impact how these particles interact with electromagnetic radiation. Retrieving microphysical properties of ice particles from scattered radiation implies solving an ill-posed inverse problem, which typically lacks constraints (Logvin et al., 2002). In the context of weather radars, scientists have been relying on the use of dual polarizations (e.g., Bringi et al., 2003; Testud et al., 2000), Doppler spectra (e.g., Cooper et al., 2017; Kollias et al., 2007), and more recently multiple frequencies (e.g., Kneifel et al., 2016; Kulie et al., 2014) in order to better constrain radar retrieval algorithms. The procedure requires the definition of a forward model simulating the scattering properties of an ensemble of hydrometeors. For ice-phased particles, these properties strongly depend on the size, mass, and morphology of individual targets, as reported by various numerical simulation studies (Baum et al., 2005; Leinonen et al., 2012; Um & McFarquhar, 2007, 2011). It is therefore essential to collect precise in situ data that can be used as a reference to refine ice crystal scattering simulations and reduce uncertainty in radar-based retrievals.

In this context, the information provided by ice crystal and snowflake imaging devices is of primary importance as they give insight into the dimension, concentration, type, and morphology of individual particles. These sensors can be deployed either on the ground or on research aircraft. Ground-based instruments are attractive because they are usually easier and cheaper to operate and can be deployed for a long duration. On the other hand, airborne imaging probes provide more insight to study the physical mechanisms at play and capture microphysical transitions (e.g., phase transition, as in Cober et al., 2001, and aggregation, as in Bailey & Hallett, 2012) as they directly sample within clouds. As these devices, commonly called optical array probes (OAPs), can collect up to several thousands of cloud particle images per minute, there is a need for automatic algorithms to process the resulting millions of images. Initially, simple particle geometrical features extracted from shadowgraph images (e.g., size, area, and perimeter) were used to develop the first hydrometeor classifiers for OAP images (Holroyd, 1987; Hunter et al., 1984; Moss & Johnson, 1994). These first studies based on dimension-related descriptors and decision-tree approaches were, however, not able to identify composite habits like aggregates or bullet rosettes. Korolev and Sussman (2000) proposed to use more advanced particle shape descriptors and classified ice crystal habit into four classes based on dimensionless ratios of geometrical measures. McFarquhar et al. (1999) applied a neural network algorithm to identify more complex habit like bullet rosette and polycrystal based on particle dimension and area ratio. Feind (2006) performed a comparative study to assess the performance of various classification algorithms and concluded that although a neural network was giving the best classification accuracy, the relevance of the particle descriptors utilized is crucial and more important that the classification method. With the progress in imaging techniques, new OAPs with higher pixel resolution became available such as the cloud particle imager (CPI) for which a habit classification program was developed (Lawson, Baker, et al., 2006). More recently, principal component analysis was applied to CPI images to identify eight distinct cloud habits with an accuracy higher than 80% (Lindqvist et al., 2012).

Ice cloud particle habit identification based on OAP images proved to be very insightful for various microphysical studies such as investigating the composition of different types of cloud (Korolev et al., 2000), documenting the microstructural properties of individual particles (Baum et al., 2005; Korolev & Isaac, 2003), and relating those with their mass and terminal velocity (Heymsfield et al., 2002, 2004; Heymsfield & Westbrook, 2010). They have also been used to estimate single-scattering properties of different habits and conduct cloud radiative simulation studies (e.g., Um & McFarquhar, 2007, 2009). In the visible spectrum, proportions of classified habits have been directly correlated with scattering phase functions measured with a polar nephelometer and showed a link between the observed dominant habits and the peaks in the phase function (Lawson, Baker, et al., 2006).

In the present contribution, a simple and efficient classification algorithm that can be applied to a broad range of OAP devices with various specifications (e.g., pixel resolution, imaging technique, and sampling volume) is proposed. As described in Praz et al. (2017), the approach was initially developed for the Multi-Angle Snowflake Camera (MASC), a ground-based snowflake imager that captures high-resolution photographs of falling snowflakes from three different angles (Garrett et al., 2012). On MASC data, the classifier achieved high accuracy (>90%) in determining the hydrometeor type among six classes, estimating the degree of riming ranging between zero (no riming) and one (graupel), and identifying if the particle was melting or not. The classification is achieved by means of a multinomial logistic regression (MLR) and compared to other notorious machine learning methods like support vector machine (SVM) and artificial neural networks (ANNs). Logistic regression is a well-established classification method (Bishop, 2006) and has successfully been applied in atmospheric research, for example, for statistical downscaling of precipitation (Fealy & Sweeney, 2007) and for improving probabilistic forecast of precipitation amounts (Wilks, 2009). As an extension to binary logistic regression for multiclass problems, MLR is a probabilistic model that assigns to each observation probabilities of belonging to different classes. MLR is a supervised model, meaning that it relies on a set of labeled data, called the training set, to identify discriminating features and in turn assign a class to new data samples. In the context of ice particle imagery, this means that a habit classification scheme and a training set composed of labeled images have to be defined beforehand.

This article applies MLR to identify ice cloud particle habit from OAP images. In contrast to previous research, the classification method is applied and evaluated on three different OAPs with a common classification scheme applied to all. The algorithm makes use of innovative particle features introduced for the purpose

Table 1
Overview of the Particle Imaging Probes Used in This Study

Instrument	Pixel resolution (μm/pixel)	Measurement range (μm)	Imaging technique	# of views	Image type
2D-S	10	10 – 1,260	linear PDA	1	binary
HVPS	150	150 – 12,000	linear PDA	1	binary
CPI	2.3	2.3 – 1,000	2D-PDA	1	256 levels gray scale
MASC	35	35 – 10,000	2D-PDA	3	256 levels gray scale

Note. The MASC, the instrument the classification method was originally developed for, is also included. Measurement ranges for the 2D-S, HVPS, and CPI are reported from Baumgardner et al. (2011). 2D-S = two-dimensional stereo; HVPS = high volume precipitation spectrometer; CPI = cloud particle imager; MASC = Multi-Angle Snowflake Camera; 2D-PDA = two-dimensional photodiode array.

of this study in addition to a large variety of descriptors already used in previous studies. A feature selection algorithm is implemented in order to identify and retain only the most relevant ones. A dedicated effort to evaluate the classifier performance and generalization properties (by means of cross validation and learning curves) is also presented. Compared to existing classification methods, an attractive property of the proposed approach is that it does not rely on any manually fixed threshold and can therefore be adapted to new problems (e.g., new field campaign, OAPs with varying resolution, and modifications of the number of classes) at a minor cost (rerun the training phase on new/extended data). The remainder of the manuscript is structured as follows. Section 2 describes the three probes utilized, the Olympic Mountain Experiment (OLYMPEX) field campaign during which data were collected and the image processing techniques used to extract the particle descriptors. The classification model and the procedure introduced to assess the classification performance are presented in section 3. Classification results are analyzed for the three probes independently and then compared on two common flight periods in section 4. The work is summarized, and key results are emphasized in a conclusion drawn in section 5.

2. Data and Methods

2.1. Ice Cloud Particle Images

Various airborne sensors have been developed for collecting information on the size, mass, and concentration of ice cloud particles. An overview of these instruments is given in Baumgardner et al. (2011, 2017). Some of them also provide additional information on the particle shape and habit as they capture two-dimensional images. These devices are commonly classified as OAPs due to their measurement system based on optical arrays. The OAPs utilized in this study are the two-dimensional stereo (2D-S) probe, the high volume precipitation spectrometer (HVPS), and the CPI. The 2D-S (Lawson, O'Connor, et al., 2006) and HVPS (Lawson et al., 1993) use a linear photodiode array scanning at a rate adapted to the particle velocity in order to reconstruct a two-dimensional image of the target based on its shadow. *Shaded pixels* are detected if the light level decreases below a certain threshold (typically 50%), resulting in binary images of ice crystals. These probes require a precise adjustment of the scanning rate proportional to the true air speed and may suffer from particle distortion effects.

The CPI operation principle is different from that of the 2D-S and HVPS in that it relies on a square photodetector array, meaning that an entire image is captured instantaneously when the device is triggered. As a result, the measurement is less sensitive to distortion effects, but discontinuous. The latter point is not critical for the current study but can be an issue in quantitative studies, which require precise estimation of particle concentration (Baum et al., 2005). In this regard, the CPI is more similar to the MASC for which the classification method was initially developed. In contrast to the 2D-S and HVPS, the CPI probe provides 256 level grayscale images, hence giving additional information on the surface structure and transparency of ice crystals. Although background noise is substantial and can vary significantly from one CPI image to another, this textural information is used to calculate additional descriptors, which were relevant for habit classification and riming degree estimation in MASC images (Praz et al., 2017). A summary of the cloud particle imaging probes used in this study as well as their main characteristics is displayed in Table 1. The MASC is also included for comparison.

The present study focuses on ice particle images collected during the OLYMPEX. OLYMPEX was a field campaign whose main objective was to provide ground-based validation support for the Global Precipitation Measurement satellite mission. The mission took place between November 2015 and February 2016 and

Table 2
Summarizing List of the 31 Features Selected for Habit Classification (15 Features for Each Probe With Some Common to Two or Three Probes)

Feature ID	Feature name	Related probe(s)
1	particle maximum dimension	2D-S
2	bounding box maximum dimension	2D-S
3	particle porous area over total area ratio	HVPS and CPI
4	particle area to circumscribed circle area ratio	HVPS and CPI
5	particle area to convex hull area ratio	HVPS, 2D-S, and CPI
6	particle area to bounding box area ratio	HVPS and CPI
7	ratio of particle outline touching frame edge	2D-S
8	fitted ellipse area	HVPS
Largest inscribed ellipse area	CPI	
10	largest inscribed / smallest circumscribed ellipse area ratio	2D-S and CPI
11	fitted ellipse / smallest circumscribed ellipse area ratio	HVPS
12	fitted ellipse eccentricity	2D-S
13	particle area to fitted ellipse area ratio	HVPS,CPI
14	morphological skeleton to particle area ratio	2D-S
15	number of corners in the perimeter	2D-S and CPI
16	standardized distance to centroid Fourier power spectrum comp. P0	CPI
17	standardized distance to centroid Fourier power spectrum comp. P2	2D-S
18	standardized distance to centroid Fourier power spectrum comp. P3	2D-S
19	standardized distance to centroid Fourier power spectrum comp. P6	HVPS, 2D-S, and CPI
20	#max(P0 to P6)	HVPS
21	distance to centroid mean	HVPS
22	distance to centroid standard deviation over mean ratio	HVPS and 2D-S
23	boolean value indicating if the particle is touching the frame	HVPS
24	normalized perimeter average line segment $\bar{s}\left(\frac{1}{12}\right)$	HVPS
25	normalized perimeter average line segment $\bar{s}\left(\frac{1}{2}\right)$	HVPS and CPI
26	normalized perimeter autocovariance acov $\left(s(\frac{1}{12}), 0\right)$	CPI
27	normalized perimeter autocovariance acov $\left(s(\frac{1}{12}), \frac{1}{12}\right)$	CPI
28	normalized perimeter autocovariance acov $\left(s(\frac{1}{4}), 0\right)$	2D-S
29	normalized perimeter autocovariance acov $\left(s(\frac{1}{4}), \frac{1}{12}\right)$	HVPS and 2D-S
30	normalized perimeter autocovariance acov $\left(s(\frac{1}{2}), \frac{1}{12}\right)$	2D-S and CPI
31	Haralick feature homogeneity	CPI

Note. 2D-S = two-dimensional stereo; HVPS = high volume precipitation spectrometer; CPI = cloud particle imager.

included a large number of ground-based and airborne sensors (Houze et al., 2017). The images utilized in this study were collected with OAPs installed on the UND Citation research aircraft. More than 8,000 ice crystal images collected with HVPS (1,410), 2D-S (4,252), and CPI (2,964) were manually selected from this data set and utilized for training the classification algorithm. In order to cover a broad range of environmental conditions, the images were extracted from various flights and time intervals randomly selected.

2.2. Image Processing and Feature Extraction

Image classification and pattern recognition methods typically require a set of variables, commonly called features or descriptors, upon which the algorithm is based. In the case of snow or ice habit classification, these features take the form of numerical values intended to be representative of the size, shape, and internal structure of the particle.

As it is difficult to assess a priori what set of descriptors is most relevant for the classification task, the methodology applied here is to extract as many descriptors from the images as possible and run a feature selection

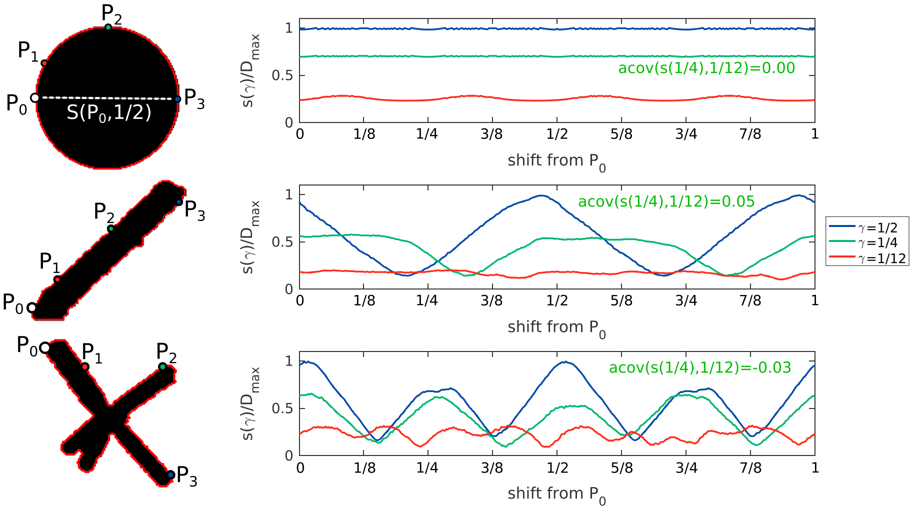

Figure 1. Illustration of the concept of normalized curvilinear distance γ and line segment $s(\gamma)$ for three distinct particle shapes recorded with the two-dimensional stereo probe. On the right panel, the three curves show the evolution of $s(\gamma)$ for three different γ value as the reference point P_i is moved along the particle perimeter. Autocovariance values for $\gamma = \frac{1}{4}$ (green curve) and lag $\phi = \frac{1}{12}$ are also displayed for each particle shape.

algorithm in order to select only the most significant ones (the selection procedure is detailed in section 4.1). In this way it is ensured that the descriptors utilized are not too correlated, which is a desired property for most classification techniques.

