D. Dunkerley (Referee)
david.dunkerley@monash.edu
Received and published: 29 December 2016

- This paper presents an analysis of a set of disdrometer data (from 2008-2014) processed to yield a 1 minute rainfall series. This was then grouped into 545 simple rainfall events using an inter-event time of 30 minutes (page 3 line 31). Half of the events were used for 'learning' using some grouping methods, and the other half for testing the grouping methods.

From the literature, 17 characteristics of rainfall events (such a duration, depth, or mean rainfall rate) were explored as criteria for characterizing events. Since some of the characteristics are correlated, the authors sought to identify a smaller set of more parsimonious variables that might be sufficient to characterise rainfall events.

In terms of method and approach, the paper seems to be generally sound and to employ useful methods that are well applied.

An aspect that I found to be missing was a discussion of the purpose for which rainfall events might need to be characterized. After all, it seems reasonable to think that this should guide the selection of parameters that might be meaningful for particular applications. For instance, in hydrology, it is well known that rainfall events cannot be characterized adequately through the use of simple descriptors such as average rainfall rate or peak intensity. Rather, measures of the time-distribution of rain and no-rain periods within an event are important, as is the position of the intensity peak(s) within the event – either early or late, for instance.

These characteristics in turn affect the partitioning of water when rain arrives at the soil surface (e.g., far more of the rain tends to become overland flow if the largest intensity peak occurs late in the rainfall event, and much less becomes overland flow if the intensity peak is early).

We agree on the fact that, given an application, some parameters clearly carry more information than others. This article mainly aims at validating a methodology able to select the most parsimonious subset of variables representative of the whole. The 23 parameters of this study are clearly not exhaustive. They encompass the parameters usually used by the authors themselves, to which are added those derived from a bibliographic study. They nevertheless allow to validate the methodology. In further studies it will be possible to improve the set of parameters to test their ability to better characterize the rain events. It is also possible to adapt the parameter set with a given application in mind.

Often parameters are the result of the knowledge of various given physical/environmental contexts. When dealing with a data set, the researcher often has to relate his case to the numerous situations encountered in the literature. It can be puzzling to choose the most appropriate parameters. The methodology allows to automatically bring out a reduced set of pertinent parameters. This article illustrates a way to discriminate among numerous parameters.

The SOM algorithm implies that each neuron of the map gather similar data, rain events in our case. By selecting the best map according to the topological error we make sure that the events gathered by a neuron are close to the events of its neighbors. By doing so we make sure that the map is well spread in the data space.
The map can be seen as a classification of the rain events. The five parameters we hold on to are the one best minimal subspace preserving the topology and thus the similarity in the original data space.
The underlying hypothesis is that the most valuable information is the one which best represent the similarity/dissimarity in the original data space.
This assumption is based on the fact that the parameters are redundant and sufficiently coherent to learn a meaningful map. This is validated by the validation process.

The following sentence is added p.4 L7

Of course this set of 17 features is not exhaustive and some other features could be added depending the application. For example is hydrology the positions of the intensity peaks inside the event could be a relevant feature.

The following sentence p1 L36 is modified

The first goal of this study is to select the most relevant features to characterize the events through a data-driven approach without taking into account the application context and thus to characterize a rain event in the most parsimonious and efficient possible way.

New formulation

The first goal of this study is to select the most relevant features to characterize the events through a data-driven approach without taking into account the a priori knowledges of the field of application and thus to characterize a rain event in the most parsimonious and efficient possible way.

- A useful typology of 5 classes of event emerges from the numerical analyses presented in the paper, and I thought that this formed a significant and useful contribution of the paper. The link to convective and stratiform precipitation is well explored. However, here again, I felt that material was missing that might have been included in the paper. For instance, orographic rainfall can exhibit very different characteristics from other forms of precipitation, such as convective rain. This can include very long event durations and very prolonged rain of consistent or rising intensity. The authors give the impression that they only recognize convective and stratiform rainfall as end members.
Likewise, they don't consider rainfall over the oceans, which has some temporal characteristics that distinguish it from terrestrial rainfall. As a result, I felt that the authors should offer some caveats about the extent to which they argue that their approach and conclusions might, or might not, be more widely applicable than to the single geographical location (and rainfall climate) represented by their disdrometer data. These caveats should be reflected in the title of the paper, which should be less sweeping, and perhaps refer to stratiform and convective rainfall, and/or to the particular study region (France) and its particular rainfall event characteristics. Of course, the authors also neglect seasonal and inter-annual changes in rainfall event character, and this also warrants mention and possibly some discussion. The authors seem to consider that their results are in some way definitive and universally applicable. I doubt this, and would like them to consider how their results might have changed had they used a larger data set, including data from different rainfall climates.

We agree, we were not clear enough on the purpose of this work. The conclusions drawn in this article follow from the data set used in the article which is representative of a region (Ile de France in France). There is no orographic rainfall ***events***
Given the size of the data set we could not take seasonal and inter-annual changes in rainfall event.
When parting the complete rain event set in two. To insure a correct generalization we made sure to take different time periods. When looking closely to the data, we notice that the 2 periods are noticeably different which ensure a better generalization. With a larger data set, it would be interesting to study the seasonal and the inter-annual information encompassed in the map.

Of course novel data such as orographic and oceanic rain events would probability not be as well taken into account by the map. Nevertheless, the map was thoroughly validated. By making sure that the topology was preserved we made sure that the behavior of the map would be as good as possible.

Even if it is far from exhaustive the data set gather a large variety of events. If the map was learned on a more exhaustive data set first the number of data would increase implying the possible use of a bigger map. It would also imply a richer panel of data most certainly with a wealth of behavior not present in the current data set. In this case, new parameters better characterizing these new rain events ought to be added to highlight their specific properties.

For example, concerning a very long and consistent Orographic rain event, the current five parameters stressed out in the article may be enough to characterize such an event. If we were to take into account something as specific as a rising intensity a dedicated parameter ought to be added.
Adding new data should modify the data density allowing the rise of a neuron specific of a new class of events. Stressing the preservation of the topology in the algorithm is done in this sense.

Once the methodology validated and vetted, we are going to work on larger data sets potentially exhaustive. It is a work in itself.
The following sentences were added in the conclusion : see below

Some caveats were added
p11 L36 : …. But sufficiently heterogeneous between them. Of course this classification is obtained for middle latitude climate. The data set used in this study is only representative of a particular region and topography (Ile de France in France, temperate climate). Data driven analysis cannot lead information on process that are not sampled in the data set, there are no orographic rainfall events nor oceanic observations which could present particular features and lead to additional specific clusters of events. The last step …..

P15 L18 (conclusion) : The present study was conducted with observations from middle latitude area in plain region. The relevance of this classification needs to be confirmed with other data set collected in different climatic areas and for different meteorological situations encountered for example in mountain or coastal areas. If the SOM was learned on a more exhaustive data set the number of data would increase allowing the use of a bigger map allowing to represent possible new behaviors which are not present in the current data set. This point will have to be addressed in future works

- I recommend some minor revision to address the above points.

The written English could be improved in places. For instance, ‘criteria’ is the plural form; when referring to a single parameter the word to use is ‘criterion’ (singular).
‘Dysfunction’ (page 3 line 36) should be ‘malfunction’.

We agree. A thorough revision of the manuscript will be done by a professional translator.

Anonymous Referee #2

Received and published: 5 January 2017

1 Summary

This manuscript proposes a data-driven approach to analyze rain-rate time series at the event time scale in order to link micro- and macro-physical properties of rainfall. After defining what is a rain event, a genetic algorithm is combined with a self-organizing map (SOM) method to identify the most informative descriptors of a rain-rate time series for a parsimonious approach. The obtained 5 descriptors are then used with the corresponding self-organizing map in order to “project" the initial 23-dimensional space into a 2-dimensional (map of neurons). An unsupervised clustering technique (hierarchical ascending clustering) is then applied to identify clusters in the neurons of the SOM, corresponding to clusters of rainfall events. Using 2 clusters, the usual convective/stratiform dichotomy is retrieved, and the authors proposed to use up to 5 clusters.

Taking advantage of the fact that the employed time series come from a disdrometer, the links between the rain descriptors or the neurons of the SOM and two important parameters of the drop size distribution (DSD) are investigated. In this way some relationships between rainfall macro and micro-physical properties are highlighted.

2 Recommendation

I enjoyed reading this manuscript (despite the quality of the English that must be improved) because the proposed approach is original and promising. Such characterization of rainfall events and the possible links between the macro and micro properties of rainfall are highly relevant to AMT readership and to the community in general. I have some relatively minor comments/suggestions listed below, I hence recommend to send the manuscript back to the authors for minor revisions.

3 General comments

1. The dimensionality reduction is well explained, but I did not find a quantification of the amount of information lost in the process. The obtained 5 descriptors and the corresponding SOM are optimal with respect to the criterion defined (topological error), but this optimum could be bad in absolute term (i.e., a significant amount of information is lost overall even if the selected descriptors/SOM are better than other combinations of descriptors/SOMs). I missed such discussion in Sec.3.1.

The SOM algorithm aims at giving a low quantization error. First the map, is learned with a coarse training allowing the map to spread well (with no folding) in the data space, keeping the topology of the original data space. The second step, fine tuning, insures that the neurons become local averages of the data. This step insures a good quantization of the data.

