Hydrometeor classification from polarimetric radar measurements: a clustering approach

J. Grazioli¹, D. Tuia², and A. Berne¹

¹Environmental Remote Sensing Laboratory (LTE), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
²Laboratory of Geographic Information Systems (LASIG), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Received: 24 July 2014 – Accepted: 1 August 2014 – Published: 19 August 2014

Correspondence to: A. Berne (alexis.berne@epfl.ch)

Published by Copernicus Publications on behalf of the European Geosciences Union.
Abstract

A data-driven approach to the classification of hydrometeors from measurements collected with polarimetric weather radars is proposed. In a first step, the optimal number $n_{\text{opt}}$ of hydrometeor classes that can be reliably identified from a large set of polarimetric data is determined. This is done by means of an unsupervised clustering technique guided by criteria related both to data similarity and to spatial smoothness of the classified images. In a second step, the $n_{\text{opt}}$ clusters are assigned to the appropriate hydrometeor class by means of human interpretation and comparisons with the output of other classification techniques. The main innovation in the proposed method is the unsupervised part: the hydrometeor classes are not defined a-priori, but they are learned from data. The proposed approach is applied to data collected by an X-band polarimetric weather radar during two field campaigns (totalling about 3000 h of precipitation). Seven hydrometeor classes have been found in the data set and they have been associated to drizzle (DZ), light rain (LR), heavy rain (HR), melting snow (MS), ice crystals/small aggregates (CR), aggregates (AG), rimed particles (RI).

1 Introduction

Hydrometeor classification (HC) from weather radar data refers to a family of techniques and algorithms that retrieve a qualitative information about precipitation: the dominant hydrometeor type within a given sampling volume, where the term “dominant” is used to underline that the actual hydrometeor content is usually a mixture. These methods use as input a set of quantitative measurements provided by the radar itself and some additional information coming from external sources such as vertical profiles of temperature or estimates of the 0°C isotherm height.

The classification is conducted at the spatial scale of the radar resolution volume and its inputs are usually a set of polarimetric variables like horizontal reflectivity $Z_H$,
differential reflectivity $Z_{DR}$, correlation coefficient $\rho_{hv}$ and specific differential phase $K_{dp}$ (see Bringi and Chandrasekar, 2001; Berne and Krajewski, 2013, for definitions).

The most modern HC techniques require dual-polarization (and Doppler) capabilities. These allow to acquire with a single instrument multiple measurements, each one sensitive to distinct characteristics of precipitation, and facilitate the interpretation of many microphysical processes (e.g. Seliga and Bringi, 1976; Jameson, 1983; Vivekanandan et al., 1994; Ryzhkov et al., 2005; Bechini et al., 2013; Schneebeli et al., 2013).

Different HC algorithms are available for S-band (Straka et al., 2000; Liu and Chandrasekar, 2000), C-band (Marzano et al., 2007; Dolan et al., 2013) and X-band (Dolan and Rutledge, 2009; Snyder et al., 2010; Marzano et al., 2010) radar frequencies. Recently, after years of improvements, HC also became a common product provided operationally by national meteorological services (e.g. Gourley et al., 2007; Al-Sakka et al., 2013; Chandrasekar et al., 2013).

Most HC methods are based on similar principles: they start by selecting the number and type of hydrometeor classes undergoing classification. Then, through scattering simulations, the theoretical radar observations associated with these hydrometeor classes are reconstructed. Finally, actual observations are associated (labeled) with the appropriate class according to their degree of similarity with the sets of simulations available. This last step is usually conducted by means of a fuzzy logic input-output association (e.g. Dolan and Rutledge, 2009), or by means of Bayesian (Marzano et al., 2010) and neural network (Liu and Chandrasekar, 2000) techniques. In some cases these relations rely entirely on the simulation framework available (e.g., Dolan and Rutledge, 2009). In other cases, instead, they are adapted and modified in order to adequately reproduce actual observations (e.g. Marzano et al., 2007), or according to empirically-based constraints (e.g. Al-Sakka et al., 2013).

In the present manuscript we denote with capital subscripts the variables expressed in dB ($Z_H$ and $Z_{DR}$), and with lower-case subscripts the other ones ($\rho_{hv}$, $K_{dp}$, $\Phi_{dp}$, $\Psi_{dp}$).
The typical HC techniques mentioned above have become a state-of-the-art approach, stable and robust enough to be implemented operationally. However it is important to underline that this approach has some limitations, since it relies on strong assumptions. First, the choice of the hydrometeor classes in the classification scheme is mostly subjective. Secondly, the scattering simulations (e.g. Mishchenko et al., 1996), which are usually very accurate for rainfall, are largely uncertain for ice-phase hydrometeors, because of the complex geometries, dielectric properties and particle size distributions of ice particles (Tyynela et al., 2011). Finally, it is not easy to take into account the accuracy of actual radar measurements when comparing simulations and observations. In the present paper we propose a different approach to HC, in which the classifier is built on actually measured radar data instead of on numerical simulations.

A clustering technique, i.e., a technique that is used to find patterns (groups) in data sets in an unsupervised way (see Jain et al., 1999; Xu and Wunsch, 2005; Von Luxburg, 2007, for a complete overview) is applied to a database of precipitation measurements collected by an X-band dual-polarization radar. An optimal partition of these data into $n_{\text{opt}}$ separate groups is found as a trade-off between data similarity (in terms of polarimetric variables), and spatial smoothness of the partition. The content of these groups is then interpreted a-posteriori, and a hydrometeor class is assigned to each of them.

The manuscript is structured as follows. Section 2 provides some background on clustering algorithms and Sect. 3 presents the polarimetric data employed in the study. Section 4 describes the unsupervised part of the classification method and Sect. 5 is devoted to the identification of the optimal number of clusters in the data set. Section 6 deals with the labelling of the $n_{\text{opt}}$ clusters identified, and Sect. 7 presents the summary, discussion and conclusions.

2 Background on clustering techniques

The proposed approach to HC is data-driven. The first two necessary steps are therefore to identify groups (clusters) in the available data set, and then to define as
objectively as possible the optimal number of these groups. In this section we provide some background on the clustering methods that will be employed in the following.

2.1 Hierarchical data clustering

We define as unsupervised data clustering techniques all techniques that aim at organizing a given set of objects (observations) in a certain number of groups (clusters). The shape (or functional form) of these groups, as well as their number, is unknown a-priori (Jain et al., 2000).

We consider here a particular type of clustering technique: the agglomerative hierarchical clustering (Ward, 1963, AHC hereafter). AHC is a step-wise approach that is used to group a set of $N_D$ objects into $n_c$ clusters ($n_c \leq N_D$) in a way that objects belonging to the same cluster are more similar to each other than to those belonging to the others. The technique is called agglomerative because at a step $i$:

$$n_c^i = N_D - i.$$  

(1)

This means that at the initial step ($i = 0$) single objects populates the clusters, while at each step two objects (the most similar) are merged together, thus reducing the total number of clusters by one. The method is nested, in the sense that once two samples are grouped in the same cluster, they remain clustered in all the following levels of the hierarchy.

In order to define which objects are the most similar, two criteria need to be defined (Xu and Wunsch, 2005): (i) a metric i.e., a measure of distance between objects and (ii) a merging rule. At each step $i$ the pair of objects that are situated at the closest distance (according to a certain merging rule) are merged together.
2.2 Distance metric

Let \( x \) and \( y \) be two objects, defined in a \( d \) dimensional space. They have therefore \( d \) components:

\[
\begin{align*}
    x &= \{x[1], \ldots, x[d]\} \\
    y &= \{y[1], \ldots, y[d]\}
\end{align*}
\]

A list of common distance metrics \( D(x, y) \) between \( x \) and \( y \) is provided in Table 1. Each of these metrics is designed to capture a particular type of similarity between pairs of objects. To cite a few, the “Minkowski Distance”\(^2\), is a good index of similarity when all the \( d \) components of \( x \) and \( y \) have the same order of magnitude, while the “Correlative” distance is less affected by unbalanced components but might be ill-defined when \( d \) is small.