For the grayscale images collected by the CPI probe, a total of 111 features was initially calculated based on particle dimensions, morphology, and textural structure. This list includes the 72 descriptors introduced in Praz et al. (2017), 25 descriptors based on the particle corners detection procedure detailed in Lindqvist et al. (2012), and 9 other descriptors implemented for this work. Many of these coefficients had been previously used to describe ice particles in previous studies (e.g., Garrett & Yuter, 2014; Hogan et al., 2012; Nurzynska et al., 2012, 2013; Schmitt & Heymsfield, 2014). For the sake of brevity, only the 15 descriptors preserved after the feature selection procedure are mentioned here. The list is reduced to 31 as some descriptors are common to two or three probes. A list of the features kept for the classification task is displayed in Table 2. Features 1 to 23 and 31 were used for hydrometeor classification in Praz et al. (2017) and are described there. Note that among the 15 descriptors retained for the CPI probe, only feature 31 (Haralick homogeneity, as detailed in Haralick et al., 1973) is effectively using the textural information contained in the grayscale images. Using more advanced image processing techniques to better handle the substantial noise and lack of constant contrast in CPI images may lead to more informative textural descriptors but is beyond the scope of this study.

Features 24 to 30 are introduced in the current classification framework for the first time and rely upon the concept of normalized perimeter defined by Lindqvist et al. (2012). As seven of these features proved to be relevant for habit classification, we describe here how they are established. First, the perimeter of a particle is detected and discretized in such a way that the distance between two adjacent pixels of the perimeter is constant. This distance is then normalized so that the total distance (i.e., the perimeter length) is equal to 1. By doing so,we can define a normalized curvilinear distance $\gamma \in [0, 1]$ from any point P_i of the perimeter, which translates to a distance along the perimeter (γ is called invariant angle and defined between 0 and 360 in Lindqvist et al., 2012). In the case of a circle or a square, a value of $\gamma = \frac{1}{2}$ from a point P_i is directly opposite with respect to the particle centroid. This is not necessarily the case for more complex shapes like bullet rosettes or aggregates, as illustrated in Figure 1.

From this curvilinear distance γ, a line segment $s(P_i, \gamma)$ is defined as the Cartesian distance between two points P_i and $P_{i'}$ of the perimeter separated by γ. Finally, the average line segment for all P_i elements of the perimeter $\bar{s}(\gamma)$ and its autocovariance at different curvilinear distance lags ϕ, denoted acov$(s(\gamma), \phi)$, can be calculated and used to characterize the particle morphology with a limited number of parameters. Mathematically, $\bar{s}(\gamma)$

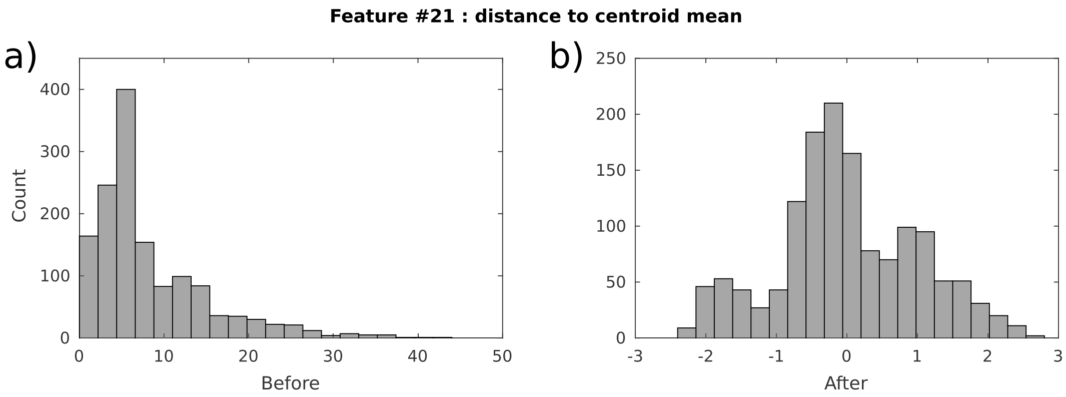

Figure 2. Distribution of feature #21 within the training set used by the high volume precipitation spectrometer classifier before (a) and after (b) the normalization and Gaussianization transforms are applied.

and acov($s(\gamma), \phi$) can be expressed as follows:

$$\bar{s}(\gamma) = \frac{1}{N_P} \sum_{i=1}^{N_P} s(P_i, \gamma), \tag{1}$$

$$\text{acov}\,(s(\gamma), \phi) = \frac{1}{N_P} \sum_{i=1}^{N_P} \left(s(P_i, \gamma) - \bar{s}(\gamma)\right) \left(s(P_i + \phi, \gamma) - \bar{s}(\gamma)\right), \tag{2}$$

where N_P is the number of points within the particle perimeter. Figure 1 provides an illustration of the concept of normalized curvilinear distance and gives values for the line segment for three different particles: a circle, a column, and a bullet rosette imaged by the 2D-S probe. In this study, 3 $\bar{s}(\gamma)$ and 15 acov($s(\gamma), \phi$) were calculated following Lindqvist et al. (2012). After feature selection, seven of them were retained for the classification task, namely, $\bar{s}(\frac{1}{12})$, $\bar{s}(\frac{1}{2})$, acov($s(\frac{1}{12})$,0), acov($s\frac{1}{12}), \frac{1}{12}$), acov($s(\frac{1}{4})$,0), acov($s(\frac{1}{4}), \frac{1}{12}$), and acov($s(\frac{1}{2}), \frac{1}{12}$). For simple geometrical shapes, $\bar{s}(\frac{1}{2})$ is equivalent to an *orientation-averaged* diameter and behaves similarly to the particle equivalent-area diameter. As highlighted by Lindqvist et al. (2012), the concepts of curvilinear distance and line segment allow the identification and regrouping of particles with a similar morphology even if their actual perimeter is significantly different.

2.3. Feature Transformation

When working with a linear model like MLR, it is important that the different dimensions of the problem, in this case the particle descriptors, have a similar range of values. The latter property was fulfilled by normalizing each feature $\boldsymbol{x}_d$ to have 0 mean and 1 variance by applying the transformation

$$\boldsymbol{x}_d^{\text{norm}} = \frac{\boldsymbol{x}_d - \mu(\boldsymbol{x}_d)}{\sigma(\boldsymbol{x}_d)}, \tag{3}$$

where $\mu(\boldsymbol{x}_d)$ and $\sigma(\boldsymbol{x}_d)$ are the mean and standard deviation within $\boldsymbol{x}_d$, respectively. As with many other methods developed in the field of machine learning, MLR utilized in the present work performs better if the input features follow a normal distribution. Even though the latter property is not a necessary condition, the benefits of having nearly Gaussian descriptor distributions are twofold. First, it reduces the amount of outliers that are difficult to deal with the current cost function. In addition, it increases the convergence properties of the optimization algorithm utilized to minimize the cost function, making the classification training phase faster and more robust. For these reasons, the descriptors retained for classification (displayed on Table 2) were passed through a simple Gaussian anamorphosis. If S_d denotes the skewness of the dth feature $\boldsymbol{x}_d$, then the applied transform can be written as

$$\mathbf{x}_d = \begin{cases} \exp(\mathbf{x}_d) & \text{if } \mathcal{S}_d < -1, \\ \mathbf{x}_d^2 & \text{if} -1 < \mathcal{S}_d < -0.75, \\ \sqrt{\mathbf{x}_d} & \text{if } 0.75 < \mathcal{S}_d < 1, \\ \log(\mathbf{x}_d) & \text{if } \mathcal{S}_d > 1. \end{cases} \tag{4}$$

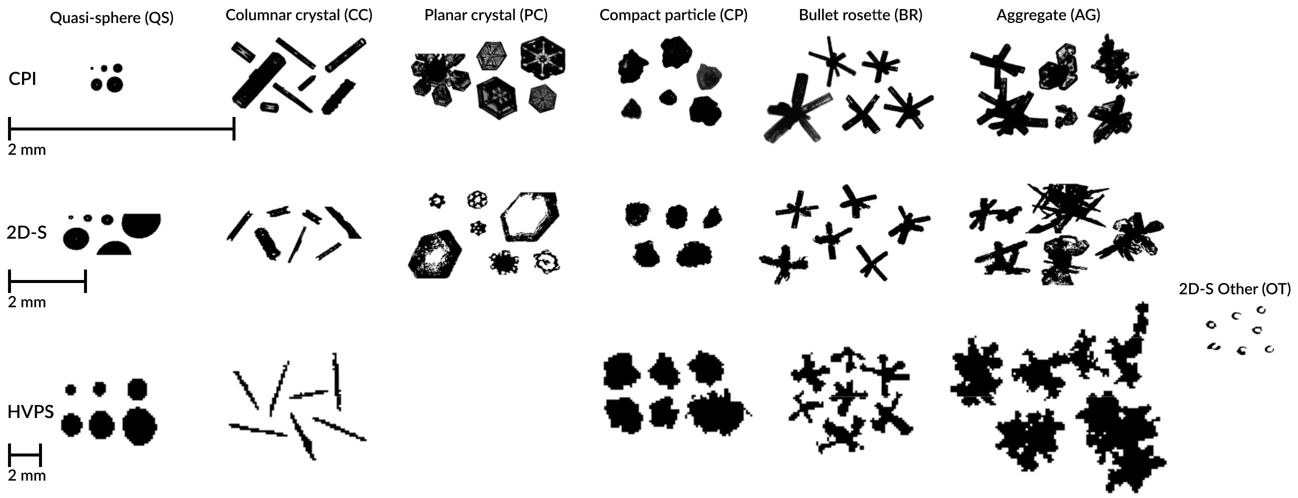

Figure 3. Illustration of the habits used in the classification scheme. From top to bottom, each row displays a few samples per habit, as imaged by the CPI, 2D-S, and HVPS probes. On each row, a pixel resolution scale is provided and highlights the very different range of size covered by the three probes, even within the same crystal habit. Note that planar crystal was discarded from the HVPS classification scheme because not enough representative particles were found in the OLYMPEX data set. CPI = cloud particle imager; 2D-S = two-dimensional stereo; HVPS = high volume precipitation spectrometer.

Applied together, these two transforms significantly increase the classification overall accuracy (OA) for the three probes with a gain varying between 5% and 10%. For instance, Figure 2 illustrates the result of such transforms on feature #21 (distance to centroid mean) utilized by the HVPS classification model. Note that both the normalization and Gaussianization steps are calculated on the pure training set only and then applied on the validation set in order to exclude any influence of validation data on the classifier during the training phase and thus avoid overfitting.

3. Ice Crystals Classification

This section details the classification methodology applied to the three airborne particle imagers introduced in section 2.1. Ice crystal habit classification is essentially a pattern recognition problem. One wants to assign a unique label to each particle recorded based on its image, the content of which is usually summarized in a finite number of numerical variables (descriptors). Because crystal habit is usually easily identifiable by visual inspection, a supervised classification approach was adopted according to the following guideline. First, the number of classes included in the model was determined. In the second step, a large number of particle images were selected and manually labeled in order to have a reference data set that can be used for training the classifier. This data set was then divided into two subsets: a pure training set used to build the classifier and a validation set intended for evaluation purposes. This step also required the selection of a classification algorithm, MLR in this work. The whole methodology was applied to the three OAP data sets individually. The following subsections explain each of these steps in more detail.

3.1. Classes Definition

A common aspect to all supervised classification problems is that they require a fixed number of classes to be defined beforehand. In some applications, this step comes conveniently as there is a clear and distinct number of classes (handwriting recognition, for instance). In cloud particle classification, however, this step is not straightforward due to the high diversity of crystal sizes and shapes observed in nature. As a result, many classification schemes including various number of distinct habits have been introduced (e.g., Korolev & Sussman, 2000; Lindqvist et al., 2012; Um & McFarquhar, 2009). In this contribution, a nested classification model is proposed in an attempt to be as exhaustive and flexible as possible. Six classes are used in the upper layer of the model: columnar crystal (CC), planar crystal (PC), bullet rosette (BR), aggregate (AG), compact particle (CP), and quasi-spheres (QS). For the 2D-S probe, an additional category named other (OT) was added in order to identify small out-of-focus images, which appeared to happen frequently. Samples collected with the HVPS, 2D-S, and CPI for each of the defined classes are illustrated in Figure 3. For the sake of readability, images from 2D-S and HVPS were artificially magnified by a factor of 1.5 and 9, respectively. The 2-mm reference scale present on each subpanel relates the size of the particles displayed to their real physical size. Because of the scale difference in the crystals recorded by the three probes, the HVPS device is not able to identify small habits ($\lesssim$500 μm) but can capture much larger graupel-like particles and aggregates, which would appear truncated on the two other probes.

At the upper level, the classification scheme uses nearly identical categories to the ones initially introduced by Magono and Lee (1966). However, columns and needles are merged together in this first level of classification. Moreover, the categories combinations of planar crystals, combinations of columnar and planars crystals and germs of snow crystals are discarded because they were not observed in sufficient quantities to reliably train a classification algorithm. Nevertheless, it should be noted that these categories could easily be reintroduced in future studies if enough training samples are provided. Finally, visual inspection of numerous samples associated with the class compact particle suggests that this category is essentially composed of compact heavily rimed and graupel particles. However, the term of compact particle was adopted because it is objectively difficult to differentiate between a graupel and a potentially less rimed irregular compact crystal from these OAPs images. This is particularly relevant in CPI images where a large amount of irregular ice particles in the range of 50 – 150 μm are classified as CP due to their apparent compact shape. Riming is unlikely in this size range, as reported by Ono (1969) who observed no riming on columnar crystals with minor axis <50 μm and only very rare occurrence of riming on planar crystals with a diameter < 300 μm.

This general classification scheme can be divided into subcategories in a nested fashion. In section 4.3, two examples of subclassification are proposed: columnar crystal into column (CC-C) and needle (CC-N), and aggregate into aggregate of bullet rosettes (AG-R) and other aggregate (AG-O). The whole classification framework being flexible and easily adjustable, different nested schemes could be considered in the context of different objectives or observations. For instance, one could consider adding a second layer of classification within planar crystals to differentiate between hexagonal plates, sectored plates, and dendrites.