The final map is a compromise between the topological error and the number of parameters. Since the topological error is computed on the total data space, it insures that the parameters used to learn the map also preserved the topology for most of the other parameters. This insures that data close with respect to the five final parameters are also close in the full/original data space. For parameters used to characterize natural events, like rain, ruled by the law of physics this implies that the map was able to learn the relationship between the parameters (even for the parameters not used in the mapping of the data space), otherwise the topology would not be preserved.

Unlike linear approaches, like principal component analysis, that uses an orthogonal transformation, the non-linear dimensionality reduction method uses a mapping quantification of the amount of information. A quantification of the amount of information lost in the process is quite complex to evaluate due to nonlinear relationships between variables. For this reason we have provided in table 4 the R2 determination coefficient for the whole variables. As it can be seen in this table the determination coefficient are rather good even for the unlearned variables. This guarantees the good behavior of the map and consequently only few information is lost.

2. How transferable to other climatic regions are the results obtained from the presented analyses? Can interested reader use the exact same SOM in other regions or it should be recomputed to adjust to the local climatology?

Of course novel data such as orographic and oceanic rain events would probability not be as well taken into account by the map. Nevertheless, the map was thoroughly validated. By making sure that the topology was preserved we made sure that the behavior of the map would be as good as possible. This point will have to be addressed in future work

The following sentences were added :

p11 L36 : …. But sufficiently heterogeneous between them. Of course this classification is obtained for mid latitude climate. The data set used in study is only representative of a particular region and topography (Ile de

France in France, mid latitude climate). Data driven analysis can not lead information on process that are not sampled in the data set, there are no orographic rainfall **events nor oceanic observations which could present particular features and lead to additional specific clusters of events.. The last step …..**

P15 L18 (conclusion) : The present study was conducted with observations from middle latitude area in plain region. The relevance of this classification needs to be confirmed with other data set collected in different climatic areas and for different meteorological situations encountered for example in mountain or coastal areas. This point will have to be addressed in future work.

3. There are many grammar and vocabulary mistakes throughout the manuscript.
The authors must have the manuscript edited by a professional or a native speaker at least. I cannot list all of them but here are a few examples: precipitation without s, clusterS (p.2, l.34), "In a second time" (p.2, l.36), punctual should be point (p.3, l.2), "is more able for detecting" (p.5, l.10), "the variables those the components" (p.5, l.36)...

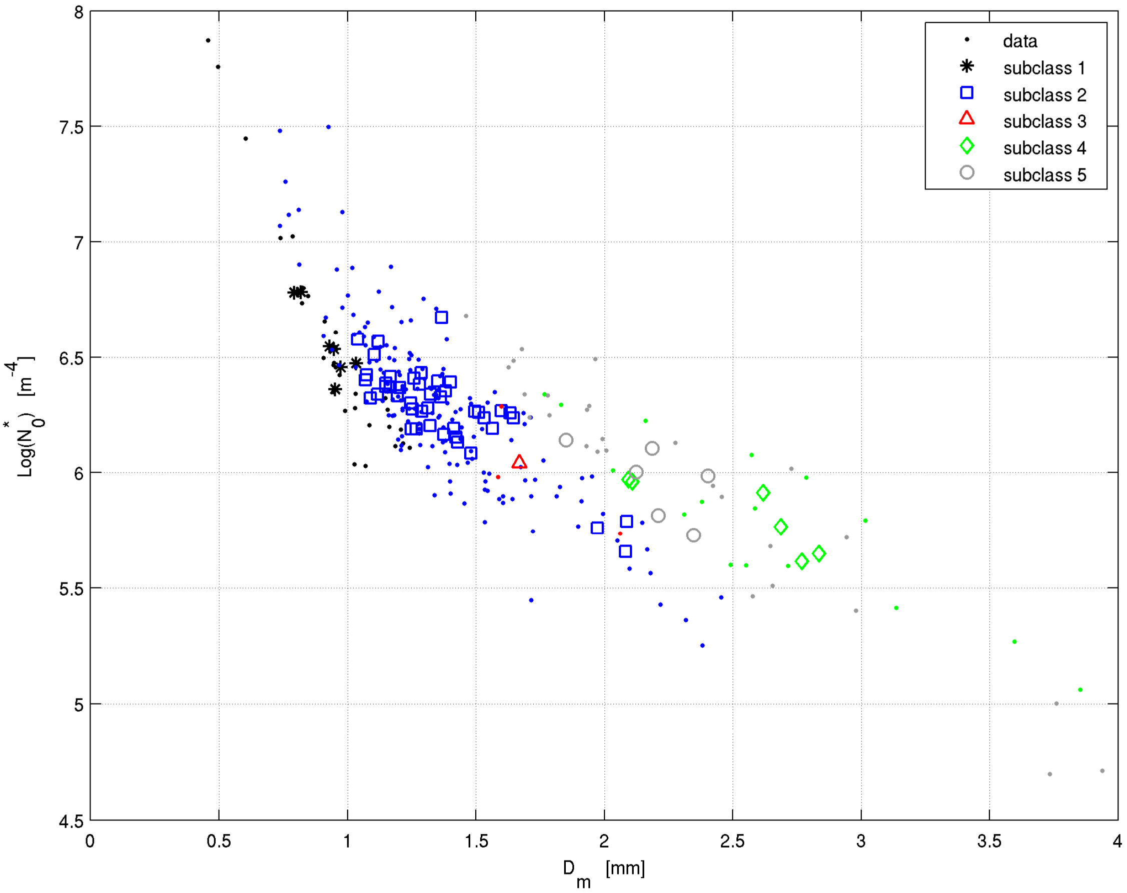
We agree. A professional translator will edit the manuscript.

4 Specific comments

1. Title: is microphysical information really derived from macrophysical information?
It is mean that some microphysical properties can be inferred thank to the knowledge of the macrophysics subclass. Hence an event belonging to subclass 4 has high Dm (2.7 mm) and low N0* (5.7)

Figure 9 shows that there is a link between a given neuron and (Nw , Dm) but we do not know how the events "attached" to a given neuron are spread in the (Nw , Dm) space.

Here is Figure 9 with added data points. The points are colored with the color of the class they belongs to. We can see that for each class the corresponding neurons are quite well spread among the data belonging to the same class.



2. P.1, l.38-39: dimensionality reduction implies more or less information loss. What can be discarded exactly may differ from one application to the other... Hence the intended application may be important.

The dimension reduction indeed implies potential information loss. Nonetheless, the large number of parameters implies redundancy. Dimensionality reduction is intended to reduce this redundancy thus part of the information loss can be considered as denoising.

P1, l.38, we replaced "to select the most relevant features" by "to select a reduced set of features"

Concerning what can be discarded or not, let's consider the unlearned IETp variable because we knew that it would not bring pertinent information on a rain event.. The resulting map is not structured with respect to the IET which is, as it should be expected, considered as noise. Thus any loss of information associated to an impertinent parameter cannot be considered in the end as a loss of information. An impertinent parameter even if learned is inconsistent with the rest of the information used to describe an event. It is thus discarded during the learning process. Hence the resulting map is not structured according to this variable.

We intended a general approach aimed toward the retrieval of microphysical information. We believe that this approach is generic and that the resulting map, due to the specificity of the algorithm, should also be useful for other applications. Either way, for a specific application, if a dedicated data set is provided, it should be used to apply the algorithm.

3. P.2, l.8: microphysics does not reduce to the DSD (which corresponds more to the microstructure of rainfall). The use of microphysics in this context is a bit ambiguous and confusing.

We agree. These sentences :
"Usually the microphysics is characterized by the raindrop size distribution, noted N(D), which is defined by the number of raindrops per unit of volume and per unit of raindrop diameter (D). Actually, information on rain microphysics is often displayed through proxies of N(D) as it will be explained further. At the present time, the rain microphysics features are not accessible by rain gauges which only provide macrophysical information."
are replaced by:
"One of the key information for remote sensing is the raindrop size distribution, noted N(D), which is defined by the number of raindrops per unit of volume and per unit of raindrop diameter (D). Actually, information on raindrop size distribution is often displayed through its proxies as it will be explained further. At the present time, these kinds of features are not accessible by rain gauges which only provide macrophysical information."

4. P.2, l.13: there are more than a few disdrometers worldwide! Please rephrase.

We agree.
"Around the world, there are few disdrometers where there are tens of thousands of rain gauges."
is replaced by
"Around the world, there are tens of thousands of rain gauges where there are a lot less disdrometers."

5. P.2, l.20: some disdrometers allow the estimation of rain rate (and other variables) at higher temporal resolution than 1 min.

"Disdrometers provide drop size distribution and consequently they allow estimating one-minute rain rates which are the measurements used to get hydrological information."

is replaced by:

"Disdrometers provide drop size distribution and consequently they allow estimating one-minute (or less) rain rates which are the measurements used in this study to get hydrological information."

6. P.3, l.10: a rain event will also strongly depend on the considered spatial and temporal resolution. You work at the point scale, but a rain event could also be defined over a given area (using model grids for instance). This should be mentioned I think.