2.3 Merging rule

The second concept to introduce is the merging rule. A merging rule defines the criteria that an object \( x \), or a cluster of objects \( C_i \) (a group of objects \( x \in C_i \)), has to satisfy in order to be merged with another cluster \( C_j \). In other words, it generalizes the concept of distance between single objects of Table 1 to distances between two clusters, or between a cluster and a single object. The merging rules tested in the present work include the single linkage (SL), complete linkage (CL), weighted pairwise average (WPA, sometimes also defined as WPGMA), and weighted centroid (WC) distance (Jain and Dubes, 1988).

\(^2\)Note that this distance reduces to the well known Euclidean distance (if \( p = 2 \)) and the “City-block” distance (if \( p = 1 \)), according to the notation of Table 1.
– SL defines the distance between two clusters \((C_i, C_j)\) as the minimum distance between couples of objects belonging to these clusters:

\[
D(C_i, C_j) = \min_{x \in C_i, y \in C_j} D(x, y)
\]  

(2)

– CL selects the maximum distance between objects:

\[
D(C_i, C_j) = \max_{x \in C_i, y \in C_j} D(x, y)
\]  

(3)

– WPA defines the distance between objects as the average distance between couples of objects belonging to the two clusters, weighted by the number of object of each sub-cluster. In this case the definition of distance between clusters, employed as merging rule, is recursive. As an example, given \(C_i = C_K \cup C_L\):

\[
D(C_i, C_j) = D(C_{K \cup L}, C_j) = \frac{n_K D(C_K, C_j) + n_L D(C_L, C_j)}{n_K + n_L}
\]  

(4)

where \(n_K\) and \(n_L\) are the number of objects contained in the clusters \(C_K\) and \(C_L\), respectively.

– WC defines the distance between clusters as the distance between the (weighted) centroids of each cluster. The centroid is the centre of mass of a cluster \(C_i\). It is computed as the average position of all the sub-clusters \(C_K \subset C_i\), weighted by the number of objects in each \(C_K\). Thus:

\[
D(C_i, C_j) = D(\bar{x}_{C_i}, \bar{x}_{C_j})
\]  

(5)

where \(\bar{x}_{C_i}\) is the weighted centroid of cluster \(C_i\), defined as:

\[
\bar{x}_{C_i} = \frac{\sum_{C_K \subset C_i} n_K \sum_{x \in C_K} x}{n_i}
\]  

(6)
Even though different distance metrics and different merging criteria can be defined, all hierarchical cluster methods start with \(N\) objects distributed into \(N\) clusters, and they end with \(N\) objects in one single cluster. The key point of any clustering method is therefore the selection of the optimal intermediate partition, named \(n_{\text{opt}}\), between the starting and the ending point. A universally applicable criterion to perform this choice does not exist, and the selection is usually based both on quantitative cluster quality indices and on the available prior knowledge about the data undergoing clustering (Kovács and Iváncsy, 2006; Wilks, 2011).

3 Data and processing

The present section provides a description of the data employed in the following analysis, and some details about data processing.

3.1 Data source

The polarimetric radar data considered here were collected with an X-band dual-polarization Doppler weather radar (MXPol), whose characteristics are summarized in Table 2.

During any measurement period MXPol was deployed at single locations for the whole duration of the measurements. This ensures that data collected during different days are referenced with respect to the exact same location. The pointing accuracy of the mechanical antenna was adjusted by means of a Sun-tracking scan, following the approach of Muth et al. (2012).

In the present work we employ radar data collected during two field deployments. The first one took place in Davos (CH), in the Swiss Alps, from September 2009 to July 2011. The radar was deployed at 2133 m a.s.l. on a ski slope dominating the valley of Davos, as shown in Fig. 1a. The altitude of the deployment site allowed to collect many observations of ice-phase precipitation when the radar itself was located above
the melting layer, and therefore not suffering from liquid-water signal attenuation. Given the complex topography, the scanning sequence included multiple Plan Position Indicator (PPI) sector scans over the valley of Davos, at elevation angles ranging from 9° to 27°, a Range Height Indicator (RHI), and a vertically-pointing PPI used for the calibration of the differential reflectivity $Z_{\text{DR}}$.

The second field deployment took place in the Ardèche region (FR) from September to November 2012, at an altitude of 605 m a.s.l. This deployment was part of the HyMeX experiment (www.hymex.org, Bousquet et al., 2014). Stratiform and convective Mediterranean precipitation events were sampled during this campaign. The scanning sequence in this case included large sector scans (spanning 200° in azimuth) conducted at elevations ranging from 3.5° to 10° (Fig. 1b). Additionally, RHIs towards different directions as well as a $Z_{\text{DR}}$ calibration zenithal PPI were collected within each scanning sequence.

## 3.2 Polarimetric data

The polarimetric variables calculated from the measurements of MXPol and employed in the following analysis are: $Z_H$ [dBZ], $Z_{\text{DR}}$ [dB], $K_{dp}$ [$^\circ$ km$^{-1}$] and $\rho_{hv}$ [–].

PPI data of $Z_H$ and $Z_{\text{DR}}$ collected in rain are corrected for attenuation using the relations linking $K_{dp}$, $Z_H$, specific horizontal attenuation $\alpha_H$ [dBZ km$^{-1}$], and differential attenuation $\alpha_{\text{dr}}$ [dB km$^{-1}$] according to the method of Testud et al. (2000). The power laws between these variables are parametrized using disdrometer measurements for the data collected in France (Fig. 1b) and using simulated realistic drop size distribution fields (Schleiss et al., 2012) for the data collected in Switzerland. The set of observations corresponding to events during which the radar was located above the melting layer were not corrected for attenuation, assuming the attenuation in dry snow to be negligible.

$K_{dp}$ is estimated from the total differential phase shift $\Psi_{dp}$ [$^\circ$] using a method based on Kalman filtering (Schneebeli et al., 2014). The approach is designed to guarantee
the independence between $K_{dp}$ estimates and other polarimetric variables, and to capture the fine-scale variations of $K_{dp}$. All the polarimetric variables are censored with a mask of signal-to-noise ratio SNR > 8 dB, and all the radar resolutions volumes potentially contaminated by ground clutter are censored as well.

4 Clustering of polarimetric radar data

Hierarchical clustering is applied to radar observations (objects) $x$, defined in the multidimensional space of the polarimetric variables. Here we present in detail our clustering approach and we apply it to the database of Sect. 3.

4.1 Data preparation

The data object $x$ is a five dimensional vector defined for each valid radar resolution volume. The components of $x$ are therefore:

$$x = \{Z_H, Z_{DR}, K_{dp}, \rho_{hv}, \Delta z\}. \quad (7)$$

The last component ($x[5] = \Delta z$) is not a polarimetric variable and it is defined as:

$$\Delta z_i = z_i - z_{0^\circ} \quad (8)$$

where $z_i$ [m] is the altitude above sea level of the $i$th resolution volume, and $z_{0^\circ}$ is the estimated altitude of the 0°C isotherm, taken as a reference. A positive $\Delta z$ refers to a measurement collected at temperature ranges where ice-phase hydrometeors are expected, while a negative one to a measurement likely taken in liquid-phase precipitation. This variable is used as a prior information for the clustering algorithm, in order to take into account the approximate environmental conditions associated to each measurement. $z_{0^\circ}$ is estimated by means of linear interpolation of the data coming from ground-based weather stations located at different altitudes and at a distance $\leq 40$ km.
from the radar location. It can also be estimated from radar data directly, when a melting layer is sampled.

The vector $\mathbf{x}$ is not yet suitable to undergo cluster analysis. Two different issues need to be tackled.