3.2. Training and Validation Set

A satisfactory classifier must be able to reproduce the same performance on an unknown data set as compared to the reference training set. For this reason, the selection of the training set is a crucial step in the development of a classification method. Each class must present a certain consistency in its statistical properties while ensuring that it is representative of the variability observed in real data. For instance, the category aggregate must contain samples of different types (aggregates of plates, columns, bullet rosettes, …) and configurations (sizes, aspect ratios, orientations, …) in order to be as generic as possible.

To achieve this goal, particle images were selected and labeled conjointly by two independent operators. The images were extracted from diverse time steps and altitudes during different flights in order to cover a broad range of atmospheric conditions. The procedure followed was to first identify ~100 samples for each habit and then progressively increase the size of the training set until the classification performance reached a certain stability. This stability criterion was assessed on a validation data set, that is, a labeled data set that was not used to train the classifier, by means of a learning curve (see Figure 6). In total, $N_{\text{ref}}^{\text{HVPS}} = 1,410$, $N_{\text{ref}}^{\text{2D-S}} = 4,252$, and $N_{\text{ref}}^{\text{CPI}} = 2,964$ samples were retained to train and validate the HVPS, 2D-S, and CPI classifiers, respectively. It should be noted that because this reference data set was collected during the OLYMPEX field campaign, applying the classifier on other data sets could require some adjustments. Depending on the latitude and environmental conditions of the observations, it might be necessary to consider adding new training samples in the different classes and/or extending the classification scheme with new categories.

3.3. Classification Method

The fast development of data clustering and machine learning algorithms observed during the past decades led to a large choice of viable classification methods to consider. For the present classification problem, three classification methods were tested and compared: MLR, SVM, and ANNs. Without investing substantial effort into fine tuning the models, no classification algorithm was found to outperform the others on every aspect. SVM seemed to produce higher classification accuracy by ~1 – 2% but at the cost of a higher dependence on the validation data chosen and a reduced stability. At first glance, ANN model was never able to reproduce the accuracy of the two other methods. This aspect could be explained by the relatively small size of the reference data set for a high-dimensional (>10) problem. MLR yielded a satisfactory classification accuracy and a higher robustness to validation data sampling and was therefore selected for this study. A brief derivation of MLR is provided here, with more details given by Bishop (2006) and others.

Logistic regression is a statistical probabilistic model and can be seen as an alternative to the linear regression where the output is bounded and can be interpreted as a probability of belonging to the different classes introduced in the problem. In linear regression, the dependent variable y_n is estimated based on the value of D independent variables, also called descriptors or features, assembled as a vector $\boldsymbol{x} \in \mathbb{R}^D$ with

$$y_n = \beta_0 + \beta_1 x_{n1} + \ldots + \beta_D x_{nD} = \tilde{\mathbf{x}}_n^T \boldsymbol{\beta}, \tag{5}$$

where $\tilde{\mathbf{x}}_n = (1, \mathbf{x}_n^T)^T$ is the augmented vector of independent variables and $\boldsymbol{\beta}$ the vector of regression weights. In binary logistic regression, a logistic transformation $\mathcal{G}(x)$ is applied to the linear model and transforms the output y_n into a probabilistic variable given by

$$p\left(y_n = 1 \mid \tilde{\mathbf{x}}_n, \boldsymbol{\beta}\right) = \mathcal{G}(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}) = \frac{\exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}\right)}{1 + \exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}\right)}, \tag{6}$$

$$p\left(y_n = 2 \mid \tilde{\mathbf{x}}_n, \boldsymbol{\beta}\right) = 1 - \mathcal{G}(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}) = \frac{1}{1 + \exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}\right)}, \tag{7}$$

where $\{1, 2\}$ are two classes. As an extension to the logistic regression for multiclass classification problems, MLR performs categorical classification of the dependent variable $y \in \{1, \ldots, K\}$. The logistic function is turned into a softmax function and the probabilities $p(y_n = k)$ of belonging to the different classes can be rewritten as

$$p\left(y_n = k \mid \tilde{\mathbf{x}}_n, \mathbf{B}\right) = \frac{\exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_k\right)}{\sum_{j=1}^{K} \exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_j\right)}. \tag{8}$$

The noticeable difference with the binary logistic regression is that K vectors of regression weights $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ have been introduced and regrouped in a matrix $\mathbf{B}$ for the sake of brevity. The whole concept of training the algorithm consists of optimizing the values contained in $\mathbf{B}$ based on a set of N_{train} pairs $\{\mathbf{x}_n, y_n\}$ of labeled data. In other words, we want to maximize the likelihood of observing $\mathbf{y} = \{y_1, \ldots, y_{\text{Ntrain}}\}$ given $\tilde{\mathbf{X}}$ and $\mathbf{B}$. In the Bayesian framework and assuming the independence of the y_n, equation (8) can be rewritten as an objective function for $\mathbf{B}$:

$$\mathbf{B}_{\text{lik}} = \text{argmax}_{\mathbf{B}} p(\mathbf{y} \mid \tilde{\mathbf{X}}, \mathbf{B}) = \text{argmax}_{\mathbf{B}} \prod_{n=1}^{N_{\text{train}}} p(y_n \mid \tilde{\mathbf{x}}_n, \mathbf{B}). \tag{9}$$

The latter equation is called a likelihood and can easily be converted to a standard cost function by applying a negative logarithm transform as follows:

$$\mathcal{C}\left(\mathbf{B}\right) = -\sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{y}_{nk} \tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_k + \sum_{n=1}^{N} \log \sum_{j=1}^{K} \exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_j\right). \tag{10}$$

Even though usable as is, this cost function is modified here to avoid two common issues that can occur in classification problems: overfitting the training data set and neglecting underrepresented categories. In order to reduce the sensibility of the model to the training data set and penalize arbitrarily large values in the vectors $\boldsymbol{\beta}_k$, a regularization parameter λ is added to the cost function following Bishop (2006; procedure also known as Ridge regression or L2 regularization). In order to deal with unbalanced training data and similarly to Praz et al. (2017), a simple factor in the cost function is applied to weight each training data pair $\{\mathbf{x}_n, y_n\}$ by a number inversely proportional to the occurrence frequency of the y_n category in the training set. This strategy is commonly used in classification problems dealing with imbalanced data (López et al., 2013). If f_n denotes the proportion of data belonging to the same class as y_n (in the reference data set) and K the total number of classes, this factor can be written as $\omega_n = 1/\left(K f_n\right)$. The inclusion of ω_n proved to be notably effective in the 2D-S and CPI classifiers where the class Planar crystal is significantly underrepresented (8% and 13% of the training set, respectively). Mathematically, these modifications translate into the following changes in the standard MLR cost function, which is now given by

$$\mathcal{C}_f\left(\mathbf{B}\right) = -\sum_{n=1}^{N} \omega_n \sum_{k=1}^{K} \tilde{y}_{nk} \tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_k + \sum_{n=1}^{N} \omega_n \log \sum_{j=1}^{K} \exp\left(\tilde{\mathbf{x}}_n^T \boldsymbol{\beta}_j\right) + \lambda \sum_{j=1}^{K} \boldsymbol{\beta}_j^T \boldsymbol{\beta}_j. \tag{11}$$

The last term in equation (11) corresponds to the L2 regularization, and its effect can be adjusted by tuning the hyperparameter λ (larger λ yields a higher degree of regularization). During the training phase, one typically adjusts λ in order to have similar classification accuracy within the training and the validation set. Note that this regularization term is not altered by the ω_n weighting. In this application, the inclusion of ω_n is improving the classification accuracies presented in section 4 by ∼1% while diminishing the variability around the obtained values by ∼50%.

The final cost function shown in equation (11) can be straightforwardly adapted to binary classification problems ($K = 2$) and is applied in section 4.3 to the CPI data to illustrate the concept of nested subclassification with two examples: the subclassification of columnar crystals into columns and needles and the subclassification of aggregates into aggregates of bullet rosettes and other aggregates. Note that learning techniques based on neural networks are ideally suited for nested classification problems (e.g., Guo & Gelfand, 1992) but require a much larger training set, implying a very time-consuming manual labeling phase.

3.4. Classification Performance Assessment

The performance of the classifiers was assessed using conventional tools such as the confusion matrix and different score indices (e.g., Witten et al., 2016). The confusion matrix is a table giving an overview of the classification strengths and weaknesses of a supervised classification model. Each row represents samples predicted to belong to a certain class, whereas each column represents the ground truth, according to the manual labeling performed beforehand. Examples of confusion matrices for the HVPS, 2D-S, and CPI classification schemes are provided in Figures 5, 7, and 8, respectively. If **M** is the confusion matrix, then M_{ij} contains the number of samples belonging to class j and classified in class i. The diagonal components M_{ii} contains the samples correctly predicted. Based on this confusion matrix, one can define three performance indices (among others): the OA, the Heidke skill score (HSS, also known as the Cohen's kappa), and the error rate (ER), defined as follows:

$$\mathrm{OA} = \frac{\sum_{i=1}^{K} M_{ii}}{N} \times 100, \tag{12}$$

$$\mathrm{HSS} = \frac{\mathrm{OA} - E}{1 - E} \times 100, \tag{13}$$

$$\mathrm{ER}_j = M_{i\neq j,j}/M_{*,j} \times 100, \tag{14}$$

where K is the number of classes, N the number of samples considered, and $M_{*,j}$ the number of samples in the jth column. The HSS can be seen as an alternative to the OA, which takes into account the number of correct predictions that could occur by chance. In equation (13), the coefficient E evaluates this fraction of correct predictions and can be written as

$$E = \frac{1}{N^2} \sum_{i=1}^{K} M_{i,*} M_{*,i}. \tag{15}$$

Note that the HSS is conventionally defined $\in [0, 1]$ but was converted to percent here for sake of consistency with the other indices. Finally, in order to have a unique error index for each classification model, the balanced error rate (BER) is defined from the ER as

$$\mathrm{BER} = \frac{1}{K} \sum_{j=1}^{K} \mathrm{ER}_j. \tag{16}$$

In order to further analyze the predictive capabilities of a classification model, it is good practice to compare the performance of the proposed model to a baseline model, called BAS hereafter. A baseline model consists of a simple yet relevant method to classify new images based on the reference training set. In this study, a BAS model based on feature centroid was used. For a classifier relying on D descriptors, the BAS simply assigns to a new sample i the class k whose centroid is the closest in the feature space. Mathematically, this translates to

$$\mathcal{K}_i = \mathrm{argmin}_k \sum_{j=1}^{D} |x_{ij} - \bar{x}_{kj}|, \tag{17}$$

where $\mathcal{K}_i$ is the class assigned to the new sample i, x_{ij} the value of the jth descriptor for sample i, and $\bar{x}_{kj}$ the average value of descriptor j within the class k (i.e., the centroid), calculated from the training set. The introduction of a baseline model is very valuable to distinguish the part of the classification performance

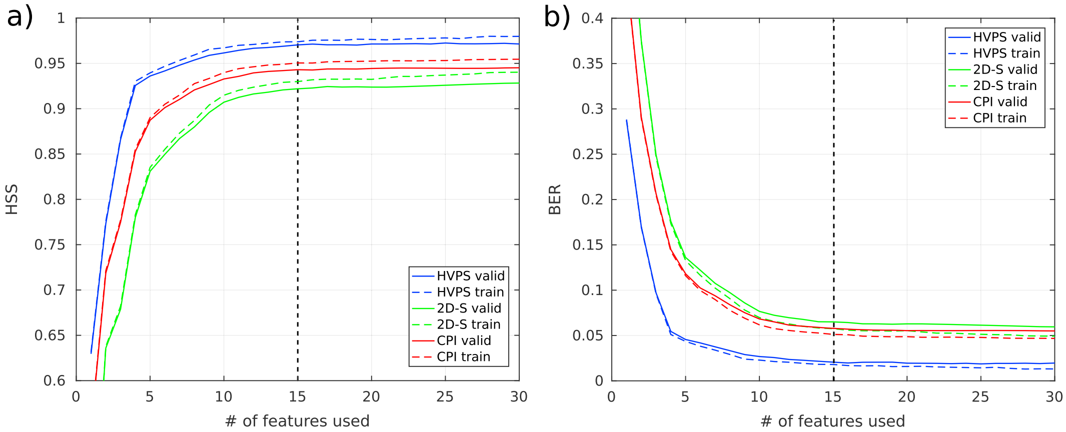

Figure 4. HSS (a) and BER (b) scores obtained when applying the forward features selection algorithm. On both panels, the training and validation scores are displayed as dashed and solid lines, respectively. Color codes are indicated in the legend. Features added after the vertical dashed black line were not kept for classification. BER = balanced error rate; HSS = Heidke skill score; HVPS = high volume precipitation spectrometer; 2D-S = two-dimensional stereo; CPI = cloud particle imager.

explained by the relevance of the selected features from the added value brought by the model itself (MLR here). Typically, a high gap in classification accuracy between the chosen model and the baseline might be an indicator of overfitting.

Finally, cross validation is used as a tool to validate the HVPS, 2D-S, and CPI models and assess how these classifiers would likely generalize to new independent and unlabeled data sets (e.g., Witten et al., 2016). The principle behind one round of k-fold cross validation is the following: the reference (labeled) data set is first partitioned into *k* subsets of equal size. Consecutively, each single subset is kept aside and used for testing the performance of the model on an unbiased sample (validation set), whereas the $k-1$ other subsets are merged and used to fit the model (training set). This procedure is usually conducted several times on different partitions to reduce variability and assess the stability of the classifier. For this study, fourfold cross validation is used as it proved to be a good trade-off between the size of the training set and stability of the test error. Different fold values between 3 and 10 were also investigated and showed very similar results in terms of classification performance.

4. Results

This section is divided into four subsections. In the first section, the 111 descriptors introduced in section 2.2 are investigated in order to select a subset containing only the most relevant ones to be employed for the classification task. In the second step, the three classification models built for HVPS, 2D-S, and CPI probes are assessed in term of accuracy and robustness as well as generalization capabilities following the methodology introduced in section 3.4. The third subsection is dedicated to the possibility to define subclassification schemes and provides examples. Finally, classification outputs from the three probes are presented and compared in terms of habit proportion for two different flights conducted during OLYMPEX.