"Indeed a rain event will depend on the sensor characteristics (specific surface caption, detection threshold, instrumental noise)."
is replaced by:
"Indeed a rain event will depend on the sensor characteristics (specific surface caption, detection threshold, instrumental noise) as well as on the spatial or temporal resolution chosen for the study."

7. P.3, l.30: how sensitive are the results to this MIT value of 30 min? As mentioned above, the spatial scale probably has an influence on the relevant MIT value.

We have run some quick test for various values of the MIT, but we have not really investigated this point and consequently we are not really in position to answer to this question. A greater MIT value will aggregate rain periods together and consequently will modify some parameters like the event duration 'De' or the Rain event Depth 'Rd'. Some others parameters can remain unchanged or not, this is the case for example for parameters Pci (which is sensitive to the variability of the rain rate). We expect that a higher MIT may aggregate events of a different type, whereas a shorter MIT tends to increase the number of events while they belong to the same group of rain cells

8. P.4, l.7: the term "descriptors" could also be used here.

Changed

9. P.4, l.27: could you provide some quantitative information about the goodness-of-fit to the normal distribution of the different transformed variables? Are they close enough to Gaussian distribution?

The transformations are empirical. Our algorithm or the PCA do not rely on the gaussianity of the data. Properly speaking, there is no statistical or probabilistic framework in this article.
(No estimator, no confidence interval or no whatsoever are involved.)
As it is written P.4, l.25 and P.4, l.27 the final distributions are "quasi-normal" or "the closest to a normal distribution". As it is stated P.4, l.26 the transformations were chosen "empirically" on the basis of the kurtosis and skewness estimation. Before using any learning algorithm, the data must be checked and transformed to be fitted for the algorithm. Here the transformations are simply used to contract/dilate the data space to guarantee a better spreading of the map among the data.
Having transformed distributions, close to normal, also helps for the interpretation of the PCA.

10. P.4, l.33: "learning data set": it is not defined... I guess it is a subset of the total data sets, but how was it obtained?

The learning and test sets are defined at the end of section 2.1 where a reference is made to the Table Tab.1..

"on the learning data set."
is replaced by
"on the learning data set (cf. end of section 2.1 and Tab.1)."

11. P.4, l.33 - p.5, l.2: is this paragraph about PCA really necessary?

The PCA is a well-established technique known from most. Many readers may wonder what could have been done with a PCA. Since it could be counterproductive the PCA is done. It is also a way to display the redundancy in the parameters. It also allows to introduce the fact that some parameters (such as IETp) are potentially non-informative.

12. P.5, l.30: vector should be denoted in bold font.
ok

13. P.5, l.32: why 60 chromosomes?
This value is not critical and any even value can be used. A smaller number of chromosomes involves more generations to converge. The GA theory shows that the algorithm converges to the same solution whatever the number of chromosomes.

14. P.5, l.35: "training data set": same as above for learning set, it is not defined...
Training data set is replaced by learning data set

15. P.5, l.36-37: "Once training each ... variables": I do not understand this sentence, it seems there is a syntax issue.

There is indeed a syntax issue.
"Once training each of the 60 Maps"
is replaced by:
"Once trained each of the 60 Maps"

16. P.6, l.6: could you provide the functional form of te(xk)?
Additional information is now provided in sections 3 and 4.
the topological error $te(x^k)$ *is computed according to eq. 2 in Uriarte and Martín (2008)*

E. Arsuaga Uriarte, and F. Díaz Martín, Topology Preservation in SOM, World Academy of Science, Engineering and Technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:9, 2008

17. P.6, l.30: "describe quite well the original space": could you provide quantitative information on this aspect? Based on what can you state this?

The discussion concerning the representativeness of the 5 selected variables is discussed in sections 4.1 & 4.2. Consequently this sentence has nothing to do here. The sentence is replaced by the following one :

At the 187th generation we get a subspace formed by 5 variables namely : Event duration D_e (#1), Standard deviation σ_R (#9), rain rate peak in event R_max (#11), Rain event depth R_d (#13), Absolute rain rate variation P_c1 (#14).

18. P.6, l.28-32: if I am correct, the 5 selected variables are transformed ones, so their physical interpretation may be slightly less straightforward than suggested in the text.

As told previously the transformations are simply used to contract/dilate the data space to guarantee a better spreading of the map among the data. It has an impact on the distances between the data. It does not affect the ordering of the data. It will not impact the interpretation of the map.

19. P.7, l.21: why 8 × 8 neurons?

64 neurons is a compromise. When we use the topological error we aim at a fine mapping of the data space. A smaller map would imply a poorer description of the data space. We have 234 rain events. 64 neurons imply more or less 3 to 4 data points to compute a neuron referent. We could not have more neurons. Choosing a square map implied not making hypotheses on the shape of the map that could have been representative only of our non-exhaustive data set.

20. P.8, l.21-27: is this valid for all climatologies or just for the one studied here (temperate mid-latitudes)?

It is not valid for all climatologies. The text has been clarified.
"These authors have noticed that successive rain and no rain periods are found to be uncorrelated, thus a rain time series can be considered by an alternation of rain event and no rain independently drawn periods. That is similar to say that inter-event time (IET) doesn't characterize the rain events. Brown et al. (1983) also investigated a possible dependence between IETp and the intra-event characteristics and they conclude that the assessment of their data gave no indication that such dependency exists."
is replaced by:
"These authors, working in temperate mid-latitudes for relatively short periods, have noticed that successive rain and no rain periods are found to be uncorrelated, thus a rain time series can be considered by an alternation of rain event and no rain independently drawn periods. That is similar to say that inter-event time (IET) doesn't characterize the rain events. For other places and other climatologies it is less clear, Brown et al. (1983) also investigated a possible dependence between IETp and the intra-event characteristics and they conclude that the assessment of their data gave no indication that such dependency exists."

21. P.10, l.1: the delineation could be illustrated in Figure 6.

Added

22. P.10, l.19: maybe you could add the coordinates (2,0.7) in the (Rm ,βL3) space to help the reader.

It has been modified.

23. P.10, l.32: "body rain disturbances"?

It has been corrected.
"the synoptic precipitations caused by the mid-latitude depressions are an example of stratiform precipitations. ***They manifest themselves in the body rainy disturbances associated with warm and cold fronts."***
is replaced by:

*“*the synoptic precipitations caused by the mid-latitude depressions are an example of stratiform precipitations **they manifest themselves in disturbances due to warm and cold fronts.*”***

24. P.12, l.11: given the definition provided in Eq.6, Dm is the mass-weighted diameter.

You are right.
“This variability is represented by two microphysics parameters namely the mean volume diameter (Dm) and the parameter N*.”
is replaced by:
“This variability is represented by two microphysics parameters namely the mass-weighted diameter (Dm) and the parameter N*.”

25. The quality of figures 2, 3 and 6 should be improved.

We agree. The resolution has been improved and the size of the font changed.

Anonymous Referee #3

Received and published: 11 January 2017

Paper revision. Title: Data driven clustering of rain events: microphysics information derived from macro scale observations. Author(s): M. D. Dilmi et al. MS No.: amt-2016-389

In this paper, the authors present a data-driven approach to analyze the rain events in order to characterize them both from a micro- and macro-physical point of view.
The authors use a disdrometer dataset of 545 events (according to their definition of event), which is divided in two uneven groups, one for learning and on for testing.
A Genetic Algorithm (GA) is combined with a self-organizing map (SOM) method to identify a number of indicators (from a list of 23 micro- and macro-variables) able to full describe each event. A previous step led to identify the independent indicators though a Principal Component Analysis (PCA). They found that the best performance is obtained by using 5 macro indicators. A hierarchical cluster analysis is then applied to the 5 indicators subset to cluster all the events in the two "commons" convective/stratiform groups and in five finer groups.

I found the manuscript interesting because the authors use a new technique (moreover derived from a different science field) and even because they aimed to link micro- and macro-physical variables to describe a rain event. Generally, the quality of figures and tables is good, but I suggest to box all the figures and to increase the quality of figure 1a. However, some questions and doubts arise reading the manuscript. I suggest the publication of the manuscript after the authors address the following points.

We partially agree. The figure 1 was re-drawn with a better resolution and boxed. We chose not to box the other figures.

- Generally, the English has to be improved. Several errors are found by reading the text and I suggest you to review the manuscript by a native speaker. Some errors are very basic and could be just writing error (i.e. "is justify" Pag.1 line 16, the lacking of "s" in some third person verbs, etc.). Please revise the expression like "one can see/consider/note. . .." and check the correct use of the singular/plural words.

We agree. A thorough revision of the manuscript will be done by a professional translator.

.

- What I mainly miss within the manuscript are two aspects: what does it happen if the Minimum Inter-event Time (MIT) is chosen different from 30 minutes? What I ask here is if the authors have been conducted a sensitivity study to show the influence of the MIT on the results. This could be very useful if the technique is applied to a different definition of event.