1. The first issue is related to the distribution of $K_{dp}$ values: at X-band, $K_{dp}$ ranges approximately from $-1$ to $25^\circ \text{km}^{-1}$ (e.g. Dolan and Rutledge, 2009; Otto and Russchenberg, 2011) but its probability distribution, calculated over a large set of observations, is positively skewed with typical modal values below $0.5^\circ \text{km}^{-1}$. This issue is tackled by log-transforming $K_{dp}$ values before performing the cluster analysis. Values of $K_{dp}$ lower than $0.005^\circ \text{km}^{-1}$ (corresponding in rain to $R \approx 0.2 \text{mm h}^{-1}$, Otto and Russchenberg, 2012) are set to $0.005^\circ \text{km}^{-1}$ and $K_{dp}$ is then transformed with a natural logarithm. By proceeding this way, negative $K_{dp}$ values with actual physical meaning (e.g., vertically aligned ice, Dolan and Rutledge, 2009) might be lost. However, it was verified that in the data set employed the occurrence of negative $K_{dp}$ values ($K_{dp} < -0.5^\circ \text{km}^{-1}$) is very low ($< 0.5\%$). Furthermore, slightly negative values ($-0.5 < K_{dp} < 0^\circ \text{km}^{-1}$), cannot be discriminated from positive ones because of the uncertainties associated with the estimation of $K_{dp}$ (see Grazioli et al., 2014a, for the expected boundaries of accuracy around $K_{dp}$).

2. The different components of $\mathbf{x}$, due to the differences in their units, have typical range of values that differ by several orders of magnitude. For instance, $Z_H$ can vary over tens of dBZ, while $Z_{DR}$ and $K_{dp}$ are smaller by one order of magnitude and $\rho_{hv}$ even by two orders of magnitude. This issue is tackled by means of data standardization (stretching). Even though a classical approach would be to use a z-score transformation, based on mean and standard deviation of a sample of data (e.g. Wilks, 2011), we selected a method based on minimum and maximum boundaries, that allows to pre-select physically relevant thresholds. The
components $x[i]^*$ of the standardized data are obtained as:

$$x[i]^* = \frac{x[i] - x_{\text{min}}[i]}{x_{\text{max}}[i] - x_{\text{min}}[i]} \quad i \in \{1, 2, 3, 4\}$$

(9)

where $x_{\text{min}}[i]$ ($x_{\text{max}}[i]$) is a minimum (maximum) value assigned to each polarimetric variable. The boundaries employed in the present study are: $-10$ to $60$ dBZ for $Z_H$, $-1.5$ to $5$ dB for $Z_{DR}$, $-5.3$ to $3 \log(\text{°} \text{km}^{-1})$ for $\log(K_{dp})$, $0.7$ to $1 [-]$ for $\rho_{hv}$ ($\Delta z$ is considered in the next paragraph). Variations on the order of $\pm 20\%$ around the proposed boundaries have a negligible impact on the results presented in Sect. 5.

$\Delta z$ is stretched within a smaller range of variation in the following way:

$$x[5]^* = \begin{cases} 
0 & \text{if } \Delta z \leq -400; \\
\kappa & \text{if } \Delta z > 400; \\
f(\Delta z) \times \kappa & \text{if } -400 < \Delta z \leq 400 
\end{cases}$$

(10)

$0 < \kappa \leq 1$.

$\kappa$ is a scaling factor and $f(\Delta z)$ denotes any monotonically increasing functional form that gives continuity to Eq. (10). Gaussian, sigmoid and logistic functions have been tested and appeared to be adequate. The reason behind a different standardization for $\Delta z$ is to reduce the weight of this non-polarimetric input in the clustering process: this parameter is intended only to flag positive and negative temperatures in a quasi-binary way and not to substitute the information provided by the polarimetric variables (therefore $\kappa$ is kept strictly $\leq 1$). $\kappa$ factors ranging between $0.3$ and $0.9$ lead to similar outputs, and an intermediate value of $0.5$ was used.

With the standardization detailed in Eqs. (9) and (10), the radar observations collected at each resolution volume are summarized by the observation vector $x^*$, whose entries are now expressed in a similar order of magnitude.
4.2 Subset undergoing clustering analysis

Agglomerative clustering algorithms are generally computationally expensive, because full distance matrices among all samples (and then groups) to be clustered are computed at each step of the hierarchical aggregation chain. Therefore we opted for the definition of the clusters using a representative subset of the data and then to assign the whole dataset to these clusters using a nearest cluster rule (e.g., Volpi et al., 2012).

About 50 precipitation events belonging to the data set of Sect. 3 were manually selected. These events cover the range of precipitation types observed with MXPol, and they are assumed to be a representative sample of mid-latitude temperate precipitation.

A random portion of data is taken from each of these 50 precipitation events, from PPI scans conducted at elevation angles between 3.5° and 10° (free of ground clutter contamination). A total of 20,000 observations $x^*$ are extracted and are used as input of the subsequent cluster analysis. Evaluation of the results over multiple choices of the initial random selection, as well as the exclusion of some manually selected precipitation events showed that the random sampling did not affect the outcome of the clustering technique presented in the next section.

4.3 Clustering algorithm: data similarity and spatial smoothness

An AHC is applied to the polarimetric data set of $x^*$ objects in order to obtain an optimal partition of the data into a set of clusters.

The technique is a trade-off between purely data-driven clustering, as it was described in Sect. 2a (i.e., that only looks for similarity in the 5-D feature space of $x^*$) and spatial smoothness of the partition in the physical space. In other words, groups of precipitation types should both contain objects that are similar among each other (data-wise) and that exhibit spatial consistency, since we assume spatial smoothness of the geographic distribution of precipitation types. Here, and in the following, we will refer to an Euclidean distance metric and WPA linkage. Similar results can be obtained with correlative distance metrics, and/or WC linkage criteria. The method developed
in the present paper is sketched in the flow chart of Fig. 2. Panel a of the figure is explained step by step in the following sections.

4.3.1 Step1 (Fig. 2a1)

Initially the 20 000 selected objects populate \( n_c = 20000 \) clusters. A first hierarchical aggregation is conducted on the data, until reaching a number of 1 000 clusters in the dataset. This step aims at merging the most similar objects before proceeding with more computationally expensive calculations.

4.3.2 Step2 (Fig. 2a2)

Given the remaining \( n_c = 1000 \) clusters, referred to as \( C_L (L = 1, \ldots n_c) \), we proceed to the classification of the entire PPI images from which the original 20 000 objects were extracted. Let \( x^*_p \notin C_L (L = 1, \ldots n_c) \) be an object taken from one of the PPI images, and not belonging to any cluster \( C_L \). This object is now classified into one of the \( n_c \) clusters available, specifically the one related to the minimal distance to it (according to the given merging rule). We proceed until all the objects of the PPI images are classified into one of the \( n_c \) clusters available. An example of a classified PPI image, corresponding to \( n_c = 7 \) can be found in Fig. 7.

At this point, we evaluate the spatial smoothness of the partition into \( n_c \) clusters. Each object \( x^*_p \) has been assigned to a cluster \( C_M (1 < M \leq n_c) \). We define now a local spatial smoothness index (SSI) associated with \( x^*_p \). This index evaluates the spatial consistency of the classification of an object with respect to the classification of neighbouring objects:

\[
SSI(x^*_p, C_M) = \frac{1}{n_{NN}} \sum_{i(p)=1}^{n_{NN}} \delta_{i(p)}
\]

(11)
where

\[
\delta_{i(p)} = \begin{cases} 
0 & \text{if } x^*_i(p) \notin C_M \\
1 & \text{if } x^*_i(p) \in C_M 
\end{cases}
\]

where \(n_{\text{NN}}\) (number of Nearest Neighbours) is the number of nearest objects considered in the construction of SSI, and \(x^*_i(p)\) indicates the \(i\)th nearest object of \(x^*_p\). In the present work \(n_{\text{NN}} = 4\), and very similar results are obtained for \(n_{\text{NN}} = 2, 4, 8\). The larger \(n_{\text{NN}}\), the further the smoothing effect for the spatial regularization will reach. The identification of the nearest neighbours is performed in polar coordinates and therefore the nearest object refers to the nearest radar resolution volume. Figure 3 illustrates this concept, by labelling the 8 nearest neighbours of a resolution volume centered on sample “0” on the figure.