4.1. Feature Selection

As explained in section 2.2, an ensemble of 111 geometric and textural particle descriptors has been introduced for the classification task. The selection of a relevant yet nonredundant set of features for the classifier to rely upon is a crucial step. It directly impacts the performance and numerical stability of the classification algorithm. Correlated descriptors have a negative influence on the convergence properties and stability of the solution (i.e., the matrix of regression weights $\mathbf{B}$) as they tend to make the input descriptors matrix $\tilde{\mathbf{X}}$ ill-conditioned and almost singular. As a result, the cost function minimization step would be prone to large numerical errors leading to a higher risk of overfitting training data. The goal here is therefore to come up with a restrained and uncorrelated set of descriptors, which minimizes the intraclass variability while maximizing the interclass variability, based on the reference data set. As a typical dimensionality reduction problem, this task can be achieved using conventional feature extraction techniques like principal component analysis (successfully applied in Lindqvist et al., 2012) or more advanced tools such as autoencoders, recently used

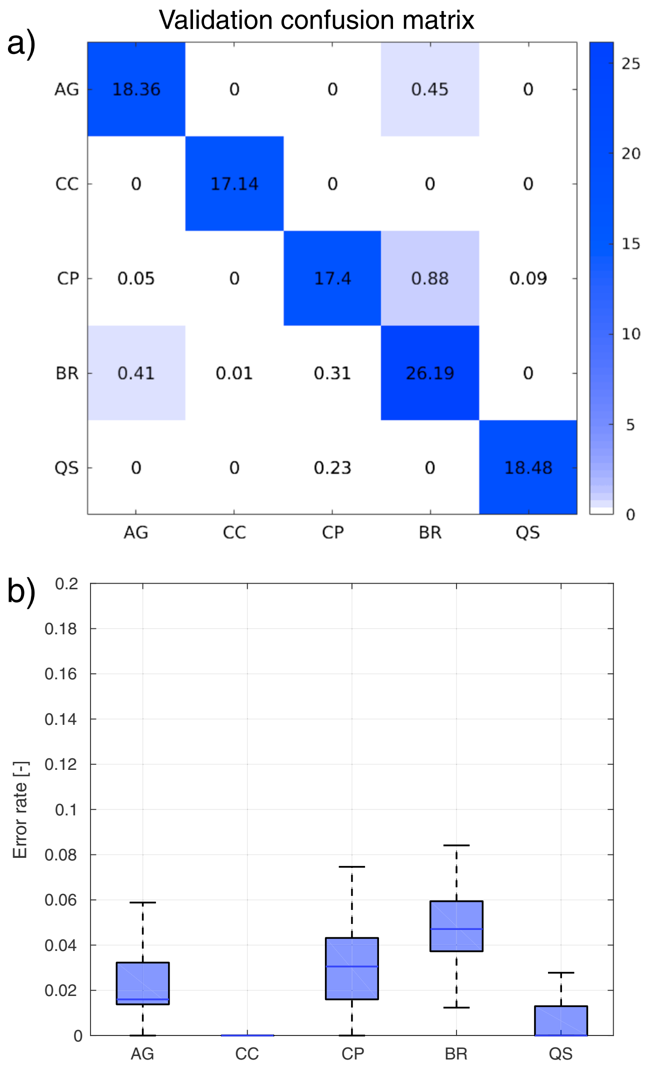

Figure 5. Overview of the classification accuracy of the multinomial logistic regression model applied to high volume precipitation spectrometer data. (a) Confusion matrix obtained on validation data averaged over 10 instances of fourfold cross validation. The *x* and *y* axes represent the true labels and predictions obtained by the classifier, respectively. Entries on the diagonal correspond to samples correctly classified. The matrix entries are normalized so that they sum up to 100%. Note the nonlinear color bar used to emphasize misclassification. (b) Box plots of the error rate by habit corresponding to the confusion matrix above. Each box plot is therefore calculated over 40 values. Solid lines represent the median values, the boxes show the 25–75% quantiles, and the whiskers extend to 1.5 times the interquartile range (or the minimum, respectively, maximum sample value if this one is included within this interval). AG = aggregate; CC = columnar crystal; CP = compact particle; BR = bullet rosette; QS = quasi-sphere.

in remote sensing imagery (Zhang et al., 2015). However, feature extraction techniques are not ideal for the present work as they usually remap the input descriptors into a new space, which provides no room for feature interpretation. In our case, it is very insightful and thus desired to be able to link relevant features to physical properties of the particle/habit as these features could in turn be used to simulate realistic ice crystals (e.g., for scattering simulation purposes) or to accurately parametrize habit in numerical weather prediction models.

Therefore, a forward feature selection technique that keeps the initial input descriptors unchanged was implemented. Based on the highest HSS value (see equation (13)), the process starts by implementing a single dimension MLR model (i.e., based on one descriptor only) . Features are then iteratively added to the model based on their discriminative power. At each step, the method tries to add every remaining descriptor to the current list and select the most relevant one. This selection is performed according to the highest averaged HSS evaluated on the validation set on the basis of 10 iterations of fourfold cross validation. More detail on the algorithm and its implementation can be found in Tang et al. (2014).

Results obtained by applying the proposed forward feature selection method are shown in Figure 4. In terms of HSS (Figure 4a), the three validation curves (blue, red, and green solid lines), corresponding to the three classifiers (HVPS, 2D-S, and CPI, respectively), exhibit a similar behavior with a strong increase of classification accuracy brought by the five first features added, a moderate improvement associated with the ~5 next features, and eventually a plateau reached by the validation curves around feature 15. After that point, the training HSSs (dashed line) seem to continue to grow at a very slow rate. Similar conclusions can be drawn from Figure 4b displaying the BER, an independent metric that is not involved in the feature selection process. For the sake of illustration, the evolution of the curves after 30 features is not shown but displays a slight decrease in the validation HSS while the training HSS is constant, indicating potential overfitting. Based on these observations, only the 15 first selected features were kept for the classification task. Table 2 gives an overview of the selected descriptors for the three probes.

4.2. Validation Using Reference Data Set

4.2.1. HVPS

The classification performance for the HVPS probe was assessed on the basis of a reference labeled data set $N_{\text{ref}}^{\text{HVPS}}$ composed of 1,410 samples. Compared to the 2-DS and CPI classifiers, this is a relatively low number of samples but somewhat compensated by the reduced number of classes included in the model ($K = 5$, PC ignored because only a few dozen occurrences were found in the OLYMPEX data set investigated). The performance and accuracy of the classifier were evaluated based on a confusion matrix and the different tools introduced in 3.4. Figure 5a shows the confusion matrix obtained by averaging 10 instances of fourfold cross validation. Each instance consists of a different random sampling of N_{ref} into four equal subsets required for cross validation. Note that the matrix has been normalized to sum up to 100% so that each entry represents a fraction of the reference data set. In this way, one can assess the proportion of the reference data set belonging to each class by calculating the sum over the rows. The nonlinear color bar is adjusted to highlight not only the diagonal but also the entries where most misclassifications occur.

As a complement to the confusion matrix, Figure 5b shows the classification ER within each class. The median, box extents, and whiskers are derived from the same 10 random instances of fourfold cross validation. Median ER values are (in ascending order) as follows: 0% for columnar crystal and quasi-sphere, 1.6% for aggregate,

Table 3
Classification Performance Overview

OAP (N_{ref})	Method	OA	HSS	BER
HVPS (1,410)	MLR	97.6 ± 0.7%	97.0 ± 0.9%	2.1 ± 0.7%
	BAS	88.6 ± 1.3%	85.7 ± 1.6%	9.8 ± 1.1%
2D-S (4,217)	MLR	93.4 ± 0.8%	92.1 ± 0.9%	6.5 ± 0.7%
	BAS	82.2 ± 1.3%	78.9 ± 1.6%	16.4 ± 1.3%
CPI (2,964)	MLR	95.3 ± 0.6%	94.2 ± 0.7%	5.9 ± 0.6%
	BAS	8506 ± 1.0%	82.4 ± 1.2%	15.8 ± 1.0%

Note. OAP = optical array probe; OA = overall accuracy; HSS = Heidke skill score; BER = balanced error rate; HVPS = high volume precipitation spectrometer; 2D-S = two-dimensional stereo; CPI = cloud particle imager; MLR = multinomial logistic regression; BAS = baseline model.

3.1% for compact particle, and 4.7% for bullet rosette. Even though satisfactory for all habits, the ER is bigger for the BR class, mostly due to the missclassification of BRs as CPs or AGs as indicated by the confusion matrix.

Classification scores indicating the OA, the HSS, and the BER (as defined in section 3.4) are reported in Table 3. They appear to be very satisfactory with a HSS value of 97%, very similar to the OA. The reference baseline model (denoted as BAS in the table) also performs well with a HSS and BER of 85% and 10%, respectively, thus indicating that the set of features utilized is appropriate and relevant to the target concept. The difference in scores between the two models (ΔHSS=∼11%, for instance) quantifies the added value brought by the MLR method itself. It could also be used to diagnose overfitting in case the BAS model cannot reproduce the classification performance of the MLR at all.

To further assess the performance, stability and convergence properties of the classifier, one can compute the so-called learning curves. They are calculated based on the following procedure: first, a subset of 25% of the reference data set is kept aside for validation; then, the size of the training set is progressively extended from 2% to 100% of the rest of the reference data in constant steps of 2%. For each step, the OA, HSS, and BER scores are calculated on the training and validation set. In this way, the evolution of the classification performance is monitored as the number of training samples is increased. HSS learning curves have been calculated for the HVPS classifier and are displayed in Figure 6a. To assess the sensitivity of the curves to random sampling effects, the procedure was repeated 10 times with different splits between training and validation subsets. The validation curve exhibits a sharp increase at the beginning (i.e., for a low amount of training data) and flattens thereafter until reaching a plateau when more than 75% of the training samples are used. The training curve displays opposite behavior, starting at nearly 100% accuracy and decreasing slightly to finally reach a plateau. This plateau is characterized by a HSS value very close to that of the validation curve, which is a desired property and indicates that the model is not overfitting the training data set. Moreover, the asymptotic nature of both curves around 100% of training samples is an indicator that enough training data are used to have a

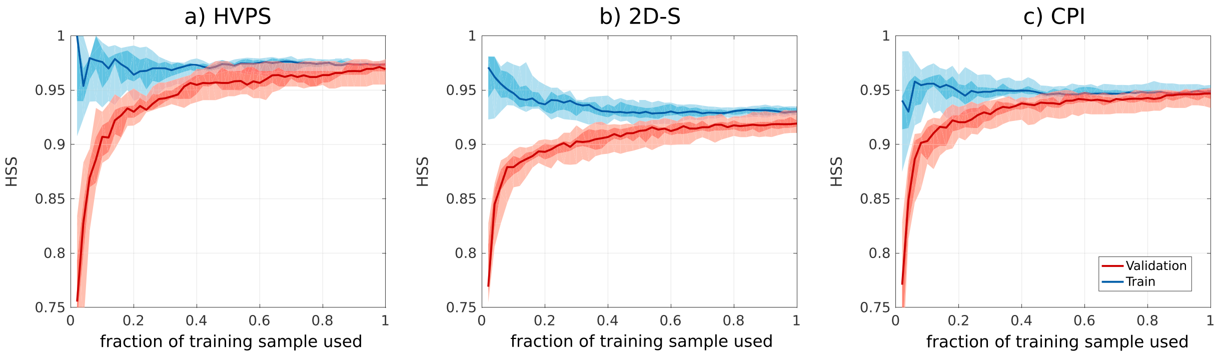

Figure 6. Learning curves for the three classification models built for (a) the HVPS, (b) the 2-DS, and (c) the CPI probes. On each panel, the evolution of the training HSS (in blue) and the validation HSS (in red) are plotted as a function of the size of the training set. The solid lines represent median values; dark shaded areas define the 25–75% interquartile range and light shaded extent to the 10–90% quantiles, on the basis of 10 iterations of train and validation subset random splitting. HSS = Heidke skill score; HVPS = high volume precipitation spectrometer; 2D-S = two-dimensional stereo; CPI = cloud particle imager.

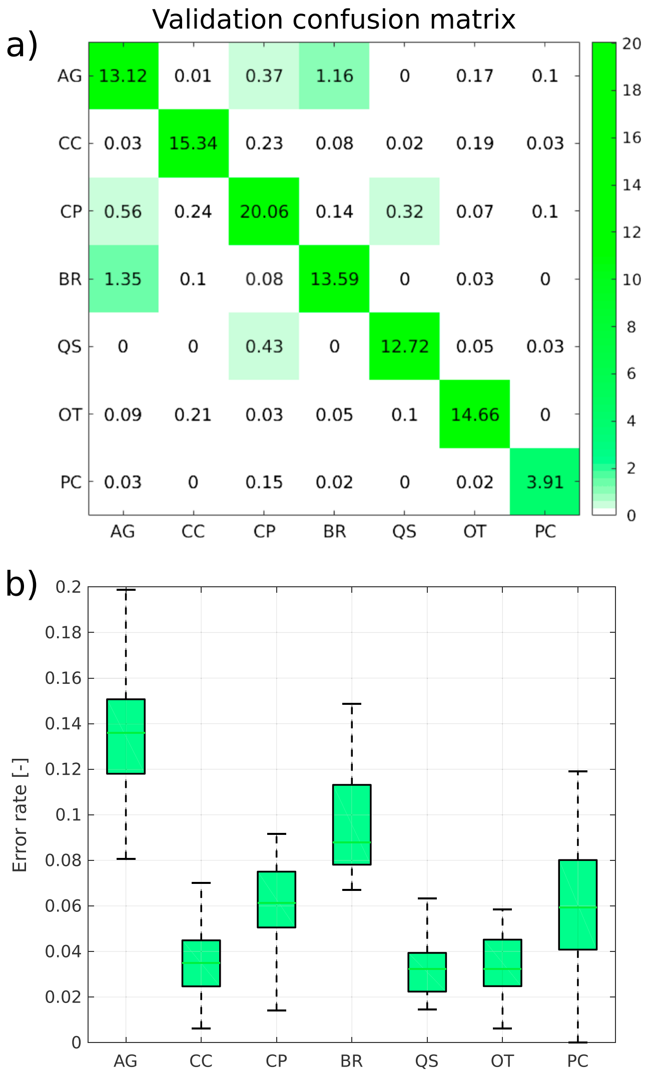

Figure 7. Same as in Figure 5 but for the two-dimensional stereo classification model.

reliable representation of the validation data set. The latter point suggests that the model should generalize well to real data. It should be noted that although the training and validation curves do not completely merge for the 2D-S probe (as they do for HVPS and CPI), the remaining gap is negligible and the parallelism of the curves indicates that the HSS scores will not vary even if the training set is increased. Finally, the restricted variation around the solid lines indicates that the model is rather insensitive to random training/validation data sampling.