Our choice was made based on the literature. Even if it cannot be called a sensitivity study, we also run some test to see if the conclusion drawn in the lite nevertheless literature were coherent with our data set.
We have not really investigated this point and consequently we are not really in position to answer to this question. A greater MIT value will aggregate rain periods together and consequently will modify some parameters like the event duration De or the Rain event Depth Rd. Some others parameters can remain unchanged or not, this is the case for example for parameters Pci (which is sensitive to the variability of the rain rate). We expect that a higher MIT may aggregate events of a different type, whereas a shorter MIT tends to increase the number of events while they belong to the same group of rain cells

The second point is related to the measurement instrument. All their analysis are based on disdrometer data that are not so widespread with respect to the rain gauges. Moreover, they found that the best performance is obtained by five macrophysical indicators, which are also a rain gauge outputs. So, do the authors think that they get the same results if the rain gauges data are analyzed? It could be interesting, if they have some rain gauges data available, to apply their technique to this type of data. It is well known that rain gauges have some problems in measuring very light and heavy precipitation. How this can impact on the results?

It is one of the perspectives of this work. It would give access to more exhaustive (as well in space and in time) data sets. We are working on this aspect. It has to be noticed that the nature of the measurements is different. (For example, a tipping bucket rain gauge is measuring the time to fill a given volume. It has some important implications for stratiform rain events.)
Since the resolution is less good it will definitely impact the results. It will strongly impact some parameters such as the Rmax and the PC parameters. Since Rmax and PC$_1$ are two of the five parameters selected by the algorithm it will impact the results. This is nevertheless a study work in itself.

- In the section 3.1, the authors describe the methodology used. Even if it is quite well explained, I suggest to slightly improve so that every reader can be able to correctly reproduce it (i.e. they should better explain what is the topological error).

We agree. Section 3 has been partially re-written.

Minor comments:

- Page 2, line 13: “Around the world, there are few disdrometers. . .”.
The expressions is incorrect even if I understand the will of the authors to highlight the high ratio between the number of disdrometers.

The sentence :
“Around the world, there are few disdrometers where there are tens of thousands of rain gauges.”
is replaced by
“Around the world, there are tens of thousands of rain gauges where there are a lot less disdrometers.”

- Figure 3: I do not understand the colorbar values. Can you better explain them?
The colorbar represents the average distance between a particular neuron and its neighbors
The caption of Fig. 3 is modified:
The color of each neuron represents the average distance between the neuron and its neighbors.

Anonymous Referee #4

Received and published: 15 January 2017

Brief Summary/General Comments:

This manuscript develops an approach to analyze and interpret rain disdrometer data by using statistical and analytical tools common in other scientific communities, but relatively unknown within the hydrological and rain microphysical communities.
As a frequent user of disdrometric data, I find the analysis presented in this paper to be fascinating and potentially transformative to future work in hydrology and rain microphysics. It is clear the methods presented here are extremely novel within this community, and it appears that there is a lot of valuable information that can possibly be gained by conducting analyses in the ways described within the text.
However, as other referees have noted, the entire manuscript needs a thorough revision by native English speakers in the interest of readability. Since the analytical methods used here are expected to be new to most of the intended readership, it is vitally important that the exposition is clear, unambiguous, and complete enough so that others can follow in the footsteps of these investigators; the tone of these sections would be most useful to the reader if they were very tutorial in nature. In particular, sections 3 and 4 (which involved fundamentally new ideas for me) require more detail and clarity to allow a reader to reproduce and/or mimic the work with similar data-sets.
As it stands, I only have a vague understanding of the details of the process that was undertaken – but the results and methods are intriguing enough that I desperately wish the manuscript could be rewritten in such a way that a diligent reader could reproduce the results. The text is currently opaque enough that it is difficult for me to determine whether or not there are technical issues related to the work.

Responses to general comments

In section 3, we have added more detailed information and references concerning the variables selection. The figure 2 has been enlarged and made more readable. In addition, the following sentence will be added in the conclusion : " the Matlab code of the genetic algorithm is available on request from the authors".

The section 3 & 3.1 is replaced by this one :

Computer-assisted variable selection is important for several reasons. Indeed, the selection of a subset of variables in a high dimension space can improve the performance of the model or its statistical properties, but also provides more robust models and reduce their complexity. In practice it is generally impossible to try all possible combinations of variables and select the best ones because of the computational cost. Many approaches exist for variables selection (Guyon and Elisseeff, 2003)**.** Among then we choose to develop a model based on genetic algorithms to search for the best variables subset. Genetic algorithms (GAs) (Holland, 1975) are stochastic optimization algorithms based on the mechanics of natural selection and genetics described by Charles Darwin. In our study, a chromosome is defined as a subset of the 23 variables. A first generation composed of a population of 60 potential chromosomes is arbitrarily chosen. The performances of each chromosome (i.e. for each 60 corresponding subset of variables) is evaluated through a fitness function *f*. The fitness function is defined in such a way that the higher it is, the more the fitness function is able to represent the whole dataset (of dimension 23) with a minimal number of variables. Based on the performances of the 60 chromosomes we create a new generation of 60 chromosomes of potential solutions using classical evolutionary operators: selection, crossover and mutation. The performances of this new generation is then evaluated. This cycle is repeated until the stop criteria is asserted. The best chromosome of the last generation provides the optimal subset of variables.

3.1 Methodology

Let's define by x^k the chromosome number k : $x^k = (x_1, x_2, \ldots, x_{23})$. x^k is a binary vector in $\{0,1\}^{23}$ space such that each component has the following meaning:

$$\forall i \in \{1,\ldots,23\},\ \begin{cases} x_i = 1 & \textit{The variable number } i \textit{ in Tab. 2 column 1 is selected} \\ x_i = 0 & \textit{The variable number } i \textit{ in Tab. 2 column 1 is not selected} \end{cases} \tag{1}$$

The word "selected" in eq. 1 means that the corresponding variable will be used both in the training step described just below in "Step 2" and for performances evaluation. Otherwise if the corresponding variable is not selected it will be used only for performances evaluation.
As previously stated the fitness function allows to provide a measure of how well a minimal subset of variables can

represent the entire data space (in dimension 23). To do that the fitness function f is defined as follow :

$$f(x^k) = \frac{1}{nb(x^k) \cdot te(x^k)} \qquad (2)$$

With

x^k : *chromosomes number k*

$nb(x^k)$: *number of selected variables in chromosome* x^k

$te(x^k)$: *topological error associated to chromosome* x^k

The aim is to minimise the number of selected variables *nb* and the topological error *te*. Consequently we seek to maximize the fitness function. The estimation of the topological error made from a Self-Organizing Maps (SOM) is somewhat complicated and requires some explanation. Self-Organizing Maps introduced by Teuvo Kohonen (Kohonen, 1982, 2001) is a popular clustering and visualization algorithm. SOM is a neural network algorithm that is based on unsupervised learning based on competitive learning (Kohonen, 1982, 2001; Vesanto and Alhoniemi, 2000). It may be considered has a nonlinear generalization that has many advantages over the conventinal features extraction methods such as Empirical Orthogonal Functions (EOF) or Principal Component analysis PCA (e.g., Liu et al., 2006). SOM applications are becoming increasingly useful in geosciences (e.g., Liu and Weisberg, 2011). As was written by Uriarte and Martín (2008) : “The SOM provides a non-linear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array. The main characteristic of the projection provided by the algorithm is the preservation of neighbourhood relations; as far as possible, nearby data vectors in the input space are mapped onto neighbouring locations in the output space”. This property allows to compute easily a topological error (see Uriarte and Martín, 2008, eq. 2). For each of the x^k chromosomes a Self-Organizing Maps *M*(x^k) is learned on the training data set. Only the selected variables are used during the learning process. Finally for each Maps *M*(x^k) the topological error $te(x^k)$ is then computed according to eq. 2 in Uriarte and Martín (2008). Section 4 provides additional information concerning Self-Organising Maps.

The Genetic Algorithm is based on the following five steps (Fig.2.):

Step 1- Initialization: (initial population)
Generate randomly, a population {x^k , k=$1, \dots, 60$} of 60 chromosomes of dimension 23.
Step 2- Evaluation: For each of the x^k chromosomes a SOM *M*(x^k) is learned. Only the selected variables are used for learning. Once training each Maps provides the topologic error *te*(x^k) allowing to calculate their fitness score *f*(x^k) by the mean of the fitness function.
Step 3- Select the best chromosome x^{Best} among the 60 chromosomes according to its fitness score previously computed on the test dataset. If x^{Best} remains the same during 50 generations then stop the procedure and select the relevant variables, i.e. the ones for which the corresponding components are equal to 1 in x^{Best}. Else go to step 4.
Step 4- Selection: Create a new population of 60 chromosomes from the current population by random sampling with replacement of chromosomes based on their probabilities calculated according to the formula:

$$\Pr(x^k) = \frac{f(x^k)}{\sum_{i=1}^{60} f(x^i)} \qquad (3)$$

Step 5- Reproduction: Mutation and Crossover possibilities in the new population.
Mutation: It consists of modifying (or not) some components of the chromosomes. The probability of mutation is in general very low and is commonly set to $p = 10^{-7}$. In our case the number of necessaries generations to reach our objective is lower than a few hundred. Consequently in our case, the probability of a mutation is highly unlikely.
Crossover: First, we randomly draw $\frac{60}{2} = 30$ couples of chromosomes from our population. Then, for each couple (x^k, x^l) (called parents) one crossover point, noted I_c, is randomly drawn in the range [1, 23] using a discrete uniform law. Two new chromosomes ($x^{k'}$, $x^{l'}$) are created in the following way:

$$\begin{cases} x^{k'} = (x_1^k, x_2^k, \dots, x_{Ic}^k, x_{I_c+1}^l, x_{I_c+2}^l, \dots, x_{23}^l) \\ x^{l'} = (x_1^l, x_2^l, \dots, x_{I_c}^l, x_{I_c+1}^k, x_{I_c+2}^k, \dots, x_{23}^k) \end{cases} \qquad (4)$$

So from two parents we generate two children allowing having a new generation with the same number of chromosomes. Finally go to step 2.