SSI ranges between 0 and 1. If all the \(n_{\text{NN}}\) objects belong to cluster \(C_M\), then SSI is equal to 1. More generally, SSI is also defined if \(x^*_p \in C_Q (Q \neq M)\). In this case the index quantifies the occurrence of \(C_M\) objects around \(x^*_p\), rather than the spatial smoothness around it.

SSI indices are calculated for each \(x^*_p\) and they are summarized in a \(n_c \times n_c\) matrix \(M\), hereafter called spatial smoothness matrix. The elements \(M_{i,J}\) of \(M\) are defined as:

\[
M_{i,J} = \sum_{p=1}^{N_i} \text{SSI}(x^*_p, C_J)
\]

where \(N_i\) is the total number of objects \(x^*_p\) satisfying the condition \(x^*_p \in C_I\). The matrix \(M\) is conceptually similar to a confusion matrix, commonly used to evaluate the goodness of categorical classifications (e.g. Wilks, 2011). Diagonal entries \(M_{i,i}\) quantify the global spatial smoothness of the cluster \(C_I\), while the off-diagonal terms \(M_{i,J} (i \neq J)\) quantify the probability of objects belonging to a cluster \(C_I\) to be surrounded by objects of the cluster \(C_J\).
In analogy with a confusion matrix, the information contained in $M$ can be further summarized by means of quality indices. As an example, Cohen’s Kappa can be used to evaluate the global smoothness of a partition of the data-set into $n_c$ clusters. Cohen’s Kappa $[-]$ is defined as:

$$\text{Kappa} = \frac{\text{SSO} - S_{\text{est}}}{1 - S_{\text{est}}}$$  \hspace{1cm} (13)$$

where:

$$\text{SSO} = \frac{\sum_{i=1}^{n_c} M_{i,i}}{N}$$  \hspace{1cm} (14)$$

and

$$S_{\text{est}} = \frac{\left[ \sum_{i=1}^{n_c} \left( \sum_{j=1}^{n_c} M_{j,i} \right) \left( \sum_{j=1}^{n_c} M_{i,j} \right) \right]}{N^2}.$$  \hspace{1cm} (15)$$

$N$ is the total sum (over rows and columns) of all the elements of $M$. Kappa ranges from $-1$ to $1$ $[-]$ and increases as the level of global spatial smoothness increases. Furthermore, it is a robust estimator in case of unbalanced clusters. In fact it takes into account the globally observed spatial smoothness (SSO), and also the contribution occurring by chance, namely $S_{\text{est}}$. Finally, the individual smoothness of each cluster $C_M$ is quantified by means of the spatial smoothness per cluster ($SS_M$) index:

$$SS_M = \frac{M_{M,M}}{\sum_{i=1}^{n_c} M_{M,i}}$$  \hspace{1cm} (16)$$

A simple but useful example to clarify the concepts of $SS_M$ and spatial smoothness is reported in Appendix A.
4.3.3 Step 3 (Fig. 2a3)

At this stage the set of observations is divided into $n_c$ clusters, and the spatial smoothness of this partition has been evaluated. A classical hierarchical approach would now proceed by merging the two most (data-wise) similar clusters, reducing the total number of clusters to $n_c - 1$ at each iteration.

In our case, we make additional use of the information provided by Eq. (16). Let the cluster $C_W$, with the lowest spatial smoothness score be defined as:

$$C_W \text{ s.t. } SS_W = \min_{L=1,...,n_c} SS_L.$$  \hfill (17)

The cluster $C_W$ is forced to disappear, and it is merged with the most similar (data-wise) one according to the linkage method and the distance metric selected.

In this way, at each step of the AHC, spatial smoothness is used to identify the cluster that exhibit the highest spatial discontinuity (lowest spatial smoothness), while data similarity is used to merge it to one of the other $n_c - 1$ available clusters. The aggregative algorithm detailed in steps 1–3 recursively repeats step 2 and step 3, until $n_c = 2$. The reader should be aware that different constraints on spatial smoothness (or a direct inclusion of spatial smoothness indicators as additional dimensions of $x^*$) could be implemented at this stage and the constraint implemented here is a specific example.

5 Selection of the optimal cluster partition

The most delicate part of any unsupervised clustering method is the selection of a partition $n_{opt}$ that should be considered as an optimal one. This evaluation is conducted by taking into account the physical knowledge available for the process under investigation, and according to a set of cluster quality indices, that evaluate quantitatively the goodness of each possible partition (cf. Fig. 2b).
5.1 Cluster quality metrics

Different spatial cluster quality indices can be defined.

1. A first one is directly Kappa (Foody, 2004), defined in Eq. (13). Kappa quantifies the global degree of spatial smoothness of a given partition.

2. A second index, defined here as the accuracy spread index (AS), can be derived from Eq. (16) as follows:

\[
\text{AS} = \max_{L \in \{1, \ldots, n_c\}} SS_L - \min_{L \in \{1, \ldots, n_c\}} SS_L.
\]  

(18)

This index evaluates the inhomogeneity of the spatial characteristics of a partition into \(n_c\) clusters. The lower this score is, the more homogeneously the \(n_c\) clusters are performing in terms of spatial scores. Lower values are therefore associated with better partitions.

Other indices can be employed to evaluate the partition from the point of view of data similarity. Most of these indices evaluate the degree of homogeneity within cluster members and the degree of heterogeneity among different clusters. In the present work we employ two indices of this kind: RMSSTD and RS (e.g. Halkidi et al., 2002a, b; Kovács and Iváncsy, 2006).

3. The root mean square standard deviation (RMSSTD) is defined as:

\[
\text{RMSSTD} = \sqrt{\frac{\sum_{L=1}^{n_c} \sum_{k(L)=1}^{N_L} (\|x^*_k(L) - \bar{x}^*_C_L\|_2)^2}{d \sum_{L=1}^{n_c} (N_L - 1)}}
\]

(19)

where \(N_L\) is the number of objects in the \(L\)th cluster \(C_L\). \(x^*_k(L)\) and \(\bar{x}^*_C_L\) are the \(k\)th object and the mean value (centroid, see Eq. 6) of \(C_L\), respectively. \(d\) is the dimensionality of \(x^*\) and \(\|\|_2\) is the 2-norm, as in Table 1. RMSSTD quantifies cluster...
heterogeneity. Therefore good partitions (homogeneous clusters), are characterized by small RMSSTD.

4. The second data-based validity index is the $R$-squared index (RS), defined as:

$$RS = \frac{SSQ_t - SSQ_w}{SSQ_t}$$

where

$$SSQ_w = \sum_{L=1}^{n_c} \sum_{k(L)=1}^{N} \left( \|x^*_{k(L)} - \bar{x^*_C_L}\|_2 \right)^2$$

$$SSQ_t = \sum_{k=1}^{N} \left( \|x^*_k - \bar{x^*}\|_2 \right)^2$$

where $SSQ_t$ is the total variance of the dataset (with $N$ member objects), summed over all dimensions and $SSQ_w$ is the average variance within each cluster. RS is therefore an index of dissimilarity between clusters and higher values indicate better clustering.

5. Eventually, a simple RKR (RMSSTD and Kappa ratio) index is defined in order to combine spatial behaviour and data similarity:

$$RKR = \frac{\text{RMSSTD}}{\text{Kappa}}$$

where both indices are stretched in order to span exactly the same order of magnitudes in the interval $n_c = 1, \ldots, 1000$. The idea behind RKR is to find a trade-off between spatial smoothness (high Kappa), and data similarity (low RMSSTD). Therefore low values of RKR are associated with better partitions.
5.2 Evaluation

Figure 4 illustrates the behaviour of the quality indices defined as a function of the number of clusters in the dataset. Even though \( n_c \) ranges between 1 and 1000, a shorter interval is shown here, as we do not expect, given the goal of the clustering technique (i.e., hydrometeor classification), that any optimal solution would appear for very large \( n_c \).