4.2.2. 2D-S

Similar analyses have been conducted to evaluate the performance of the 2D-S classifier. Results obtained for the 2D-S probe based on $N_{\mathrm{ref}}^{\mathrm{2D\text{-}S}} = 4,217$ are shown in Figure 7 as well as in Table 3. In this case, the classification is based on seven distinct categories, with the addition of planar crystal and other (out-of-focus) categories.

As for the HVPS, 10 instances of fourfold cross validation were conducted and resulted in an accuracy characterized by an OA of 93.4%, a HSS of 92.1%, and a BER of 6.5%. Compared to the scores obtained for the HVPS probe, these are slightly below but still satisfactory. This might be a consequence of the more complex classification scheme (seven categories) utilized for 2D-S classification. Moreover, the analysis of the ERs indicates that most misclassifications occur in aggregate, bullet rosette, and planar crystal categories, in descending order. In particular, the highest source of error raises from the confusion between AG and BR (see Figure 7a) as well as between AG and CP to a lesser extent. Recalling that the discrimination is achieved based on a binary silhouette of the particle only, it is not surprising that similar habits confuse the classifier. This ambiguity between AG, BR, and CP was also noticed by the operators while manually labeling the data. Nevertheless, MLR provides an interesting tool to assess the confidence in the predicted habit in the form of the probabilities of belonging to each class introduced in the model. Indeed, once a MLR model is trained, it is straightforward to calculate these probabilities using equation (8). As an example, the average probability of belonging to the class AG has been calculated for all particles predicted as AG by the model, based on one iteration of cross validation. Results obtained show a clear distinction between correct and erroneous predictions, with average probabilities $\bar{p}(y_n = \mathrm{AG},\ \mathrm{true}) > 99\%$ and $\bar{p}(y_n = \mathrm{AG},\ \mathrm{false}) \simeq 75\%$. For applications requiring higher confidence in the predictive output, one can therefore impose a certain threshold to the probability in order to select only the most reliable predictions. If the 2D-S probe could be reconfigured to match images of the same particle collected by both orthogonal photodiode arrays, one could potentially improve the reliability of the method by performing the classification on the two projected images independently and merging the probabilities a posteriori in order to assign one unique habit per particle, according to the methodology applied in Kennedy et al. (2018) for MASC data.

Learning curves have also been processed for the 2D-S classification model and are displayed in Figure 6b. They exhibit a very similar behavior with respect to the HVPS model, yielding the same conclusions about the stability of the prediction output as well as the completeness of the training set.

4.2.3. CPI

The CPI classifier was tested and validated following the same methodology as for the HVPS and 2D-S probes, based on $N_{\mathrm{ref}}^{\mathrm{CPI}} = 2,964$ labeled samples. In contrast to the previous probes, the CPI provides 256 levels grayscale images allowing textural descriptors to be calculated. Interestingly, the feature selection process led to the inclusion of only one texture-based descriptor in the 15 kept for the classification task: the Haralick homogeneity (Haralick et al., 1973) Moreover, the evaluation of classification performance with and without this feature showed that the latter was not crucial as it improved the HSS by less than 3%. One can therefore conclude that the added value brought by the textural descriptors introduced in this study has a very limited impact for habit classification. This is in line with the findings from MASC hydrometeor classification

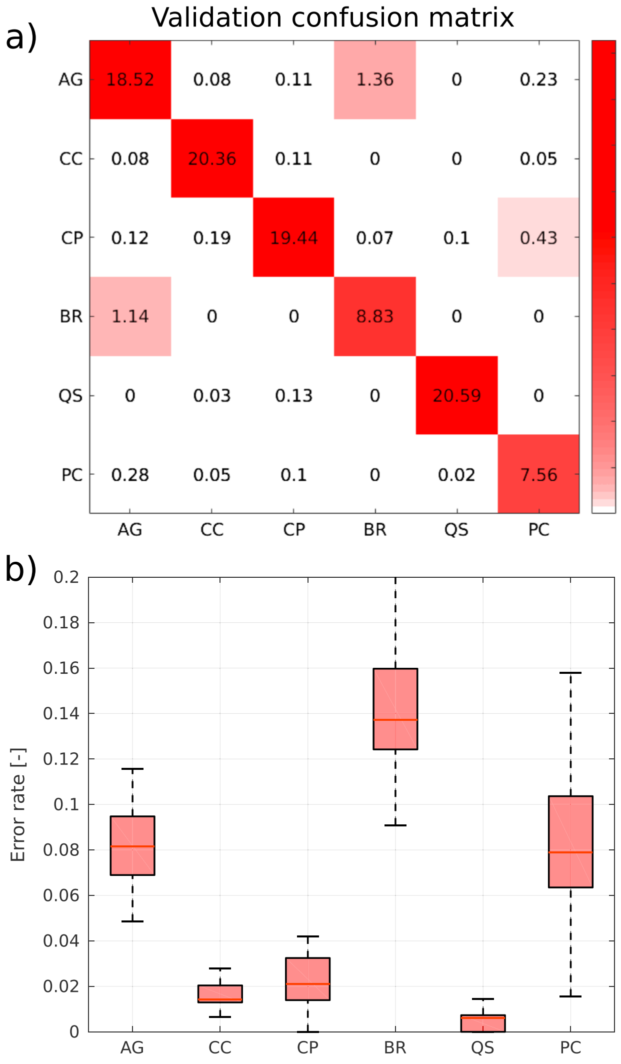

Figure 8. Same as in Figure 5 but for the cloud particle imager classification model.

(Praz et al., 2017) where the textural information appeared to be relevant only for estimating the degree of riming and detecting melting snow. Riming degree identification was not achievable and would require dedicated effort beyond the scope of this study due to some limitations intrinsic to the CPI device: noise in the photodiodes, lack of constant background intensity threshold and highly variable contrast, average brightness, and focus from image to image. Nonetheless, the presence of transparent areas within the ice particle boundaries (resulting in a low value of Haralick homogeneity) proved to be a relevant criterion to distinguish between some planar crystals and compact particles.

Figures 8 and 6c present the CPI classification results in the same manner as outlined for the HVPS and 2D-S imaging devices. Similar conclusions to the 2D-S classification model can be drawn, namely, a slightly reduced classification accuracy (OA = 95.3%, HSS = 94.2%, and BER = 5.9%) compared to HVPS and a higher misclassification rate between AG and BR. In terms of learning curves, the CPI curves are comparable to the previous ones.

Finally, it is worth highlighting the importance and relevance of two descriptors common to the three classifiers, namely, the particle area to convex hull area ratio and the sixth component of the standardized distance to centroid Fourier power spectrum. Even though several past studies have used area ratio to describe the microphysical properties of ice particles, those works have generally introduced one single area ratio, namely, the ratio of particle area to a reference circle (e.g., Grazioli et al., 2014; Heymsfield et al., 2002; Korolev & Isaac, 2003; McFarquhar & Heymsfield, 1996). With eight area ratios retained for the classification task (see Table 2), the present study shows that combining different area ratios allows for better identification of the particle habit (compact vs. complex shape; rectangular, polygonal, or spherical outline, etc.).

4.3. Potential for Subclassification

As mentioned in section 3.1, the logistic regression framework offers the opportunity to refine the habit detection by introducing subclassification(s) in a nested fashion. In other words, one can train a binary/MLR model on the output of the initial classifier by selecting all particles belonging to a certain category and refining the classification scheme into two or more subhabits. Coupled with information collected by other in situ sensors, this gives the possibility to get more insight into the environmental conditions under which specific habits grow and in turn improve the calculation of cloud radiative effects (Bailey & Hallett, 2009).

As a proof of concept, two binary subclassification schemes are introduced in the present study. The first aims to discriminate aggregates of bullet rosettes from other aggregates. Thus, 88 aggregates of bullet rosettes and 123 other aggregates (mostly aggregates of plates) were manually identified and labeled from images initially recognized as aggregates. In a second step, a forward feature selection was conducted in order to identify the most relevant features for the subclassification. It appeared that a model based on only two descriptors achieves high classification accuracy characterized by a HSS of 95.3% evaluated on validation data. This score was not significantly improved by introducing additional descriptors. Figure 9 illustrates the classification results by displaying a scatterplot of the two first descriptors picked by the forward feature selection algorithm, namely the particle fractal index defined as $F = 2\ln\left(\frac{P}{4}\right)/\ln A$ (Grazioli et al., 2014) and the normalized perimeter autocovariance $\mathrm{acov}\left(s(\frac{1}{12}), 0\right)$.

Note that the fractal index was not utilized in the initial classification scheme and is therefore not present in Table 2. As the descriptors were normalized for feature selection, the axis units are arbitrary and uninformative. The use of a single descriptor (i.e., the fractal index) discriminates very well between the two habits. The second feature is not very informative by itself, as illustrated by the left-side histogram, but advantageously bends the decision boundary in the two-dimensional space created.

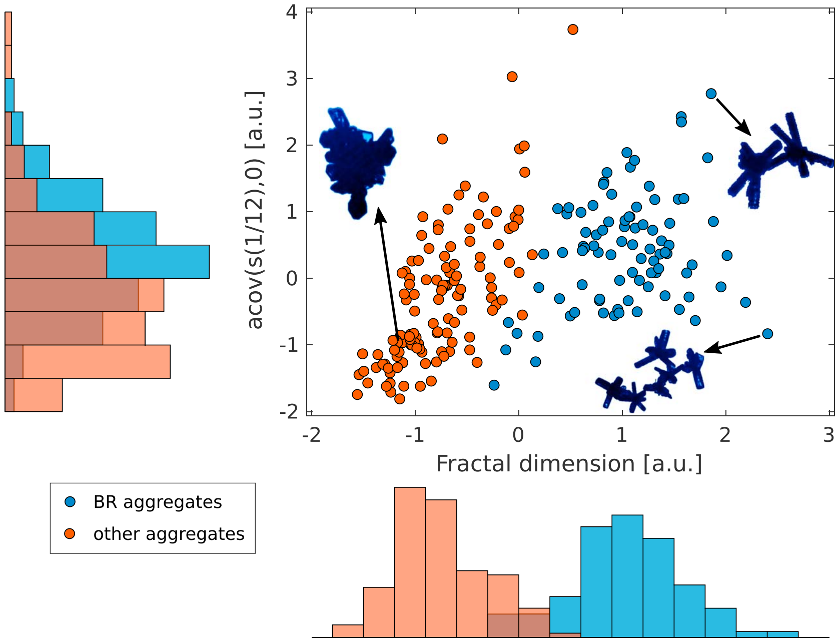

Figure 9. Scatterplot of the two descriptors selected by the feature selection algorithm for the subclassification of aggregates into aggregates of bullet rosettes and other aggregates, on the basis of 211 reference samples. The two descriptors have been normalized to have 0 mean and 1 standard deviation. The histograms represent the marginal distributions with respect to each descriptor independently. BR = bullet rosette.

The second example of subclassification separates columnar crystals into columns and needles, as initially introduced by Magono and Lee (1966). Needle-like crystals are characterized by a longer and more slender body compared to standard columns (Libbrecht, 2005). Results of the subclassification based on a reference data set composed of 90 columns and 81 needles are illustrated in Figure 10. In this case, the inclusion of a single descriptor, the aspect ratio (AR) derived from the smallest rectangle encompassing the particle (bounding box), yields a complete separation between the two habits. However, for the sake of consistency with Figure 9,

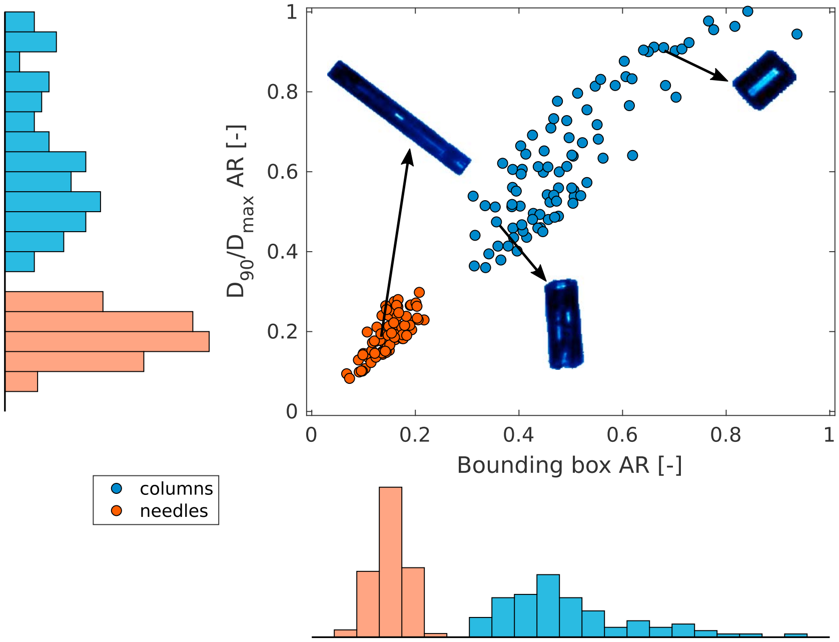

Figure 10. Same as in Figure 9 but for the subclassification of columnar crystals into columns and needles, as based on 171 reference samples.