Figure 2: Diagram for the selection of variables based on a Genetic Algorithm associated with Kohonen Maps and a fitness function

In Section 4 : Some additional information are provided. The following sentences replace lines 11 to 27 p7

A SOM is a topological map composed of neurons. In our case, a neuron is a vector of dimension 23 containing the 23 variables defined previously in Tab. 2. Each neuron has 6 neighboring neurons. SOM is an unsupervised neural network trained by a competitive learning strategy that performs two tasks: vector quantization and vector projection. Different from K-means, SOM uses the neighborhood interaction set to learn the topological structure hidden in the data. In addition to the best matching referent vector (neuron), its neighbors on the map are updated, resulting in regions where neurons in the same neighborhood are very similar. It can be considered as an algorithm that maps a high dimensional data space to a two dimensional space called a map. A map can be used at the same time both to reduce the amount data by clustering and for projecting the data in a non-linearly way to a regular grid (the map grid).

In this study we used the toolbox developed by "the SOM Toolbox Team" available at the following address: http://www.cis.hut.fi/somtoolbox/. A SOM with 8×8 = 64 neurons is considered here. This choice corresponds to a compromise. A smaller map would not be able to distinguish fine details, whereas a large map would not make sense given the number of observations available.

After training by the GA algorithm described the previous section the obtained map $M(x^{Best})$ can be used to affect to any event the best matching referent vector (neuron) according to the 5 selected variables associated to the chromosomex^{Best}. Hence the obtained $M(x^{Best})$ map can be considered as an optimal representation of the initial data set.

Figure 3 shows the distances matrix. For each neuron the color indicates the mean distance between a neuron and its neighbors. The value at the center of each neuron represents the number of rain events of the training data set captured by the corresponding neuron. All neurons capture rain events and a bit more than half of them capture between 3 and 5 rain events which is close to the value that would be obtained ($234 / 64 \cong 4$) if the rain events were uniformly distributed on the map.

The following references were added :

Guyon, I. and André Elisseeff; An Introduction to Variable and Feature Selection (Kernel Machines Section)3(Mar):1157--1182, 2003.

E. Arsuaga Uriarte, and F. Díaz Martín, Topology Preservation in SOM, World Academy of Science, Engineering and Technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:9, 2008

Kohonen, T. (2001). Self-Organizing Maps. Springer-Verlag, ISBN 3-540-67921-9, New York, Berlin, Heidelberg

Vesanto, J. & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, Vol. 11, 586–600, ISSN 1045-9227

Liu, Y.; Weisberg, R. H. & Mooers, C. N. K. (2006). Performance evaluation of the self- organizing map for feature extraction, Journal of Geophysical Research, Vol. 111, C05018, doi:10.1029/2005JC003117, ISSN 0148-0227

Liu, Y.,and R.H. Weisberg (2011) A review of self-organizing map applications in meteorology and oceanography. In: Self-Organizing Maps-Applications and Novel Algorithm Design, 253-272.

Specific Comments:

In addition to the presentation concerns outlined above (and by the other reviewers), I do have some basic scientific questions to add to the discussion.

1. As noted by other reviewers, I am curious as to how much the choice of a 30-minute MIT interval used to define a rain event effects the results. Similarly, how much does the minimum detection threshold influence the results?
Our choice was made based on the literature. Even if it cannot be called a sensitivity study, we also run some test to see if the conclusion drawn in the lite nevertheless literature were coherent with our data set.
We have not really investigated this point and consequently we are not really in position to answer to this question. A greater MIT value will aggregate rain periods together and consequently will modify some parameters like the event duration De or the Rain event Depth Rd. Some others parameters can remain unchanged or not, this is the case for example for parameters Pci (which is sensitive to the variability of the rain rate). We expect that a higher MIT may aggregate events of a different type, whereas a shorter MIT tends to increase the number of events while they belong to the same group of rain cells.

The chosen threshold (0.1 mm/h) allows to detect very light rain while avoiding false detections due to dust or insects. This threshold can play an important role concerning the rain support but has very little influence on whole of the others paramaters used in this study

2. Given that this method is new to some readers, could it briefly be described what happens with this method if it is

employed with non-transformed data? The normalization method used to make each of the variables pseudo-normal sounds quite practical, but I always wonder what biases this can introduce in fundamental non-linear processes like rainfall. What would happen if such a transformation is not utilized?

We used 7 transformations. Among the seven (see Table 3) 6 transformations are nonlinear. There are simply used to contract/dilate the data space to guarantee a better spreading of the map among the data. Without any transformation the resulting map would be less representative of the data.

Yes there are quite practical, properly speaking, there is no statistical or probabilistic framework in this article (No estimator, no confidence interval or no whatsoever are involved.)

The nonlinear transformations are intended to avoid any biais.

3. The instrument you utilized had a 1-minute integration time. There are other detectors with sub-second integration times, and one can always coarsen data. What influence does integration time have on the method proposed? Does one find the same basic results (for the entire process if the full analysis is done with a different integration time-scale? This is not merely of academic interest here, since scale-matching is of vital importance within the hydrological community where instruments with very coarse resolution (e.g. radar) are often "ground-truthed" with point-detector measurements that can have sub-millisecond resolutions.

From a practical point of view we could simply aggregate the data if they are at a resolution finer than 1 minute. Otherwise :

Almost the 23 parameters are derived from rain rate parameter. Hence the vectors representing an events will be different from the ones at a 1-minute resolution. We expect to find the same basic results conditioned on the loss of information due to a coarser resolution. For resolution closed to 1 minute the expected results should be close.

It is one of the perspectives of this work. It would give access to more exhaustive (as well in space and in time) data sets. We are working on this aspect. It has to be noticed that the nature of the measurements is different. (For example, a tipping bucket rain gauge is measuring the time to fill a given volume. It has some important implications for stratiform rain events.)

Since the resolution is less good it will definitely impact the results. It will strongly impact some parameters such as the Rmax and the PC parameters. Since Rmax and PC_1 are two of the five parameters selected by the algorithm it will impact the results. This is nevertheless a study work in itself.

4. As noted by other reviewers, the conclusions seem a bit of an over-reach for a study done using one type of instrument, with one type of MIT choice, one minimum detection threshhold, at one location, with one integration time. I believe that if this approach was applied by a number of investigators at different locations using data from a number of different instruments and a variety of parameter choices it is possible that a very powerful result could be attained. I would love to work on such a study, but – due to the quality of written English in this paper – I can't follow the method closely enough to do so at this time.

I would like to petition the authors to give this paper a substantial re-write and have native English speakers edit it. In the interest of readability, I would encourage figure captions to be more descriptive. Finally, I would love to have access to either code or other materials necessary to replicate the work (so, as other reviewers have pointed out, unspecified details like the choices made in te(x^k) are known to a practitioner).

If these changes could be implemented, I would promise the authors that not only will this paper be read, but some of us will use the results in their own work – myself included.

We agree. A professional translator will edit the manuscript. Additional information is now provided in sections 3 and 4. References are added to help to replicate the work.

A. Parodi (Referee)

antonio.parodi@cimafoundation.org
Received and published: 15 January 2017

The authors present an interesting study on data driven clustering of rain events based on the analysis of characteristics related to rain rates or rain accumulation (macro physical information) and to the raindrop size distribution (microphysical information). The research results are interesting, relevant and timely.
Some questions, and possibly suggestions for future work, are however presented

hereafter:

1) the authors adopt a Minimum Inter-event Time (MIT) of 30 minutes. Did they perform any sensitivity of the presented results versus lower (15 minutes) and higher (60 and 120 minutes) threshold?

Indeed, the study of the sensitivity to MIT has to be done, but regarding the future work our priority is to extend the study to measurements carried out by rain gauge with a time resolution of 5 mn (more common and accessible)
We have not really investigated this point and consequently we are not really in position to answer to this question. A greater MIT value will aggregate rain periods together and consequently will modify some parameters like the event duration De or the Rain event Depth Rd. Some others parameters can remain unchanged or not, this is the case for example for parameters Pci (which is sensitive to the variability of the rain rate). We expect that a higher MIT may aggregate events of a different type, whereas a shorter MIT tends to increase the number of events while they belong to the same group of rain cells

2) a total of 23 indicators for macro physical description of rain events were defined, however the references provided in order to support their choices are rather limited. Please improve.

Line 161 section 2.1 added references:

Haile 2011 Brown 1985 Dris 1989

Added reference :

Driscoll, E. D., Palhegyi, G. E., Strecker, E. W., & Shelley, P. E. (1989). Analysis of storm events characteristics for selected rainfall gauges throughout the United States. *US Environmental Protection Agency, Washington, DC.*

Most parameters (mean, maximum, standard deviation,...) are common statistical descriptors for witch there is no particular references.