An optimal solution is selected here when \( n_c = n_{\text{opt}} = 7 \) clusters. In fact, we can observe that \( n_c = 7 \) corresponds to a local minimum both for AS, and RKR. This suggests that, at this level of aggregation, the spatial behaviour of each cluster is homogeneous (low AS), and the global trade-off between data similarity and spatial smoothness is optimal (local minimum RKR). Furthermore, for \( n_c = 7 \), Kappa is high (\( \approx 0.8 \)) and RS is slightly larger than 0.5. By looking at the evolution of RS, we observe that for \( n_c \leq 7 \) the index tends to decrease at a higher rate. This suggests that \( n_c = 7 \) is an equilibrium point, where further merging of the clusters leads to high losses in terms of inter-cluster differences. Finally, we also observed that for any value \( n_c < 7 \), clusters related to positive \( \Delta z \) start to merge with clusters related to negative \( \Delta z \).

6 From unlabeled clusters to hydrometeor classes

This section is devoted to the interpretation of the output of the clustering algorithm (Fig. 2c).

6.1 Global characteristics of the clusters

The seven clusters identified contain a set of observations (or objects), that have been grouped according to spatial smoothness and data similarity. These clusters are defined in the 5-dimensional space given by the dimensions of \( x^* \). A possible way to visualize the clusters is to display pairs of 2-dimensional projections of the objects \( x \).
Some of these projections are displayed in Fig. 5, where the seven clusters are color-coded.

Three clusters contain data collected always at positive $\Delta z$ (negative temperatures), three clusters contain data collected always at negative $\Delta z$ (positive temperatures) and one cluster contains mostly data collected where $\Delta z \approx 0$, even though with a larger spread. In the following, we will proceed by interpreting the clustering results separately for clusters appearing on average at $\Delta z \leq 0$ and $\Delta z > 0$.

6.2 Clusters at positive temperature

Three clusters (red, green and dark blue) in Fig. 5 are identified at positive temperatures. It is therefore assumed that they are related to liquid-phase precipitation. In order to properly associate each of them to a more specific liquid precipitation category, further analysis is performed. At first, all data classified into one of these three categories are extracted from the whole field campaign of HyMex 2012 ($\approx 200$ h of precipitation recorded) from PPIs taken at elevation angles ranging between $3.5^\circ$ and $10^\circ$. Then, the rainfall rate $R$ [mm h$^{-1}$] associated with each measurement is computed by means of the following relations (Otto and Russchenberg, 2012):

$$R = \begin{cases} 13\zeta_{\text{dp}}^{0.75} & \text{if } Z_H > 30\text{dBZ} \\ \left(\frac{\zeta_H}{243}\right)^{1/1.24} & \text{if } Z_H \leq 30\text{dBZ} \end{cases}$$

where $\zeta_H = 10^{0.1Z_H}$ [mm$^2$ m$^{-3}$]. The distribution of $R$ stratified for each class is summarized in Table 3. The green cluster is characterized by extremely low rainfall intensity, and therefore it is associated hereafter to a hydrometeor class named Drizzle (DZ). The dark blue cluster is characterized by low to intermediate rainfall intensity, and therefore it is associated to a category named light-rain (LR). Finally, the red cluster contains by far the highest rainfall intensities, and it is named hereafter heavy-rain (HR).

As an additional test, the objects contained in the 7 clusters are classified with the Hydrometeor classification scheme derived from Dolan and Rutledge (2009), hereafter...
The classification of the DZ, LR and HR clusters obtained with the present method is compared with the output of DR2009 in Fig. 6. Even though DR2009 does not provide three “liquid-phase” hydrometeor classes we can observe that HR is entirely classified as rain, LR is mainly classified as rain, and DZ is almost entirely associated with drizzle (in agreement with the choice made).

Figure 7 illustrates a case where DZ, LR and HR are classified on the same PPI radar image, when a high intensity convective line was approaching the radar location from the west side of the domain, leading to a layer of high values of $Z_H$ (panel b), $Z_{DR}$ (panel c) and $K_{dp}$ (panel d), and to a complete extinction of the signal behind this area.

### 6.3 Cluster around 0 °C temperature

The yellow cluster of Fig. 5 appears on average around the 0 °C isotherm, and it is interpreted as melting snow (MS). Figure 8 shows an example of classification output, where a melting layer is clearly visible in the polarimetric observations. MS is delimiting the transition between ice-phase and liquid-phase hydrometeors, in the same way as it may be interpreted by visual inspection of the vertical stratification of $Z_H$ (Fig. 8, panel b), $Z_{DR}$ (panel c), and $\rho_{hv}$ (panel e). $K_{dp}$ (panel d) seems not to play a particular role in the classification of MS.

### 6.4 Clusters at negative temperatures

The clusters identified at negative temperature (dark green, pink, and light blue in Fig. 5), are more difficult to interpret and therefore more information is needed in order to associate them with an appropriate hydrometeor class. We proceed as follows: at first we examine the behaviour of the polarimetric variables within these three clusters, then we compare the classification with the output of the DR2009 algorithm, and eventually we compare the classification with the results from an in-situ based HC technique (Grazioli et al., 2014b) applied to two-dimensional video disdrometer (2DVD).
data, with quantitative information provided by the 2DVD itself and with the output of a numerical weather prediction model (COSMO).

6.4.1 Polarimetric signatures

Figure 9 presents the distribution of the polarimetric variables $Z_H$, $Z_{DR}$, $K_{dp}$, $\rho_{hv}$, and additionally the relative altitude $\Delta z$ for the three “ice-phase” clusters.

By looking at panel a, we can observe a clear $Z_H$ signature. $Z_H$ takes lower values for the light-blue cluster (mode $\approx 12$ dBZ), slightly higher values for the pink-cluster (mode $\approx 15$ dBZ) while the dark-green cluster is characterized by the highest $Z_H$ values (mode $> 20$ dBZ). This suggests that hydrometeor concentrations, size, and/or density are larger in this last cluster.

$Z_{DR}$, shown in panel b, exhibits a different trend. It is on average lower (but much more scattered) for the pink and light blue clusters, while slightly larger and less scattered for the dark-green one. This suggests that the light-blue and pink clusters are characterized by a higher variability of particle shapes, even though rather spherical on average (the mode of $Z_{DR}$ is around 0.1 dB for these two clusters). The green cluster exhibits a clear mode around 0.6 dB, indicating therefore a certain degree of geometrical anisotropy. The variability of shapes of the light-blue cluster can also be observed in panel d, where it is shown that $\rho_{hv}$ for this cluster departs often from unity.

An interesting feature is observed in panel c, where it is clear that the light-blue cluster has no $K_{dp}$ signature (all values are close to 0° km$^{-1}$), the pink cluster is characterized by values generally lower than 1° km$^{-1}$ and the dark-green one can reach relatively large values of 2.5° km$^{-1}$. $K_{dp}$ depends on size, concentration, shape and density of the particles in the radar resolution volume and therefore the dark green cluster contains on average more hydrometeors and/or hydrometeors of larger size and density.
Eventually, by looking at panel e, it can be observed that the dark-green cluster is found over a broad range of altitudes (temperatures), and that the light-blue cluster generally appears at colder temperatures than the pink one.

From this analysis, we observed that the three clusters exhibit distinct polarimetric signatures, which led us to assume the following: the light blue cluster corresponds to individual crystals and small aggregates (denoted CR), the pink cluster corresponds to aggregates (AG). Eventually, the dark green cluster corresponds to heavily rimed particles (RI). These assumptions are discussed in the following.