Table 4
Overview of Flight Information and Environmental Data Related to the Two Flight Periods Selected for the Comparative Analysis Conducted in Section 4.4

Date	2D-S/HVPS	CPI	Elevation (m agl)	*T* (°C)	RH w.r.t. ice (%)	LWC (g/m^3)
12 November 2015 (F1)	19h20–19h21	19h20–19h25	1,100–1,600	−1–4	90–105	peak at 0.3
1 December 2015 (F2)	00h45–00h50	00h47–00h48	7,000	−28–30	79–85	< 0.01

Note. 2D-S = two-dimensional stereo; HVPS = high volume precipitation spectrometer; CPI = cloud particle imager; LWC = liquid water content; agl = above ground level; w.r.t = with respect to.

a scatterplot of the first two features selected is presented. The second feature, called D_{90}/D_{max} in this study, corresponds to the ratio between the largest dimension perpendicular to D_{max} and D_{max}.

Note that in this case the axes were not renormalized as AR values are naturally bounded and can be easily interpreted or compared to other studies. In the literature, different definitions of AR have been used (e.g., D_{90}/D_{max} in Korolev & Isaac, 2003; Hogan et al., 2012; ellipse fit based AR in Garrett et al., 2015; Jiang et al., 2017) and comparisons between those could be problematic if the obtained values are significantly different. In the present case, we clearly see that even on rather simple shapes like columnar crystals, the use of different AR definitions leads to different values. This observation appears to be more pronounced on more compact columns where the AR defined as D_{90}/D_{max} is systematically larger than the bounding box AR.

4.4. Application to Independent Flight Periods

In order to evaluate the potential of the classification method on independent data as well as to compare and validate the output obtained with the different probes, the classification was performed on two data sets collected during two flights carried out by the UND Citation aircraft during OLYMPEX. Due to the large amount of time required to extract, process, and classify the high number of images collected by the HVPS and 2D-S probes, the analysis was limited to two flight periods of 1 min each collected in contrasting environmental conditions. Results are presented in the form of proportions of different habits identified by the classification algorithm applied to each of the three probes. For a more detailed analysis, one would have to investigate both shape and size of the imaged particles using state-of-the-art processing techniques to calculate particle size distribution for each probe (e.g., Heymsfield et al., 2013; Jackson & McFarquhar, 2014; Protat et al., 2011) but this is beyond the scope of this contribution.

The first interval (F1) consists of 1 min extracted at 19h20 from a flight performed on 12 November 2015. At that time, the aircraft was climbing steadily from 1,100 to 1,600 m above ground. The sample is characterized by an average temperature of −2 ° C, an average relative humidity with respect to ice of 95% and a peak in liquid water content (LWC) reaching ~0.3 g/m^3 (compared to < 0.05 g/m^3 in the minutes before and after), as retrieved with a King probe. An overview of the selected time intervals is reported in Table 4. As the sampling volume of the CPI probe is much lower than that of the HVPS and 2D-S, the time interval was extended to 5 min

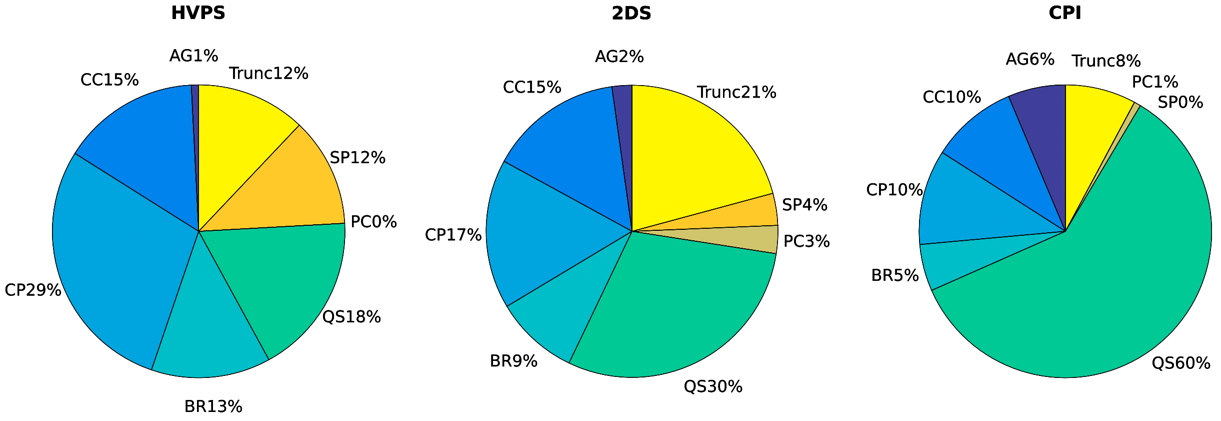

Figure 11. Classification output for the three probes installed on the Citation aircraft during OLYMPEX flight on 12 November 2015. The results displayed are for 1-min period between 19h20 and 19h21 UTC (extended to 19h20–25 for the cloud particle imager probe). The three panels show the proportions of each habit detected as pie charts. Trunc label (yellow) stands for particles that appeared truncated on the optical array probe image. HVPS = high volume precipitation spectrometer; 2D-S = two-dimensional stereo; CPI = cloud particle imager.

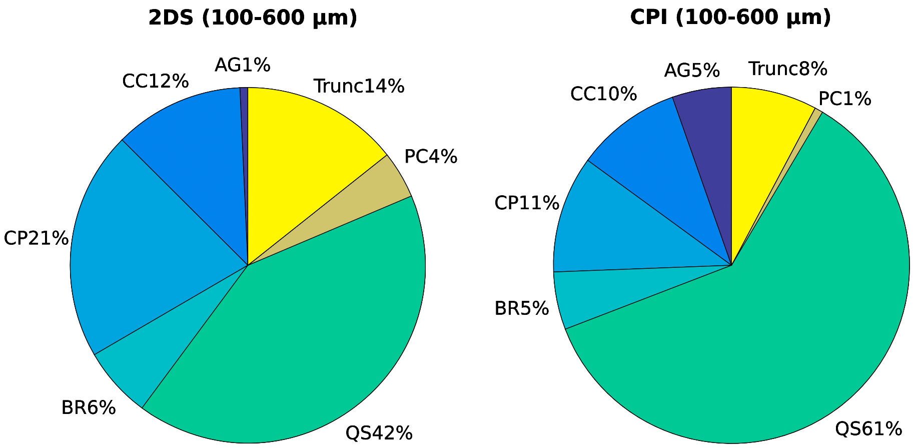

Figure 12. Classification output of the 2D-S and CPI probes for F1 restricted to a common size range between 100 and 600 µm. 2D-S = two-dimensional stereo; CPI = cloud particle imager.

for the CPI in order to have enough images to conduct a comparative study. Fluctuations in environmental conditions during these 5 min were limited and are reported in Table 4.

Results obtained for F1 in terms of proportions of particle habit identified by the MLR models are presented in Figure 11. As pointed out in several studies (e.g., Baumgardner et al., 2011; Lawson, 2011; McFarquhar et al., 2007), airborne probes suffer from many artifacts like truncated targets, out-of-focus images, or shattering of crystals into small fragments at the inlet of the detection system. In order to mitigate the impact of some of these artifacts on the proportions obtained, a few basic filters were applied beforehand. First, particles smaller than the first habit classified as *non-SP* were discarded. This filter is necessary to remove the extremely large amount of images consisting of less than five shaded pixels present in the HVPS and 2D-S data set (93% and 99%, respectively). Then, in order to identify truncated crystals, the ratio of the perimeter length touching the edge of the frame was calculated and particles characterized by a ratio > 0.3 were flagged as truncated. Finally, CPI images with high noise in the background or very low contrast were discarded as well. Even though the latter filter is not crucial here as the two time intervals were selected during periods where the CPI was providing high quality images, it becomes essential when performing the classification at the flight scale as bad quality CPI images are rather frequent (10% of the whole CPI data set for the flight on 12 November 2015).

The proportions of the different habits identified are displayed in Figure 11 in the form of pie charts. At first glance, the displayed distributions look quite different. Before proceeding further, it is important to note here that the three considered probes are characterized by significantly different specifications such as pixel resolution, observable size range, and sampling volume. For instance, more than 90% of the CPI and 2D-S particles collected and classified within F1 exhibit a $D_{\max} < 0.5$ mm. This corresponds to 1–3 of the HVPS photodiodes

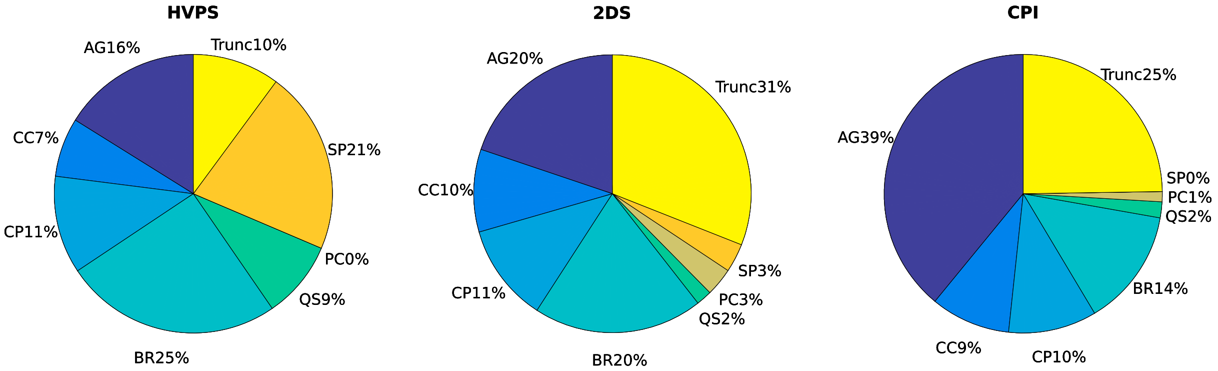

Figure 13. Same as in Figure 11 but for the OLYMPEX flight on 12 November 2015. Results shown correspond to a time period of 1 min between 00h47 and 00h48 UTC (extended to 00h45–50 for the CPI probe). HVPS = high volume precipitation spectrometer; 2D-S = two-dimensional stereo; CPI = cloud particle imager.

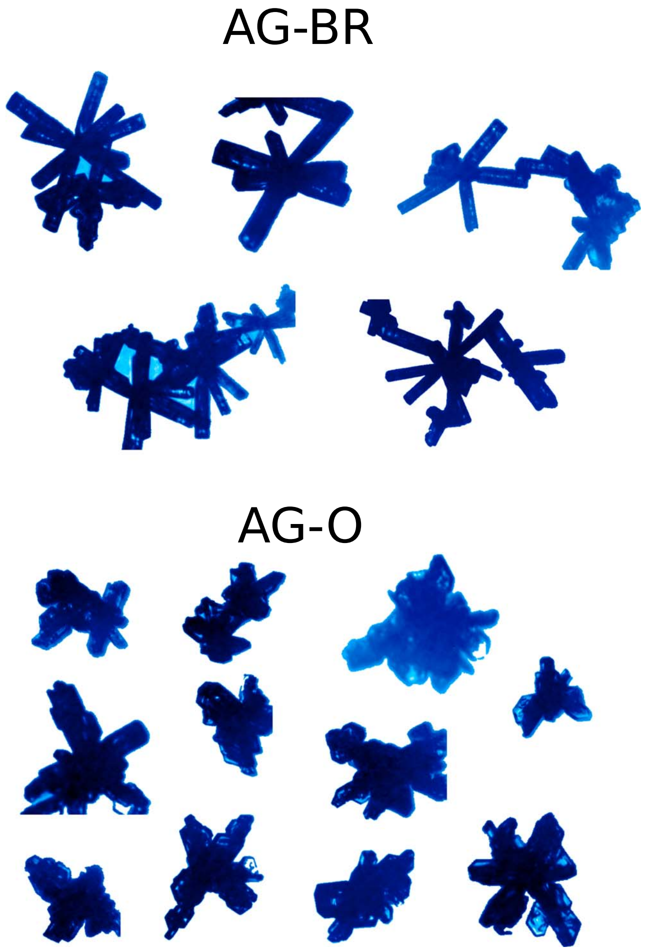

Figure 14. Illustration of ice particles classified as aggregates collected on 1 December 2015 between 00h45 and 00h50 UTC. Habit identified as aggregates of bullet rosettes (AG-BR) by the subclassification scheme are shown on the top. Other aggregates (AG-O) are on the bottom.

shadowed, a size range hence too small to identify habits with this device. As a result, consistency between the raw classification output from the three probes is not expected. Rather, the proportions should be interpreted in a complementary manner as each device can identify habits in a different size range. In order to somehow verify the consistency between the different classifications, the habit proportions obtained for the 2D-S and CPI were restricted to a common size range between 100 and 600 μm. In this interval, both devices are expected to measure a lot of particles entirely contained within the optical array (i.e., not truncated habits) and 100 μm is also sufficiently large to get enough 2D-S photodiodes shadowed for a reliable particle identification. Habit proportions given by the 2D-S and CPI in the restricted size range are illustrated in Figure 12 and are in much better agreement. A few discrepancies remain between the classes CP and QS. Since the QS habit is characterized by small diameters (typically <50 μm), the large difference in the probe resolution could explain these discrepancies.

The three classifiers diagnose a large number of small quasi-spheres, which correlates well with the peak observed in the LWC recorded by the King probe during that period. Moreover, the amount of QS detected by the CPI is noticeably higher compared to the other probes. Given the finer resolution of the CPI and the fact that >95 % of QS habit detected by the CPI were <100 μm, one could assume that this peak in LWC consists of droplets too small to be resolved by the 2D-S and HVPS probes. However, as highlighted in previous studies using the CPI (e.g., McFarquhar et al., 2013; Nousiainen & McFarquhar, 2004), the distinction between small quasi-spherical ice particles and liquid droplets is delicate as the resolution limit of the instrument is approached. A good indicator of the presence of liquid droplets is given by the particle area to circumscribed circle area ratio α. McFarquhar et al. (2013) studied fluctuations of α in small cloud particles as a function of the ratio between the LWC and the total water content, and concluded that habits characterized by $\alpha > 0.9$ are most likely liquid droplets. In the present case, 40% and 10% of the QS habit identified are characterized by α values >0.8 and 0.9, respectively. Moreover, all these images were collected between 19h20–21 when the King probe indicates a peak in the LWC. It is interesting to note that all QS >60 μm detected were characterized by $\alpha > 0.9$ and a bright spot at the center of the particle. The presence of a bright spot was automatically detected following the procedure detailed in Appendix A. It is a typical signature observed in liquid droplet images when the particles are illuminated from behind (Saylor et al., 2002) and is commonly referred to as specular reflection. Finally, the nonnegligible amount of compact particles identified with the three probes could potentially attest to the presence of particles rimed after collecting supercooled liquid water droplets.