3) also the details provided about the 23 indicators computation are rather poor. Please improve

Among the 23 indicators described some of them are very classical like the event duration or Rain event depth. Some are less classical they are described in the provided references. We voluntarily chose not to detail certain indicators in order to shorten the paper.

4) by checking the position of the test site, namely “Site Instrumental de Recherche par Télédétection Atmosphérique” (SIRTA1) in Palaiseau (France), i noticed that it is not far from the Trappes sounding site. Then i was wondering if the authors plan to explore possible relationship between the identified characteristics related to rain rates or rain accumulation (macro physical information) and to the raindrop size distribution (microphysical information) versus the vertical thermodynamical structure observed for the identified events in the period 2008-2014. For example

Molini, L., Parodi, A., Rebora, N., Craig, G. C. (2011). Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. Quarterly Journal of the Royal Meteorological Society, 137(654), 148-154.

For about 81 events, a time-scale for convective adjustment was computed, based on gridded hourly precipitation rates derived from rain-gauge data and ECMWF analysis (ERA-Interim) of convective available potential energy (CAPE). Values of the convective adjustment time-scale, τ c, shorter than 6 h indicate convection that is responding rapidly to to the synoptic environment (equilibrium), while slower time-scales indicate that other, presumably local, factors dominate.
It would be interesting to see if and how a local convective adjustment time-scale (computed using Trappes data) is related to the results of this study.

added p10 l25:
The hypothesis that the two categories of precipitation events corresponding to different dynamical regimes can be identified solely on the basis of hydrometeorological variables confirms the study of Molini et al (2011). These authors have shown the agreement between the hydrometeorological classification (based on duration and extent of events from rain gauge network data) and dynamical classifications (the convective adjustment time-scale identified to distinguish between equilibrium and non-equilibrium convection from ECMWF analysis)
Added reference :
Molini, L., Parodi, A., Rebora, N., Craig, G. C. (2011). Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. Quarterly Journal of the Royal Meteorological Society, 137(654), 148-154.

5) in a recent paper

Bühl, J., Leinweber, R., Görsdorf, U., Radenz, M., Ansmann, A., Lehmann, V. (2015). Combined vertical-velocity observations with Doppler lidar, cloud radar and wind profiler. Atmospheric Measurement Techniques, 8(8), 3527-3536.
it is explored the potential of combined vertical-velocity observations with Doppler lidar, cloud radar and wind profiler. In this respect, even if the sodar located at the Charles de Gaulle Airport is relatively far away (50 km), did the authors consider as possible to explore relationship (if any) between updraft and downdraft velocity versus the microphysical analysis performed in this study?
I would be very curious to test, in a real-word situation, the results we got years ago in these papers about the relationship between raindrop diameter and updraft velocities:
Parodi, A., Emanuel, K. (2009). A theory for buoyancy and velocity scales in deep moist convection. Journal of the Atmospheric Sciences, 66(11), 3449-3463.
Parodi, A., Foufoula Georgiou, E., Emanuel, K. (2011). Signature of microphysics on spatial rainfall statistics. Journal of Geophysical Research: Atmospheres, 116(D14).

We are agree with the reviewer that the demonstration of the existence of a relationship between hydrometeorological variables and atmospherical processes can be exploited by carrying out the joint analyses of the two types of observations. We have thus explore the potential of combined measurements in Mercier et al. 2016 . We have built a 4-D-VAR data assimilation algorithm for retrieving vertical DSD profiles and vertical wind under the bright band from observations coming from a micro-rain radar (MRR) and from a co-located disdrometer, associated with a vertical advection model. This is also intended as a tool to better characterize rainfall microphysical processes. The coupling is done via a 4-D-VAR data assimilation algorithm. The dynamical model and the geometry of the problem are quite simple. They do not allow for the moment the complexity implied by all rain microphysical processes to be encompassed (evaporation, coalescence breakup and horizontal air motion are not taken into account). In the end, the model is limited to the fall of droplets under gravity, modulated by the effects of vertical winds et by the evaporation. Because of the limitations of the model, the retrieval algorithm is currently only suitable to study stratiform, light rain events. Some improvements are needed to provide an algorithm suitable for various weather situations (tropical rain and convective events, for instance). This is an ongoing work

Mercier F., Chazottes A., Barthès L., Mallet C. 4-D-VAR assimilation of disdrometer data and radar spectral reflectivities for raindrop size distribution andvertical wind retrievals Atmospheric Measurement Techniques, European Geosciences Union, 2016, 9 (7), pp.3145-3163. <10.5194/amt-9-3145-2016> - insu-01233557

Data-driven clustering of rain events: microphysics information derived from macro scale observations

Mohamed Djallel Dilmi[1], Cécile Mallet[1], Laurent Barthes[1], Aymeric Chazottes[1]

[1]LATMOS-CNRS/ UVSQ/ UPSay, 11 boulevard d'Alembert, 78280 Guyancourt, France

Corresponding author: Laurent Barthes (laurent.barthes@latmos.ipsl.fr)

Keywords: rain events, rain microphysics, classification, Genetic algorithm, clustering, Self-Organising Maps, unsupervised learning, Hierarchical clustering

Abstract. Rain time series records are generally studied using rainfall rate or accumulation parameters, which are estimated for a fixed duration (typically 1 min, 1 hour or 1 day). In this study we use the concept of "rain events". The aim of the first part of this paper is to establish a parsimonious characterisation of rain events, using a minimal set of variables selected among those normally used for the characterization of these events. A methodology is proposed, based on the combined use of a Genetic Algorithm (GA) and Self Organising Maps (SOM). It can be advantageous to use a SOM, since it allows a high dimensional data space to be mapped onto a two-dimensional space while preserving, in an unsupervised manner, most of the information contained in the initial space topology. The 2D maps obtained in this way allow the relationships between variables to be determined and redundant variables to be removed, thus leading to a minimal subset of variables. We verify that such 2D maps make it possible to determine the characteristics of all events, on the basis of only five features (the event duration, the peak rain rate, the rain event depth, the standard deviation of the rain rate event, and the absolute rain rate variation of the order of 0.5). From this minimal subset of variables, hierarchical cluster analyses were carried out. We show that clustering into two classes allows the conventional convective and stratiform classes to be determined, whereas classification into five classes allows this convective / stratiform classification to be further refined. Finally, our study made it possible to reveal the presence of some specific relationships between these five classes and the microphysics of their associated rain events.

1. Introduction

The analysis of "precipitation events" or "rain events" can be used to obtain information concerning the characteristics of precipitation at a particular location, and for a specific application. This is a convenient way to summarize precipitation time series in the form of a small number of characteristics that make sense for particular applications.

The concept of a precipitation event is not new, and has been used for many years (Eagleson, 1970; Brown et al., 1984). A wide variety of definitions, varying according to the context of each study, has been reported in the literature (Larsen and Teves, 2015). Moreover, when a rain rate time series (generally based on rain gauge records) is broken down into individual rainfall events, a wide variety of their characteristics, such as average rainfall rate, rain event duration and rainfall event distribution (known as hydrological information), can be computed for each event. Our analysis of the literature has led to the identification of seventeen features used to characterize rainfall, which makes it quite difficult to compare different studies. The first goal of the present study is to select a reduced set of features characterizing rainfall events, through the use of a data-driven approach, without taking *a priori* knowledge of the field of application into account, thereby characterizing rainfall events in the most parsimonious and efficient manner.

The second goal is to assess, without using any *a priori* criteria, whether the rain events are still correctly clustered by the most relevant observed features. Indeed, atmospheric process specialists distinguish between stratiform and convective events, arguing that the physical processes involved in their evolution are different. The goal here is to check that a small sample of variables, derived from spot measurements to describe rain events, can allow this distinction to be made, and ultimately be used to refine it. Hydrological (hereafter referred to as “macrophysical”) information makes use of rain gauge measurements to characterize rain events. This information is defined in order to characterize the features of global events, but not to provide any information concerning the raindrop microphysics of the event. Nevertheless, in many applications such as remote sensing, knowledge of the microphysics is essential. One key parameter in remote sensing is the raindrop size distribution, noted as N(D), which is defined by the number of raindrops per unit volume and per unit raindrop diameter (D). Information related to the raindrop size distribution is often derived from its proxies, as explained in section 5. Such features are not currently accessible through rain-gauge measurements, which provide macrophysical information only. However, more expensive devices referred to as disdrometers can provide both hydrological and microphysical information. There are currently several tens of thousands of rain gauges operated throughout the world, in locations equipped with a far smaller number (if any) of disdrometers. As described later in this paper, it is possible to retrieve some microphysical information from the hydrological data. As a consequence, rain gauge data could provide valuable information in microphysics studies, through the use of a statistical approach to indirectly infer the missing microphysics information. In the following, the terms "macrophysical" or “hydrological” information are associated with characteristics related to rain rates or rain accumulation, whereas the term "microphysical" is associated with the characteristics of the raindrop size distribution.

In the present study, we use a data-driven approach to study the relationships between different rain properties. As disdrometers provide drop size distributions, they allow one-minute (or shorter) rain rates to be estimated, and in the present study these can be used to derive the hydrological information of interest, which is coherent with the data that would be provided by standard rain gauges. Through the combined interpretation of microphysical and hydrological information, we are also able to analyse the microphysical properties of the rain-event clusters provided by our algorithm. This makes it possible to retrieve (unobservable) microphysical information from rain gauge measurements.