6.4.2 Comparison with DR2009

As a second step, DR2009 is applied to the objects classified into CR, AG and RI clusters (Fig. 10). It can be seen that, while the DR2009 “Aggregates” category is dominant both for the light blue (CR) and pink (AG) clusters, there is a larger component of vertically aligned ice (a type of small crystals) in the light-blue cluster, as well as a large component of crystals. The dark green cluster is split between aggregates ($\approx 15\%$), vertically aligned ice ($\approx 15\%$) and rimed particles (HDG + LDG $\approx 70\%$). These observations are in agreement with the previous assumptions.

6.4.3 Comparison with 2DVD classification output

An additional comparison is conducted with the output of a HC scheme developed for two-dimensional video disdrometers (2DVD, e.g. Kruger and Krajewski, 2002, for more information on the instrument). This method, hereafter called HC-2DVD is described in Grazioli et al. (2014b).

HC-2DVD is applied to the two-dimensional images collected by a 2DVD, and it provides an estimate of the dominant category of hydrometeor that was recorded by the instrument over time intervals of 60 s. Thus the classification is conducted spatially at the point-scale (disdrometer measurement), but it is averaged in time. HC-2DVD discriminates between eight hydrometeor classes: Small particle-like (SP), Dendrite-like (D),
Column-like (C), Graupel-like (G), Rimed particle-like (RIM), Aggregate-like (AG), Melting snow-like (MS), and Rain (R). The “-like” is added to underline that this approach identifies the dominant type of hydrometeor, and that there is in general a mixture of different hydrometeors that do not necessarily exhibit the pristine shape.

Here we compare HC-2DVD with the output of the clustering algorithm over all the snowfall events collected during the campaign of Davos 2009–2011 (Sect. 3). As observed in Fig. 1a, one 2DVD was located, during the campaign, at a projected distance of about 5.2 km from the location of MXPol. The PPI of lowest elevation and not contaminated by clutter was taken at 9° elevation with a repetition interval of five minutes. This PPI is used to compare with HC-2DVD.

Before discussing the comparison it must be kept in mind that: (i) the closest radar resolution volume center was about 400 m above the 2DVD (the spatial match is not perfect), (ii) the sampling times and volumes of the two instruments are different. Therefore this comparison is rather qualitative.

The comparison is conducted on a subset of about 30 snowfall events selected by means of visual inspection. The events with large temporal and spatial variability or poor/suspicious data quality are removed from the analysis, in order to avoid additional uncertainties in the comparison. Radar resolution volumes situated within 150 m in horizontal distance from the 2DVD location are selected and compared with the HC-2DVD output and a buffer of ±2 min is applied in order to match multiple 2DVD observations with a single radar scan.

Figure 11 presents the results obtained. We can observe that the characteristics of the CR (light-blue) and AG (pink) clusters are similar. The main difference is that HC-2DVD tends to associate slightly more small particles, dendrites and graupel to CR, and slightly more aggregates to AG. Interestingly, the RI (dark-green) cluster is mainly associated by HC-2DVD to aggregates, rimed particles and columns. Furthermore RI, among these three clusters, has the lowest component in terms of small particles.

The large presence of columns in the RI cluster is not in contrast with the assumption of rimed particles. In fact, solid columns and needles in particular, are produced...
in temperature ranges overlapping the ones at which rimed particles are frequently observed. Needles are produced efficiently around −5 °C, and solid columns between −5 °C and −10 °C, while rimed particles can be observed from slightly positive temperatures until −20 °C (Pruppacher and Klett, 1997; Dolan and Rutledge, 2009). Finally, the dominance of aggregates in the RI cluster in Fig. 11 can be explained by the presence of large aggregates undergoing the early stages of riming (whose shape is not yet smooth enough to be classified as rimed particles or graupel by HC-2DVD).

6.4.4 Comparison with 2DVD in terms of snowfall intensity

Additionally to the qualitative information provided by HC-2DVD, the 2DVD observations can be used in a quantitative way to investigate the relation between the content of the three clusters and the intensity of snowfall. Here we quantify the snowfall intensity by means of an equivalent flux (EF), defined as:

\[
EF = \frac{4\pi \sum_{i=1}^{N(\Delta t)} De_i^3}{3\Delta t A}
\]

(23)

where \(\Delta t\) is the temporal resolution at which EF is calculated (1/60 h here), \(A\) is the measurement area of the instrument [mm\(^2\)], \(N(\Delta t)\) is the number of particles recorded in \(\Delta t\), and \(De_i\) the equivalent diameter of a snowflake [mm], as defined in Huang et al. (2010). Given the assumptions in the estimation of \(De\), EF can be erroneous in absolute terms and therefore it is used here to compare the content of the three clusters in relative terms only. Figure 12 shows the distribution of EF for the three clusters (CR, AG, RI). It can be seen that the snowfall intensity differs among these three. CR exhibits the lowest intensities, AG intermediate ones, and RI the highest intensities. This is the expected behavior of rimed particles, characterized by larger densities and therefore larger fall velocities and EF.
Summarizing the interpretation conducted until now, the cluster CR is composed mainly of individual crystal and small aggregates, AG is associated with aggregates, while RI is considered as an indicator of riming.

### 6.4.5 CR, AG, RI: classification example

Figure 13 presents an example of ice-phase precipitation. The temperature at the location of MXPol \((y = 0, x = 0\) in the figure), was about \(-5^\circ\)C at the selected time-step. In panel a it can be observed a stratification of the precipitation into three layers. At higher altitudes (> 4 km above radar location) there is a thin layer classified mostly as crystals (CR). The crystals turn into aggregates (AG), that dominates the precipitation in a second layer (between 3.5 and 4 km above MXPol), and finally (below 3.5 km above MXPol) RI dominates the precipitation. RI is characterized by larger values of \(K_{dp}\) (up to \(2^\circ\)km\(^{-1}\)) and \(Z_H\) (up to 25 dBZ). CR instead is characterized by low values of \(Z_H\) and \(\rho_{hv}\) (as low as 0.9 [--]), and very low values of \(K_{dp}\). AG in this example exhibits polarimetric signatures somehow intermediate between CR and RI.

For illustrative purpose, the situation corresponding to Fig. 13 was simulated using the numerical weather model COSMO (see online at http://www.cosmo-model.org). The model, operationally used by MeteoSwiss, was run at about 2 km resolution with forcing from MeteoSwiss reanalysis. As shown in Fig. 14, COSMO predicts the presence of supercooled liquid water (QC) at the higher altitudes, and graupel (QG) mixed with snow (QS) below it. Both the presence of supercooled liquid water in the clouds, and the explicit presence of graupel in precipitation are in agreement with the layer of rimed particles RI classified in Fig. 13a.

### 7 Summary and conclusions

A novel approach to hydrometeor classification from polarimeteric weather radar was presented in this manuscript. The method is applied to a dataset of radar data collected...
by an X-band polarimetric radar. The approach is not based on numerical scattering simulations, and the number and type of hydrometeor classes is not defined a-priori.

The available amount of polarimetric observations is reduced to a representative subset. This subset undergoes a hierarchical clustering algorithm with spatial constraints, that groups similar observations according to both the (data-driven) similarity and the spatial smoothness of each partition. This means that we made the assumption of smooth spatial transitions between hydrometeor types. Following this strategy, an optimal number of 7 clusters is found. Three clusters are found at positive temperatures and they are interpreted as drizzle (DZ), light rain (LR) and heavy rain (HR), according to the rainfall rate associated with them (and according to the comparison with the algorithm DR2009, for verification purposes). One cluster appears systematically around 0°C temperatures and is associated with melting snow (MS). Finally, three clusters are found at negative temperatures and they are interpreted as (dominated by) crystals/small aggregates (CR), aggregates (AG) and rimed particles (RI).