Results obtained for the second time interval considered (F2) are presented in Figure 13. During that period extracted from a flight performed on 1 December 2015, the aircraft was flying at a constant altitude of 7,000 m. The local environmental conditions were characterized by an average temperature of −29 °C, a relative humidity of ~80% with respect to ice and a low LWC <0.01 g/m^3. Compared to the previous case, the particles observed are globally larger, most of them being characterized by $500 < D_{\max} < 1{,}000$ μm for each probe. Similarly to F1, the proportions of SP and truncated particles are highly variable and probe dependent. In terms of habit, the classifiers all identify a large number of AG and BR, in good agreement with the observations of Bailey and Hallett (2009) at similar temperature and humidity ranges. As expected from the low temperature and LWC, almost no QS habit are observed except by the HVPS (10%). These 10% may be composed of out-of-focus particles appearing rounded by blur or small particles with an insufficient number of photodiodes shadowed to resolved their shapes. The pie charts also indicate a larger amount of AG habit detected by the CPI. For large particles, the CPI device collected a lot of truncated images, which could potentially lead to some misclassifications of BR identified as AG. Recalling that for both 2D-S and CPI the highest misclassification rate was observed between AG and BR (see Figure 8) and that even manual identification was ambiguous sometimes, a perfect agreement is therefore not expected. Moreover, if we omit small particles and truncated images and merge AG and BR classes, one obtains very similar proportions for the three probes (HVPS = 60%, 2-DS = 60%, and CPI = 70%). In order to further investigate this issue as well as to test the applicability of the subclassification scheme proposed in section 4.3, the AG habit is subclassified into aggregates of BR and other aggregates. The results obtained indicate that among the 232 images classified as aggregate, 26% are identified as aggregates of bullet rosette. Furthermore, manual inspection of the other aggregates revealed that they are composed mainly of aggregates of plates, as illustrated in Figure 14. These findings are consistent with Bailey and Hallett (2009) who observed a transition in habit around −30 °C from BR to platelike crystals. Figure 14 also highlights the large variability of crystal shapes within the same class and therefore the difficulty to classify complex ice particles into a finite set of discrete habits. This motivated some cloud microphysical studies (e.g., Morrison & Milbrandt, 2015) to treat particle properties in a continuous manner rather than assuming a collection of discrete habits as it is done traditionally. In that effort, the classification approach proposed in the present study can be used to process large measurement data sets and precisely document the intraclass and interclass variability of the main particle properties.

In summary, the two flight periods analyzed here show very distinct microphysical properties of snow crystals in terms of habit. Moreover, the particle types revealed by the classifiers are in good agreement with the literature and what one could expect in such environmental conditions. Combined with state-of-the-art PSD retrieval techniques specifically designed for OAPs (McFarquhar et al., 2017), the proposed particle habit classification technique is suitable for a detailed shape and size analysis on a broad spectrum of particle dimension.

5. Conclusions

An automatic ice cloud habit classification model was developed and applied to three distinct OAPs commonly used on board research aircraft. The three probes, namely, HVPS, 2D-S, and CPI, cover a wide range of image resolutions, particle sizes, and sample volumes. Adapted from a technique initially developed for a ground-based snowflake imager (the MASC), the classification makes use of a customized MLR model in order to identify particle habit among six classes: columnar crystal, planar crystal, bullet rosette, aggregate, compact particle, and quasi-sphere. Two subclassification schemes were also proposed in order to distinguish columns from needles as well as to separate aggregates composed of bullet rosettes from those composed of other shapes, leading to eight classes in the end. Since the classification framework is highly versatile, new classes can be easily added or removed, but the procedure requires manual selection and labeling of a few hundred images for every new class included in the scheme. Some modifications and retraining may also be required for classifying images obtained in different environmental conditions. The classification is based on 15 features selected from 98 geometrical and 13 texture-based (for CPI only) descriptors introduced for this purpose. The principal conclusions drawn from this study are as follows:

- The three classifiers showed very satisfactory performance, characterized by overall accuracies and HSSs (calculated on validation data) of 97.6% and 97.0%, 93.4% and of 92.1%, and 95.3% and 94.2% for the HVPS, 2D-S, and CPI probes, respectively. This suggests that the MLR framework can be further applied to any other sensor providing two-dimensional images of snow crystals (binary or gray scale).
- Stability and representativeness of the classifiers were investigated and validated by means of learning curves. It was shown that the amount of data used to train the classification models was sufficient, suggesting that the models should reliably generalize to new data.
- Depending on the desired level of detail, subclassifications can be achieved in a nested fashion. This feature is particularly relevant for determining the composition of crystal aggregates. Subclassification performed on a flight period also proved to be relevant to study under which specific conditions different habits are observed.
- The forward feature selection algorithm highlighted two descriptors common to the three classifiers and therefore essential for habit classification: the particle area to convex hull area ratio and the sixth component of the standardized distance to centroid Fourier power spectrum. Moreover, each classifier included at least two other types of area ratio (e.g., based on circumscribed circle area, bounding box area, and smallest encompassing ellipse area), suggesting that combining different area ratio definitions allows for a better identification of the particle habit. Features assessing the degree of self-similarity in the particle perimeter were shown to be relevant in each classifier.

- Classification output extracted from data collected by the UND Citation research aircraft equipped with the three probes was used to conduct a comparative analysis. Results showed that the habits observed by the three probes were consistent and in good agreement with the environmental conditions recorded, although the exact proportions differed. Consistency in these proportions is, however, not expected as each probe can only identify habits in the specific size range of particles it images, thus providing complementary information. Combined with state-of-the-art PSD retrieval techniques, the proposed particle habit classification technique is therefore suitable for a detailed shape and size analysis applied on a broad spectrum of particle dimension.
- The textural information brought by the grayscale images collected by the CPI probe appear to have only a marginal impact on the classification accuracy (<3% improvement in the HSS). In contrast to MASC images where textural descriptors were utilized to estimate the degree of riming of the particles (Praz et al., 2017), background noise and contrast within CPI images are too variable to automatically retrieve information on riming. If these issues could be fixed in the future, the possibility to identify riming degree would offer a great potential to study riming mechanisms and particle morphological evolution as well as to relate these observations with the amount of supercooled liquid water measured by other sensors.

In summary, the present study introduces a general framework based on the MLR model to identify and classify ice particle habit based on two-dimensional images. The method proved to perform well on a broad range of imaging devices and will allow for a systematic and consistent classification of airborne particle images, relevant for microphysical studies in the future. Indeed, the classification provides relevant information to investigate habit growth and transition as the local environmental conditions and cloud properties vary. It can also be used to refine cloud parametrization in radiative studies. In both ground and airborne studies, habit classification is useful to compare and validate hydrometeor classification algorithms based on radar products.

Appendix A: Automatic Detection of the Bright Spot Located at the Center of Water Droplets in CPI Images

As mentioned in section 4.4, CPI images of liquid droplets are frequently characterized by the presence of a small bright spot located at the center of the image. It is a typical signature observed in liquid droplet images when the particles are illuminated from behind (Saylor et al., 2002). The presence of a bright spot provides a clear and unambiguous way to distinguish between small quasi-spherical ice particles and water droplets. It is therefore desirable to detect the presence of such a bright spot automatically. The procedure for achieving this is outlined in this appendix.

First, the particle outline is detected and background noise is removed using standard image processing techniques. In the second step, the brightest pixel in the 3×3 window around the particle centroid is calculated and denoted C^*. The geometric distance d_{ij} to C^* is then computed for every pixel located within the particle outline. The pixel intensity I_{ij} is then plotted as a function of its distance from C^* in Figure A1. Three radii $r_1,\ r_2$ and r_3 are then defined. r_1 represents the maximum extent that the bright spot can take and was set to 3 pixels in this study. r_2 and r_3 delimit a reference annular area within the particle and are defined as

$$r_2 = 0.25 \cdot \max(d_{ij}),\ r_3 = 0.75 \cdot \max(d_{ij}). \tag{A1}$$

The detection of the bright spot relies on the comparison of pixels with $d_{ij} < r_1$ to pixels with $r_2 < d_{ij} < r_3$. By denoting these two ensembles I_{BR} and I_{ref}, respectively, the bright spot detection index γ_{BR} is constructed as

$$\gamma_{\mathrm{BR}} = \frac{\tilde{I}_{\mathrm{BR}} - \tilde{I}_{\mathrm{ref}}}{\min\left(\tilde{I}_{\mathrm{BR}}, \tilde{I}_{\mathrm{ref}}\right)}, \tag{A2}$$

$$I_{\mathrm{BR}} = \left\{ I_{ij} \text{ such as } d_{ij} < r_1 \right\}, \tag{A3}$$

$$I_{\mathrm{ref}} = \left\{ I_{ij} \text{ such as } r_2 < d_{ij} < r_3 \right\}, \tag{A4}$$

where $\tilde{I}$ is the median pixel intensity in the ensemble I. In this way, the sign of γ_{BR} indicates whether a bright spot is detected, and its numerical value gives an estimate of the brightness of this spot. In order to avoid false

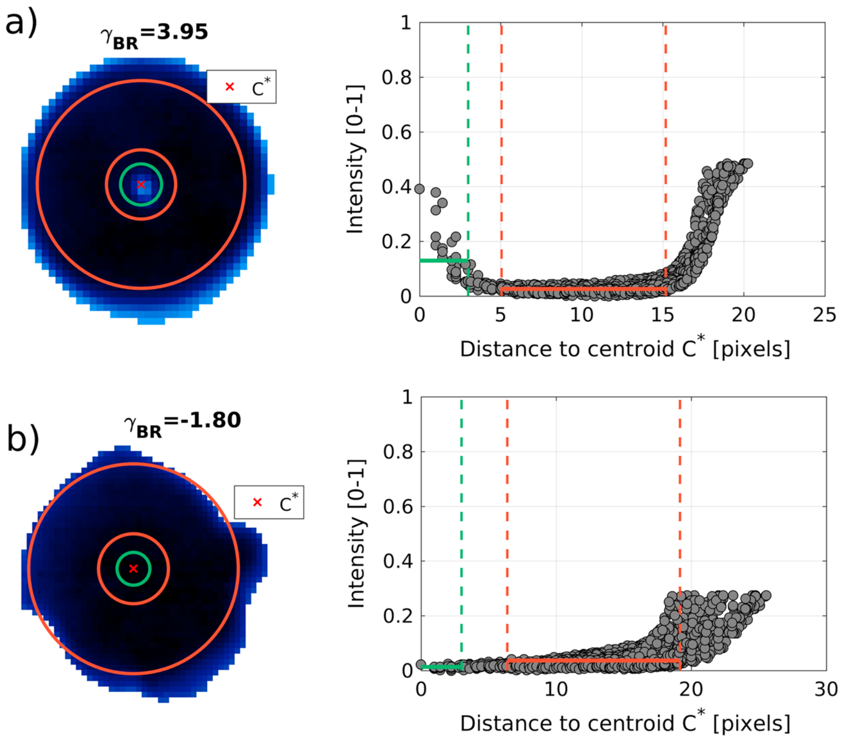

Figure A1. Illustration of the bright spot automatic detection procedure applied to two cloud particle imager images collected during F1. In both examples, the ensemble I_{BR} is delimited by the green circle on the left column and its median intensity value is illustrated as a bold green line on the right column. Similarly, the ensemble I_{ref} is delimited by the red annulus on the left column and its median intensity value is illustrated as a bold red line on the right column. (a) A typical example of a water droplet with a bright spot and (b) a compact particle without bright spot.

detection that could happen with highly transparent planar crystal for instance, the following set of criteria is used to identify the presence of a bright spot:

$$\text{water droplet bright spot} \Leftrightarrow \begin{cases} \gamma_{BR} > 0.5, \\ \alpha > 0.8, \\ \sigma\left(I_{ref}\right) < 0.04, \end{cases} \tag{A5}$$

with α the particle area to circumscribed circle area ratio and $\sigma\left(I_{ref}\right)$ the standard deviation of the pixel intensity calculated within the ensemble I_{ref}. This method was able to detect the water droplets identifiable by eye (highly spherical shape and presence of a bright spot) in the CPI labeled data set (i.e., 64 instances among 2,964 images) with a BER lower than 2%.

Acknowledgments
We are very grateful to Jussi Tiira from University of Helsinki for sharing with us and explaining how to use the *IC-PCA* libraries utilized to calculate the particle descriptors based on the notion of normalized curvilinear distance and line segment. This work was supported by the Swiss National Foundation (SNF), project 20021_157210 and by NASA Precipitation Measurement Missions grant NNX16AD80G. The MATLAB codes used to process and classify ice crystal images can be made publicly available upon request to the authors. OLYMPEX OAP data are available online, http://ghrc.nsstc.nasa.gov, from the NASA EOSDIS Global Hydrology Resource Center Distributed Active Archive Center Huntsville, Alabama, USA (Poellot et al., 2017).

References

Bailey, M. P., & Hallett, J. (2009). A comprehensive habit diagram for atmospheric ice crystals: Conformation from the laboratory, AIRS II, and other field studies. *Journal of Atmospheric Sciences*, *66*(9), 2888–2899. https://doi.org/10.1175/2009JAS2883.1

Bailey, M., & Hallett, J. (2012). Ice crystal linear growth rates from −20 to −70 °C: Confirmation from wave cloud studies. *Journal of the Atmospheric Sciences*, *69*(1), 390–402.

Baum, B. A., Heymsfield, A. J., Yang, P., & Bedka, S. T. (2005). Bulk scattering properties for the remote sensing of ice clouds. Part I: Microphysical data and models. *Journal of Applied Meteorology*, *44*(12), 1885–1895.

Baumgardner, D., Abel, S., Axisa, D., Cotton, R., Crosier, J., Field, P., et al. (2017). Cloud ice properties: In situ measurement challenges. *Meteorological Monographs*, *58*, 9–1.