From a single rain-rate times series, observed with a one minute time resolution, we seek to answer the following questions:

- among the large number of hydrological variables described in the literature, which are the most significant?

- does the resulting description of rain events allow different types of rain event to be discriminated?

- what (unobserved) microphysical properties of an event, or type of rain event, can be inferred from its macrophysical description?

Our paper is structured as follows. Section 2 presents the data used in our study, and lists various hydrological parameters that are commonly found in the literature. Seventeen macrophysical variables are identified, requiring appropriate normalisation. Section 3 presents our methodology, which is based on the use of a genetic algorithm (GA) implementing a self-organizing map (SOM also referred to as a topological map). This unsupervised approach is used to select a small subset of variables from the 17 identified variables, allowing a parsimonious characterisation of rainfall events to be applied. An exploratory statistical analysis of rainfall events is provided. In section 4, the rainfall events are grouped in clusters, and are divided into two classes. It is then shown that this grouping of the dataset corresponds to the standard convective / stratiform classification. We then propose a five-subclass classification, which corresponds to a refinement of the two initial groups. In section 5 we include some additional microphysical features of rainfall events, allowing the microphysical properties of the five previously defined event classes to be studied. Our conclusions are presented in section 6.

2. The disdrometer datasets - data processing methodology

This research relies on the analysis of raindrop measurements obtained with a Dual-Beam Spectropluviometer (DBS) disdrometer, first described by Delahaye et al. (2006). This instrument allows the arrival time, diameter and fall velocity of incoming drops to be recorded. As the capture area of the sensor is 100 cm^2 its observations can be considered, in spatial terms, to be "point-like". In the present study the integration time T_{int} was set to one minute, and the raindrop measurements were used to estimate the corresponding one-minute rain-rate time series $RR_t(t)$. In order to eliminate false raindrop detections that could be generated by dust or insects, a threshold T_0=0.1 $mm.h^{-1}$ was applied. Rain rates lower than T_0 are thus set to zero. This conventional threshold is also chosen to ensure coherency with previous studies (Verrier et al., 2013; Llasat et al., 2001). In the present study, we worked with two datasets recorded during the period between July 2008 and July 2014, at the "Site Instrumental de Recherche par Télédétection Atmosphérique" (SIRTA[1]) in Palaiseau, France.

2.1 Rain event definition

In everyday life, it is common knowledge that rain starts at a certain moment, and stops some time later. However, due to its discreet nature, rain (which generally consists in a very large number of raindrops) is not an easy concept to define. Indeed, the exact definition of a rain event will depend on the sensor's characteristics (specific surface capture, detection threshold, instrumental noise), as well as the spatial or temporal resolution chosen for the study. This definition may also depend on the purpose of the study, and thus on the scientific community behind it. There is thus a wide range of criteria used to break down precipitation records into rain events. For this reason, it is important to define and apply an unambiguous definition of a "rain event".

In this study, the pattern produced by the one-minute rain rate time series RR_t(t) can be simplified by grouping non-null rain rates into a set of separate "primitive events" (Brown et al., 1985). On the basis of an assigned Minimum Inter-event Time (MIT) (Coutinho et al., 2014) each rain rate value, corresponding to a specific one-minute period of observation, is assigned to a given rainfall event, i.e. either the rainfall event in progress, or a subsequent event that is considered to be independent and "new". The MIT could also be defined as the duration of a dry period D_{dry} following which the next occurrence of non-null rainfall marks the beginning of a new event. For dry periods shorter than the MIT, rain rates from either side of this period are considered to belong to the same "composite event". Various authors have proposed different values of MIT that ensure event independence. Llasat (2001) noted that: "*The definition of an episode is quite subjective. In this case it was felt possible to distinguish between two different episodes, when the time which elapses between them without rainfall exceeds 1h, which ensures that, the two episodes come from different 'clouds'*". Bocquillon and Moussa (2014) wrote: "*the constant rain observations on less than thirty minutes represent only 5% of all the rainy periods. The representative threshold of the discretization of the data is 30 minutes to an hour.*"

Dunkerley (2008 a, b) carried out an analysis of the Inter-Event Time (IET) in order to check the influence of this variable on the definition of rainfall events, and its influence on the average rainfall rate. As emphasized in this study, when determining a value for the MIT, it is crucial to find an appropriate compromise between the independence of rain events and the intra-event variability of rain rates. The choice of MIT thus has a direct impact on the macrophysical characteristics that are ultimately determined by the analysis. Other researchers have proposed to use MIT values of 20 minutes, 1hour or even 1day (see Dunkerley, 2008a for a detailed list). In the present study it was decided to set the MIT to 30 minutes. This is in agreement with the value used by Coutinho et al. (2014), Haile et al. (2011), Dunkerley (2008a, b), Balme et al. (2006) and Cosgrove and Garstang (1995).

[1] http://sirta-dev.ipsl.jussieu.fr/joomla/index.php/85-article-sans-categorie/71-sirta-home-page

When applied to our dataset, this choice leads to the identification of 545 rain events, which can be divided up into two subsets, i.e. one for learning and the other for testing (Tab. 1). The learning dataset is composed of observations collected over a two-year period between 2013 and 2014, with an availability of 96.4%, whereas the test dataset collected during the 2008 – 2012 period contains periods with missing data, due to a malfunction of the recording device.

Table 1: Observation periods and availability of DBS observations, and numbers of rain events for the learning and test datasets

2.2 Macrophysical description of rain events

Rain events contain a wealth of information, which generally needs to be condensed into a limited set of well-chosen features. However, there is no conventional or commonly accepted list, or specific set of macrophysical features that can be used to accurately describe and summarize an event. In the present study it was thus decided to consider a large number of features, allowing the macrophysical rain event information described in the literature to be correctly represented. Seventeen characteristics were selected and identified (Llasat, 2001; Moussa, 1991), and are listed in Tab. 2. Some of these are parameter-dependent, such as P_c which uses 3 values of the parameter c. These 3 values lead to 3 Pc indices, namely Pc_1, Pc_2, Pc_3. Finally, a total of 23 descriptors were defined and numbered from 1 to 23 (column 1 in Tab. 2).

Table 2: The 23 variables identified in the literature, used for the characterization of rain events

Among the 23 indicators (hereafter referred to as variables) corresponding to the previously defined features, some are very well known. These include the event duration (D_e), the quartile (Q_i), the mean event rain rate (R_m) and the standard rain rate deviation (σ_R). Other less traditional parameters such as the parameter β_L (indicator for the convective nature of the rain, see Llasat (2001)), the absolute rain rate variation of order c (P_c), or the absolute rain rate variation ($P_{s,c}$). Some variables that are usually used to describe time series, such as the fractal dimension, multi-fractal parameters, trend, seasonality, and autocorrelation, require a long series of data and are not well suited to an event-by-event analysis. This set of 17 features is not exhaustive, and some other features could also be included, depending on the application. One example is the case of hydrology, for which the positions of the intensity peaks inside the event could be a relevant feature. Although, for events comprising a very small number of samples (very low value of variable D_e), the computation of some indicators (σ_R, Q_i) is questionable, in the present study the 23 variables were computed for each of the 545 rain events.

2.3 PCA analysis and normalization step

It is important to note that very few of these 23 variables are compatible with the probabilistic assumptions generally associated with exploratory statistical methods. They are often highly variable, with highly skewed distributions, and therefore do not have normal distributions. It is thus more difficult to make direct use of standard statistical methods with this data, as it may lead to misleading interpretations (Daumas, 1982). It is thus necessary to introduce an additional step in order to transform the original distributions into *quasi* normally distributed distributions. The most suitable type of normalizing transformation for each of these variables was selected empirically, by testing 7 different possible transformations (Tab. 3). For each variable, the retained transformation is that leading to a distribution having the strongest similarity to a normal distribution, i.e. with a kurtosis close to 3, and a skewness close to 0. For each indicator, the selected transformation is provided in the last column of Tab.2.

Table 3: Transformations used to normalize the variables listed in Table 2.

Following the normalisation step, Principal Component Analysis (PCA) was carried out on the learning data set (cf. end of section 2.1 and Tab.1). It follows that the two principal axes contain 73% of the total information, whereas the first 5 principal axes are needed to represent 90% of the total information. The IET$_p$ variable (#7) is very well correlated with axis 5, whereas the other variables are not. This means that there is no linear relationship between IET$_p$ and the other variables. For this reason, this variable was not considered as a possible candidate, in the variable selection process, during the remainder of the study. The results obtained in Section 4.2 confirm that there is no relationship between this variable and the other 22 variables. The correlation circle on axes 1 & 2 (Fig. 1.a.) shows that among the 23 variables, 16 are well correlated with the axis (close to a unit circle) and are distributed in approximately 5 groups (hereafter referred to as PCA groups). A first PCA group (G$_1$) can be identified by the variables, which are grouped close to the first axis, and are well correlated with it. As an example, this is the case for the variables σ_R (#9), P_{C_N} (#17 – 18) and β (#21 to 23). A second PCA group (G$_2$) comprises the variables R_{max} (#11) and P_{C3} (#16), just above axis 1. The third PCA group (G$_3$) is formed by the variable P_{C2} (#15) only. The fourth PCA group (G$_4$) comprises the variables P_{c1} (#14) and Ps,c (#20) and is well correlated with axis 2. The last PCA group (G$_5$) is formed by the variable D$_e$ (#1). The correlation circle on axis 1 & 3 (fig. 1.b.) shows that the variables Q$_1$ (#4), Q$_2$ (#5) and M$_0$ (#10) are quite well represented by these two axes. A similar remark can be made for variables D$_d$ (#3) and β$_{L1}$ (#21) on axes 1 & 4 (not shown).