The proposed approach is the first attempt, using unsupervised classification, to move the starting point of a classification algorithm away from scattering simulations conducted over an arbitrarily defined number of hydrometeor classes to the identification of relevant clusters in the data themselves.

Among the advantages of the approach, we remind that it is independent from systematic biases that may affect the polarimetric observations, and the data-driven approach ensures that the identified clusters take into account the accuracy of the instrument. Finally, the method is adaptable to other radar systems and can be tuned to include other constraints about the spatial smoothness of the partition or temporal consistency. The main limitations of the method are related to the interpretation of the content of the clusters, that might not be trivial especially if no ground reference is available for comparison. Additionally, the method is as representative as the available data-base is, and the clusters identified are a-priori valid only for the instrument employed to collect the data. We nevertheless expect the number of clusters to be very similar for similar radar systems and datasets.
It is interesting to note that the method exploits the hypothesis of spatial smoothness of the hydrometeor classes only in general terms, when the full data-base of observations is hierarchically clustered. Future work could be devoted to extend the constraints involving spatial smoothness also to newly classified images, once the clusters have been already defined and interpreted or to include the spatial information directly in the distance metric of the clustering process. Also, this clustering approach (or some steps of the approach) could be employed as a support to fuzzy-logic based classification methods, to improve or adapt the membership functions according to the clustering outputs, in specific datasets.

Appendix A: Spatial smoothness evaluation: a visual example

In this section, we provide a visual example for the interpretation of the $SS_M$ index presented in Sect. 4.3.2. Let us consider the situation of Fig. A1, that represents a simplified and hypothetical partition of three clusters in a limited physical space of $5 \times 5$ pixels.

It is evident that clusters $C_1$ and $C_2$ are the most spatially coherent, while $C_3$ is spatially discontinuous. This is reflected in their spatial scores. In fact $SS_1 = 0.68 [-]$, $SS_2 = 0.63 [-]$ and $SS_3 = 0 [-]$. Therefore, referring to Fig. 2a and step 3, the cluster $C_3$ will be forced to merge with the one which is most similar to it (data-wise), either $C_1$ or $C_2$, leading to a new partition into only 2 clusters with a better spatial score.

However, it must be underlined that: (i) this is only a very simple and spatially limited example (i.e., the edges here have an impact on the results), (ii) the actual geometry of neighbouring radar observations (defined in polar coordinates) is not the one depicted in this example, but the one shown in Fig. 3.

Appendix B: DR2009 algorithm

The algorithm denoted as DR2009 in this paper is based on the work of Dolan and Rutledge (2009), with some adaptations. In this Appendix, we provide the exact...
parametrization of the membership functions for the fuzzy-logic scheme, as well as the weights assigned to each polarimetric variable. The input variables of the algorithm are $Z_H$ [dBZ], $Z_{DR}$ [dB], $K_{dp}$ [° km$^{-1}$], $\rho_{hv}$ [–] and $\Delta z$ [m] and their weights in the fuzzy logic scheme are 0.25, 0.25, 0.25, 0.08 and 0.17, respectively. The hydrometeor classes available are aggregates (AG), crystals (CR), drizzle (DZ), high density graupel (HDG), low density graupel (LDG), rain (R), vertical ice (VI) and wet snow (WS, not present in Dolan and Rutledge, 2009). The membership function employed for all the polarimetric inputs is a membership beta function $\beta$, while for $\Delta z$ a trapezoidal one is used. $\beta$ is defined as:

$$\beta = \frac{1}{1 + \left(\frac{x-m}{a}\right)^{2b}} \quad (B1)$$

where $x$ is the considered polarimetric variable, $m$ is the midpoint, $a$ is the width and $b$ the slope. Table 4 summarizes the values of the parameters for each polarimetric variable and each hydrometeor class.

The trapezoidal membership function $T$ employed for $\Delta z$ takes instead the form of:

$$T = \begin{cases} 
0 & \text{if } x < l_1; \\
\frac{x-l_1}{l_2-l_1} & \text{if } l_1 < x \leq l_2; \\
1 & \text{if } l_2 < x \leq r_1; \\
\frac{r_2-x}{r_2-r_1} & \text{if } r_1 < x \leq r_2; \\
0 & \text{if } x > r_2 
\end{cases} \quad (B2)$$

where $l_1$, $l_2$, $r_1$ and $r_2$ define the four vertices of the trapezoid. The values for these parameters are reported in Table 5.

**Acknowledgements.** This work is a contribution to the HyMeX programme. The authors acknowledge Météo-France for supplying temperature data and the HyMeX database teams (ES-DPRI/IPSL and SEDOO/Observatoire Midi-Pyrénées) for their help in accessing the data.
authors are also thankful to all the members of EPFL-LTE that were involved during the field campaigns in the deployment and maintenance of the instruments, and later in the processing of the data. The work was supported by the Swiss National Science Foundation under grant 136827. The authors are also thankful to Daniel Wolfensberger (EPFL-LTE) for the COSMO2 reanalysis of Sect. 6.4.5.

References


Kovács, F., and Iváncsy, R.: Cluster Validity Measurement for Arbitrary Shaped Clusters, in: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Kno-
Hydrometeor classification from polarimetric radar measurements: a clustering approach

J. Grazioli et al.
Table 1. Example of commonly used distance metrics $D(x, y)$. The notation $\|x\|_p$ refers to the $p$-norm of $x$: $\|x\|_p = \left( \sum_{i=1}^{d} |x[i]|^p \right)^{1/p}$.

<table>
<thead>
<tr>
<th>$D(x, y)$</th>
<th>Expression</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minkowski</td>
<td>$|x - y|_p$</td>
<td>$p$: free parameter</td>
</tr>
<tr>
<td>Superior</td>
<td>$\max_{i=1,...,d}</td>
<td>x[i] - y[i]</td>
</tr>
<tr>
<td>Cosine</td>
<td>$\frac{x^T y}{|x|_1 |y|_1}$</td>
<td>$^T$: transpose</td>
</tr>
<tr>
<td>Correlative</td>
<td>$\sqrt{\frac{1-r(x, y)}{2}}$</td>
<td>$r$: Pearson correlation coefficient</td>
</tr>
</tbody>
</table>
Table 2. Main characteristic of the X-band polarimetric weather radar MXPol. Additional information on the instrument can be found in Scipion et al. (2013).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radar Type</td>
<td>Pulsed</td>
</tr>
<tr>
<td>Frequency</td>
<td>9.41 [GHz]</td>
</tr>
<tr>
<td>Polarization</td>
<td>H-V orthogonal</td>
</tr>
<tr>
<td>Transmission/Reception</td>
<td>Simultaneous (STAR)</td>
</tr>
<tr>
<td>3 dB beamwidth</td>
<td>1.45 [°]</td>
</tr>
<tr>
<td>Max. range</td>
<td>30–35 [km]</td>
</tr>
<tr>
<td>Range resolution</td>
<td>75 [m]</td>
</tr>
</tbody>
</table>
Table 3. Rainfall rate $R$ [mm h$^{-1}$] associated with the three clusters appearing at positive temperatures. Some relevant quantiles (Q5, Q25, Q50, Q75, Q95) of the full distribution are given here.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Q5</th>
<th>Q25</th>
<th>Q50</th>
<th>Q75</th>
<th>Q95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>0.01</td>
<td>0.04</td>
<td>0.12</td>
<td>0.25</td>
<td>2.5</td>
</tr>
<tr>
<td>Dark blue</td>
<td>0.1</td>
<td>2.15</td>
<td>4.8</td>
<td>7.5</td>
<td>15</td>
</tr>
<tr>
<td>Red</td>
<td>4.3</td>
<td>12</td>
<td>22</td>
<td>32</td>
<td>57</td>
</tr>
</tbody>
</table>
Table 4. Parametrization of the membership beta functions $\beta$ employed in the DR2009 algorithm.