Baumgardner, D., Brenguier, J. L., Bucholtz, A., Coe, H., DeMott, P., Garrett, T. J., et al. (2011). Airborne instruments to measure atmospheric aerosol particles, clouds and radiation: A cook's tour of mature and emerging technology. *Atmospheric Research*, *102*(1), 10–29. https://doi.org/10.1016/j.atmosres.2011.06.021

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bringi, V. N., Chandrasekar, V., Hubbert, J., Gorgucci, E., Randeu, W. L., & Schoenhuber, M. (2003). Raindrop size distribution in different climatic regimes from disdrometer and dual-polarized radar analysis. *Journal of Atmospheric Sciences*, *60*(2), 354–365.

Cober, S. G., Isaac, G. A., Korolev, A. V., & Strapp, J. W. (2001). Assessing cloud-phase conditions. *Journal of applied meteorology*, *40*(11), 1967–1983.

Cooper, S. J., Wood, N. B., & L'Ecuyer, T. S. (2017). A variational technique to estimate snowfall rate from coincident radar, snowflake, and fall-speed observations. *Atmospheric Measurement Techniques*, *10*(7), 2557.

Fealy, R., & Sweeney, J. (2007). Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach. *International Journal of Climatology*, *27*(15), 2083–2094.

Feind, R. E. (2006). Comparison of three classification methodologies for 2D probe hydrometeor images obtained from the armored T-28 aircraft, South Dakota School of Mines and Technology.

Garrett, T. J., Fallgatter, C., Shkurko, K., & Howlett, D. (2012). Fall speed measurement and high-resolution multi-angle photography of hydrometeors in free fall. *Atmospheric Measurement Techniques*, *5*(11), 2625–2633. https://doi.org/10.5194/amt-5-2625-2012

Garrett, T. J., & Yuter, S. E. (2014). Observed influence of riming, temperature, and turbulence on the fallspeed of solid precipitation. *Geophysical Research Letters*, *41*, 6515–6522. https://doi.org/10.1002/2014GL061016

Garrett, T. J., Yuter, S. E., Fallgatter, C., Shkurko, K., Rhodes, S. R., & Endries, J. L. (2015). Orientations and aspect ratios of falling snow. *Geophysical Research Letters*, *42*, 4617–4622. https://doi.org/10.1002/2015GL064040

Grazioli, J., Tuia, D., Monhart, S., Schneebeli, M., Raupach, T., & Berne, A. (2014). Hydrometeor classification from two-dimensional video disdrometer data. *Atmospheric Measurement Techniques*, *7*(9), 2869–2882. https://doi.org/10.5194/amt-7-2869-2014

Guo, H., & Gelfand, S. B. (1992). Classification trees with neural network feature extraction. *IEEE Transactions on Neural Networks*, *3*(6), 923–933.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, *6*, 610–621.

Heymsfield, A. J., Bansemer, A., Schmitt, C., Twohy, C., & Poellot, M. R. (2004). Effective ice particle densities derived from aircraft data. *Journal of the atmospheric sciences*, *61*(9), 982–1003.

Heymsfield, A. J., Lewis, S., Bansemer, A., Iaquinta, J., Miloshevich, L. M., Kajikawa, M., et al. (2002). A general approach for deriving the properties of cirrus and stratiform ice cloud particles. *Journal of Atmospheric Sciences*, *59*(1), 3–29.

Heymsfield, A. J., Schmitt, C., & Bansemer, A. (2013). Ice cloud particle size distributions and pressure-dependent terminal velocities from in situ observations at temperatures from 0 to −86 °C. *Journal of the Atmospheric Sciences*, *70*(12), 4123–4154.

Heymsfield, A. J., & Westbrook, C. (2010). Advances in the estimation of ice particle fall speeds using laboratory and field measurements. *Journal of the Atmospheric Sciences*, *67*(8), 2469–2482.

Hogan, R. J., Tian, L., Brown, P. R., Westbrook, C. D., Heymsfield, A. J., & Eastment, J. D. (2012). Radar scattering from ice aggregates using the horizontally aligned oblate spheroid approximation. *Journal of Applied Meteorology and Climatology*, *51*(3), 655–671.

Holroyd, E. W. (1987). Some techniques and uses of 2D-C habit classification software for snow particles. *Journal of Atmospheric and Oceanic Technology*, *4*(3), 498–511.

Houze, R. A., McMurdie, L. A., Petersen, W. A., Schwaller, M. R., Baccus, W., Lundquist, J., et al. (2017). *The Olympic Mountains Experiment (OLYMPEX)*, vol. 98, pp. 2167–2188.

Hunter, H. E., Dyer, R. M., & Glass, M. (1984). A two-dimensional hydrometeor machine classifier derived from observed data. *Journal of Atmospheric and Oceanic Technology*, *1*(1), 28–36.

Jackson, R. C., & McFarquhar, G. M. (2014). An assessment of the impact of antishattering tips and artifact removal techniques on bulk cloud ice microphysical and optical properties measured by the 2D cloud probe. *Journal of Atmospheric and Oceanic Technology*, *31*(10), 2131–2144.

Jiang, Z., Oue, M., Verlinde, J., Clothiaux, E. E., Aydin, K., Botta, G., & Lu, Y. (2017). What can we conclude about the real aspect ratios of ice particle aggregates from two-dimensional images? *Journal of Applied Meteorology and Climatology*, *56*(3), 725–734.

Kennedy, P., Thurai, M., Praz, C., Bringi, V., Berne, A., & Notaroš, B. M. (2018). Variations in snow crystal riming and ZDR: A case analysis. *Journal of Applied Meteorology and Climatology*, *57*(3), 695–707.

Kneifel, S., Kollias, P., Battaglia, A., Leinonen, J., Maahn, M., Kalesse, H., & Tridon, F. (2016). First observations of triple-frequency radar doppler spectra in snowfall: Interpretation and applications. *Geophysical Research Letters*, *43*, 2225–2233. https://doi.org/10.1002/2015GL067618

Kollias, P., Clothiaux, E. E., Miller, M. A., Albrecht, B. A., Stephens, G. L., & Ackerman, T. P. (2007). Millimeter-wavelength radars—New frontier in atmospheric cloud and precipitation research. *Bulletin of the American Meteorological Society*, *88*(10), 1608–1624. https://doi.org/10.1175/BAMS-88-10-1608

Korolev, A., & Isaac, G. (2003). Roundness and aspect ratio of particles in ice clouds. *Journal of the atmospheric sciences*, *60*(15), 1795–1808.

Korolev, A., Isaac, G. A., & Hallett, J. (2000). Ice particle habits in stratiform clouds. *Quarterly Journal of the Royal Meteorological Society*, *126*(569), 2873–2902.

Korolev, A., & Sussman, B. (2000). A technique for habit classification of cloud particles. *Journal of Atmospheric and Oceanic Technology*, *17*(8), 1048–1057.

Kulie, M. S., Hiley, M. J., Bennartz, R., Kneifel, S., & Tanelli, S. (2014). Triple-frequency radar reflectivity signatures of snow: Observations and comparisons with theoretical ice particle scattering models. *Journal of Applied Meteorology and Climatology*, *53*(4), 1080–1098. https://doi.org/10.1175/JAMC-D-13-066.1

Lawson, R. P. (2011). Effects of ice particles shattering on the 2D-S probe. *Atmospheric Measurement Techniques*, *4*(7), 1361.

Lawson, R. P., Baker, B. A., Zmarzly, P., OâĂŹConnor, D., Mo, Q., Gayet, J.-F., & Shcherbakov, V. (2006). Microphysical and optical properties of atmospheric ice crystals at South Pole station. *Journal of Applied Meteorology and Climatology*, *45*(11), 1505–1524.

Lawson, R. P., O'Connor, D., Zmarzly, P., Weaver, K., Baker, B., Mo, Q., & Jonsson, H. (2006). The 2D-S (stereo) probe: Design and preliminary tests of a new airborne, high-speed, high-resolution particle imaging probe. *Journal of Atmospheric and Oceanic Technology*, *23*(11), 1462–1477.

Lawson, R. P., Stewart, R. E., Strapp, J. W., & Isaac, G. A. (1993). Aircraft observations of the origin and growth of very large snowflakes. *Geophysical research letters*, *20*(1), 53–56.

Leinonen, J., Kneifel, S., Moisseev, D., Tyynelä, J., Tanelli, S., & Nousiainen, T. (2012). Evidence of nonspheroidal behavior in millimeter-wavelength radar observations of snowfall. *Journal of Geophysical Research*, *117*, D18205. https://doi.org/10.1029/2012JD017680

Libbrecht, K. G. (2005). The physics of snow crystals. *Reports on Progress in Physics*, *68*(4), 855–895. https://doi.org/10.1088/0034-4885/68/4/R03

Lindqvist, H., Muinonen, K., Nousiainen, T., Um, J., McFarquhar, G., Haapanala, P., et al. (2012). Ice-cloud particle habit classification using principal components. *Journal of Geophysical Research*, *117*, D16206. https://doi.org/10.1029/2012JD017573

Logvin, A., Ligthart, L., & Kozlov, A. (2002). Methods for solving inverse problems in radar remote sensing. In *Microwaves, Radar and Wireless Communications, 2002, MIKON-2002. 14th International Conference on*, *2*, pp. 681–685. IEEE

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Magono, C., & Lee, C. W. (1966). Meteorological classification of natural snow crystals. *Journal of the Faculty of Science, Hokkaido University. Series 7*, *2*(4), 321 – 335.

McFarquhar, G. M., Baumgardner, D., Bansemer, A., Abel, S. J., Crosier, J., French, J., et al. (2017). Processing of ice cloud in situ data collected by bulk water, scattering, and imaging probes: Fundamentals, uncertainties, and efforts toward consistency. *Meteorological Monographs*, *58*, 11.1 – 11.33. https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0007.1

McFarquhar, G. M., & Heymsfield, A. J. (1996). Microphysical characteristics of three anvils sampled during the central equatorial Pacific experiment. *Journal of the atmospheric sciences*, *53*(17), 2401 – 2423.

McFarquhar, G. M., Heymsfield, A. J., Macke, A., Iaquinta, J., & Aulenbach, S. M. (1999). Use of observed ice crystal sizes and shapes to calculate mean-scattering properties and multispectral radiances: CEPEX April 4, 1993, case study. *Journal of Geophysical Research*, *104*(D24), 31,763 – 31,779.

McFarquhar, G. M., Um, J., Freer, M., Baumgardner, D., Kok, G. L., & Mace, G. (2007). Importance of small ice crystals to cirrus properties: Observations from the Tropical Warm Pool International Cloud Experiment (TWP-ICE). *Geophysical research letters*, *34*, L13803. https://doi.org/10.1029/2007GL029865

McFarquhar, G. M., Um, J., & Jackson, R. (2013). Small cloud particle shapes in mixed-phase clouds. *Journal of Applied Meteorology and Climatology*, *52*(5), 1277 – 1293.

Morrison, H., & Milbrandt, J. A. (2015). Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *Journal of the Atmospheric Sciences*, *72*, 287 – 311.

Moss, S., & Johnson, D. (1994). Aircraft measurements to validate and improve numerical model parametrisations of ice to water ratios in clouds. *Atmospheric research*, *34*(1-4), 1 – 25.

Nousiainen, T., & McFarquhar, G. M. (2004). Light scattering by quasi-spherical ice crystals. *Journal of the Atmospheric Sciences*, *61*(18), 2229 – 2248. https://doi.org/10.1175/1520-0469(2004)061<2229:LSBQIC>2.0.CO;2

Nurzynska, K., Kubo, M., & Muramoto, K.-i. (2012). Texture operator for snow particle classification into snowflake and graupel. *Atmospheric Research*, *118*, 121 – 132. https://doi.org/10.1016/j.atmosres.2012.06.013

Nurzynska, K., Kubo, M., & Muramoto, K.-i. (2013). Shape parameters for automatic classification of snow particles into snowflake and graupel. *Meteorological Applications*, *20*(3), 257 – 265. https://doi.org/10.1002/met.299

Ono, A. (1969). The shape and riming properties of ice crystals in natural clouds. *Journal of the Atmospheric Sciences*, *26*(1), 138 – 147.

Poellot, M. R., Heymsfield, A. J., & Bansemer, A (2017). *GPM ground validation und citation cloud microphysics OLYMPEX, dataset available online from the NASA Global Hydrology Center DAAC, Huntsville*. Alabama, U.S.A. https://doi.org/10.5067/GPMGV/OLYMPEX/MULTIPLE/DATA201

Praz, C., Roulet, Y.-A., & Berne, A. (2017). Solid hydrometeor classification and riming degree estimation from pictures collected with a Multi-Angle Snowflake Camera. *Atmospheric Measurement Techniques*, *10*(4), 1335 – 1357. https://doi.org/10.5194/amt-10-1335-2017

Protat, A., Mcfarquhar, G. M., Um, J., & Delanoë, J. (2011). Obtaining best estimates for the microphysical and radiative properties of tropical ice clouds from twp-ice in situ microphysical observations. *Journal of Applied Meteorology and Climatology*, *50*(4), 895 – 915.

Saylor, J., Jones, B., & Bliven, L. (2002). A method for increasing depth of field during droplet imaging. *Review of scientific instruments*, *73*(6), 2422 – 2427.

Schmitt, C. G., & Heymsfield, A. J. (2014). Observational quantification of the separation of simple and complex atmospheric ice particles. *Geophysical Research Letters*, *41*, 1301 – 1307. https://doi.org/10.1002/2013GL058781

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37.

Testud, J., Le Bouar, E., Obligis, E., & Ali-Mehenni, M. (2000). The rain profiling algorithm applied to polarimetric weather radar. *Journal of Atmospheric and Oceanic Technology*, *17*(3), 332 – 356.

Um, J., & McFarquhar, G. M. (2007). Single-scattering properties of aggregates of bullet rosettes in cirrus. *Journal of applied meteorology and climatology*, *46*(6), 757 – 775.

Um, J., & McFarquhar, G. M. (2009). Single-scattering properties of aggregates of plates. *Quarterly Journal of the Royal Meteorological Society*, *135*(639), 291 – 304.

Um, J., & McFarquhar, G. (2011). Dependence of the single-scattering properties of small ice crystals on idealized shape models. *Atmospheric Chemistry and Physics*, *11*(7), 3159 – 3171.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution mos forecasts. *Meteorological Applications*, *16*(3), 361 – 368.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques.

Zhang, F., Du, B., & Zhang, L. (2015). Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(4), 2175 – 2184.