<table>
<thead>
<tr>
<th>Variable</th>
<th>class</th>
<th>$a$</th>
<th>$m$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_H$</td>
<td>Aggregates (AG)</td>
<td>17.0</td>
<td>16.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Aggregates (AG)</td>
<td>0.7</td>
<td>0.7</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Aggregates (AG)</td>
<td>0.2</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Aggregates (AG)</td>
<td>0.011</td>
<td>0.989</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Crystals (CR)</td>
<td>22.0</td>
<td>−3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Crystals (CR)</td>
<td>2.6</td>
<td>3.2</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Crystals (CR)</td>
<td>0.15</td>
<td>0.15</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Crystals (CR)</td>
<td>0.015</td>
<td>0.985</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Drizzle (DZ)</td>
<td>29.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Drizzle (DZ)</td>
<td>0.5</td>
<td>0.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Drizzle (DZ)</td>
<td>0.18</td>
<td>0.18</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Drizzle (DZ)</td>
<td>0.007</td>
<td>0.992</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>High density graupel (HDG)</td>
<td>11.0</td>
<td>43.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>High density graupel (HDG)</td>
<td>2.5</td>
<td>1.2</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>High density graupel (HDG)</td>
<td>5.1</td>
<td>2.5</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>High density graupel (HDG)</td>
<td>0.018</td>
<td>0.983</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Low density graupel (LDG)</td>
<td>10.0</td>
<td>34.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Low density graupel (LDG)</td>
<td>1.0</td>
<td>0.3</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Low density graupel (LDG)</td>
<td>2.1</td>
<td>0.7</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Low density graupel (LDG)</td>
<td>0.007</td>
<td>0.993</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Rain (R)</td>
<td>17.0</td>
<td>42.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Rain (R)</td>
<td>2.8</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Rain (R)</td>
<td>12.9</td>
<td>12.6</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Rain (R)</td>
<td>0.01</td>
<td>0.99</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Vertical ice (VI)</td>
<td>28.5</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Vertical ice (VI)</td>
<td>1.3</td>
<td>−0.8</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Vertical ice (VI)</td>
<td>0.08</td>
<td>−0.1</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Vertical ice (VI)</td>
<td>0.035</td>
<td>0.965</td>
<td>1.0</td>
</tr>
<tr>
<td>$Z_H$</td>
<td>Wet snow (WS)</td>
<td>20.0</td>
<td>30.0</td>
<td>3.0</td>
</tr>
<tr>
<td>$Z_{DR}$</td>
<td>Wet snow (WS)</td>
<td>1.4</td>
<td>2.2</td>
<td>3.0</td>
</tr>
<tr>
<td>$K_{dp}$</td>
<td>Wet snow (WS)</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>$\rho_{hv}$</td>
<td>Wet snow (WS)</td>
<td>0.135</td>
<td>0.835</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 5. Parametrization of the trapezoidal membership function $T$ employed for the non polarimetric variable $\Delta z$ [m].

<table>
<thead>
<tr>
<th>Variable $\Delta z$</th>
<th>Class</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$r_1$</th>
<th>$r_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta z$</td>
<td>Aggregates (AG)</td>
<td>0</td>
<td>500</td>
<td>20 000</td>
<td>25 000</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Crystals (CR)</td>
<td>0</td>
<td>500</td>
<td>20 000</td>
<td>25 000</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Drizzle (DZ)</td>
<td>$-25 000$</td>
<td>$-20 000$</td>
<td>$-100$</td>
<td>0</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>High density graupel (HDG)</td>
<td>$-600$</td>
<td>100</td>
<td>20 000</td>
<td>25 000</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Low density graupel (LDG)</td>
<td>$-600$</td>
<td>100</td>
<td>20 000</td>
<td>25 000</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Rain (R)</td>
<td>$-25 000$</td>
<td>$-20 000$</td>
<td>$-100$</td>
<td>0</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Vertical ice (VI)</td>
<td>$-50$</td>
<td>0</td>
<td>20 000</td>
<td>25 000</td>
</tr>
<tr>
<td>$\Delta z$</td>
<td>Wet snow (WS)</td>
<td>$-1000$</td>
<td>$-700$</td>
<td>700</td>
<td>1000</td>
</tr>
</tbody>
</table>
Figure 1. Maps of the two field deployments of MXPol considered in this study. (a) Deployment in Davos (CH), (b) deployment in Montbrun (FR). The yellow lines indicates the extent of the PPI sector scans, while the white lines indicates the directions of the RHI scans. Red circles are used to mark the locations of instruments directly employed in the study, while blue squares are used for Parsivel-type disdrometers (employed only to parametrize polarimetric power laws). The source of the aerial view of (a) is http://www.geo.admin.ch, and of (b) is http://www.geoportail.gouv.fr/.
Figure 2. Flow chart of the clustering algorithm presented in Sect. 4.
Figure 3. Schematic representation of the position of the 8 nearest neighbours of the radar resolution volume labeled with the number 0.
Figure 4. Evolution of Kappa, RKR, RS and AS as a function of the number of clusters in the data-set. The vertical line at $n_c = 7$ shows the selected final number of clusters.
Figure 5. Three examples of 2-D projection of the 7 clusters found in the dataset. Different colors refer to different clusters. (a): $Z_{\text{DR}}$ vs. $\rho_{hv}$, (b): $Z_H$ vs. $K_{dp}$, (c): $\Delta z$ vs. $\rho_{hv}$. 

8509
Figure 6. Classification output of the DR2009 HC method applied to the three clusters appearing at positive temperatures. On the x-axis are indicated the categories from DR2009 and on the y-axis the proportion of cluster members belonging to a given category. All the bars of the same color sum up to one.
Figure 7. Hydrometeor classification and polarimetric observation from a PPI sector scan collected the 24 September during HyMeX SOP 2012 at 02:12 UTC with an elevation angle of 3.5°. (a) Hydrometeor classification with the clustering approach, (b) $Z_h$ [dBZ], (c) $Z_{DR}$ [dB], (d) $K_{dp}$ [$^\circ$ km$^{-1}$], (e) $\rho_{hv}$ [-]. The spatial coordinates originates at the radar location.
Figure 8. Hydrometeor classification and polarimetric observation from an RHI collected the 29 September during HyMeX SOP 2012 at 13:28 UTC. (a) Hydrometeor classification with the clustering approach, (b) $Z_H$ [dBZ], (c) $Z_{DR}$ [dB], (d) $K_{dp}$ [$^\circ$ km$^{-1}$], (e) $\rho_{hv}$ [-]. The spatial coordinates originate at the radar location (the altitude of the radar is 605 m).
Figure 9. Distribution within the three clusters found in the ice-phase of: (a) \( Z_H \) [dBZ], (b) \( Z_{DR} \) [dB], (c) \( K_{dp} \) [° km\(^{-1}\)], (d) \( \rho_{hv} \) [-], (e) \( \Delta z \) [km].
Figure 10. As in Fig. 6, for the three clusters identified at freezing temperatures.
Figure 11. As in Fig. 6, for the three clusters identified at freezing temperatures and referring to the comparison with HC-2DVD.
Figure 12. Histogram density and cumulative density of snowfall intensity, as quantified by the equivalent flux EF [mm h$^{-1}$], measured by the 2DVD and associated with the three clusters identified at freezing temperatures.
Figure 13. As in Fig. 8, for the snowfall event of the 26 March 2010, at 15:31 UTC in Davos (CH). The altitude of the radar is here 2133 m.
Figure 14. Mixing ratios of hydrometeor contents obtained with the COSMO2 numerical weather model along the RHI transect of MXPol (Same as figure 13), the 26 March 2010 at 15:29 UTC. Mixing ratios are given for cloud ice (QI), snow (QS), rain (QR), cloud water (QC), graupel (QG). The domain of figure 13 is highlighted by a black rectangle.
**Figure A1.** Schematic example to illustrate the spatial smoothness of a classification of a 5 × 5 pixels domain among three available classes. The classes are both color-coded and labeled as 1, 2 and 3.