Finally, PCA analysis clearly shows that, within each PCA group, many variables are highly inter-correlated, i.e. linearly dependant on each other. This means that several variables could be removed with no substantial loss of information. This leads to the following question: which variables can be removed in order to retain the most parsimonious subset of variables representative of the full dataset? The PCA extracts summary variables, which are a linear combination of original variables, but do not allow for the selection of variables. To answer to this question, we propose a method for the global selection of variables, which seeks to identify the relevant variables in a dataset. As it appears to be intuitively more advantageous to select variables with a physical sense, rather than using dimension reduction methods (e.g. PCA which is more suitable for the detection of linear relationships), the proposed method is based on the use of a genetic algorithm. The following section provides a brief introduction into the concept of genetic algorithms, and shows how they can be advantageously used for the selection of variables in the context of the present study.

Figure 1: PCA on the learning data set based on the 23 variables described in Tab. 3. Left: correlation circle on axes 1 & 2. Right: correlation circle on axes 1 & 3. All of the variables are normalised according to the last column of Table 2.

3. Variable selection using a genetic algorithm

Computer-assisted variable selection is important for several reasons. Indeed, the selection of a subset of variables in a high-dimensional space can improve the performance of the model or its statistical properties, but also provides more robust models and reduces their complexity. In practice, it is not generally possible to try all potential combinations of variables, and to select the best of these, as a consequence of the enormous computational cost associated with such an approach. Among the many different variable-selection techniques described in the literature (Guyon and Elisseeff, 2003), we chose to develop a model based on the use of genetic algorithms, to search for an optimal subset of variables. Genetic algorithms (GAs) (Holland, 1975) are stochastic optimization algorithms based on the mechanics of natural selection, and the genetics described by Charles Darwin. In our study, a chromosome is defined as a subset made up from our 23 variables. A first generation composed of a population of 60 potential chromosomes is arbitrarily chosen. The performance of each chromosome (i.e. for each corresponding subset of 60 variables) is evaluated through a fitness function f. This fitness function is defined in such a way that the higher its value, the greater the fitness function's ability to represent the full dataset (of dimension 23), using the smallest possible number of variables. On the basis of the performance of these 60 chromosomes, we create a new generation of 60 potential-solution chromosomes, using classical evolutionary operators: selection, crossover and mutation. The performance of this new generation is then evaluated. This cycle is repeated until a predefined stop criterion is satisfied. The best chromosome from the current generation then provides the optimal subset of variables.

3.1 Methodology

We define by x^k the chromosome number k: $\boldsymbol{x^k} = (x_1, x_2, \ldots, x_{23})$. $\boldsymbol{x^k}$ is a binary vector in $\{0,1\}^{23}$ space such that each component has the following meaning:

$$\forall i \in \{1, \ldots, 23\}, \begin{cases} x_i = 1 & \textit{The variable number i in Tab. 2 column 1 is selected} \\ x_i = 0 & \textit{The variable number i in Tab. 2 column 1 is not selected} \end{cases} \quad (1)$$

The word "selected" in eq. 1 means that the corresponding variable will be used, both in the learning step described in "Step 2" below, and for performance evaluation. Otherwise, if the corresponding variable is not selected it will be used only for performance evaluation.

As previously stated, the fitness function allows a measure to be provided of how well a minimal subset of variables can represent the entire data space (in dimension 23). The fitness function *f* is thus defined as follows:

$$f(\boldsymbol{x^k}) = \frac{1}{nb(\boldsymbol{x^k})\ te(\boldsymbol{x^k})} \quad (2)$$

where

$\boldsymbol{x^k}$: *chromosomes number k*

$nb(\boldsymbol{x^k})$: *number of selected variables in chromosome* x^k
$te(\boldsymbol{x^k})$: *topological error associated to chromosome* x^k

As the aim of this approach is to minimise the number of selected variables *nb* and the topological error *te*, we seek to maximize the fitness function. The estimation of the topological error made from a Self-Organizing Map (SOM) is somewhat complicated, and requires some explanation. The notion of a Self-Organizing Map, introduced by Teuvo Kohonen (Kohonen, 1982, 2001), makes use of a popular clustering and visualization algorithm. SOM is a neural network algorithm based on unsupervised learning, derived from the technique of competitive learning (Kohonen, 1982, 2001; Vesanto and Alhoniemi, 2000). It may be considered as a nonlinear generalization, which has many advantages over the conventional feature extraction techniques such as Empirical Orthogonal Functions (EOF), or Principal Component analysis PCA (e.g., Liu et al., 2006). SOM applications are becoming increasingly useful in geosciences (e.g., Liu and Weisberg, 2011). As stated by Uriarte and Martín (2008): "The SOM provides a nonlinear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array. The main characteristic of the projection provided by this algorithm is the preservation of neighborhood relationships; as far as possible, nearby data vectors in the input space are mapped onto neighboring locations in the output space". This property makes it straightforward to compute a topological error (see Uriarte and Martín, 2008, eq. 2). For each of the $\boldsymbol{x^k}$ chromosomes, a Self-Organizing Map *M*($\boldsymbol{x^k}$) is learned on the learning dataset. Only the selected variables are used during the learning process. Finally, for each Map *M*($\boldsymbol{x^k}$), the topological error $te(\boldsymbol{x^k})$ can be computed in accordance with eq. 2 in Uriarte and Martín (2008). Section 4 provides additional information concerning Self-Organising Maps.

The Genetic Algorithm is based on the following five steps (Fig.2.):

Step 1- Initialization: (initial population)

Randomly generate a population $\{\boldsymbol{x^k}, \text{k}=1, \dots, 60\}$ of 60 chromosomes of dimension 23.

Step 2- Evaluation: For each of the x^k chromosomes, a SOM *M*($\boldsymbol{x^k}$) is learned. Only the selected variables are used for learning. Once the learning has been completed, the test dataset and the 23 variables are used on each of the 60 maps to estimate their topological error *te*($\boldsymbol{x^k}$) allowing their fitness score *f*($\boldsymbol{x^k}$) to be computed.

Step 3- Select the best chromosome $\boldsymbol{x^{Best}}$ from the full set of 60 chromosomes according to the fitness score previously computed with the test dataset. If $\boldsymbol{x^{Best}}$ remains unchanged over a period of 50 generations, stop the procedure and select the most relevant variables, i.e. those for which the corresponding components are equal to 1 in $\boldsymbol{x^{Best}}$. Otherwise, go to step 4.

Step 4- Selection: Create a new population of 60 chromosomes from the current population, by randomly sampling with replacement chromosomes based on their probabilities, determined using the formula:

$$\Pr(\boldsymbol{x^k}) = \frac{f(\boldsymbol{x^k})}{\sum_{i=1}^{60} f(\boldsymbol{x^i})} \qquad (3)$$

Step 5- Reproduction: Mutation and Crossover possibilities in the new population.

Mutation: This consists in modifying (or not) certain components of the chromosomes. The probability of mutation is in general very low, and is commonly set to $p = 10^{-7}$. In the present case, the number of generations needed to reach the objective is less than a few hundred, such that the probability of a mutation is very low.

Crossover: In an initial step $\frac{60}{2} = 30$ pairs of chromosomes are randomly drawn from the population. Then, for each pair ($\boldsymbol{x^k}$, $\boldsymbol{x^l}$) (called parents) one crossover point, noted I_c, is randomly drawn over the range [1, 23], using a discrete uniform law. Two new chromosomes ($x^{k'}$, $x^{l'}$) are created as follows:

$$\begin{cases} \boldsymbol{x^{k'}} = (x_1^k, x_2^k, \dots, x_{Ic}^k, x_{I_c+1}^l, x_{I_c+2}^l, \dots, x_{23}^l) \\ \boldsymbol{x^{l'}} = (x_1^l, x_2^l, \dots, x_{I_c}^l, x_{I_c+1}^k, x_{I_c+2}^k, \dots, x_{23}^k) \end{cases} \qquad (4)$$

Thus, from two parents, two children are generated, allowing a new generation to be produced with the same number of chromosomes. Finally, the algorithm returns to step 2.

Figure 2: Diagram for the selection of variables based on a Genetic Algorithm associated with Kohonen Maps

3. 2 Parsimonious description of a rain event

The genetic algorithm is applied to our datasets in order to obtain an optimal subset of variables forming a subspace, which can (in a certain sense) provide relatively accurate information concerning the global space, whilst having the particularity of containing non-redundant information. At the 187th generation the algorithm produces a subspace comprising 5 variables, namely : Event duration D_e (#1), Standard deviation σ_R (#9), maximum rain rate during event R_{max} (#11), Rain event depth R_d (#13), and Absolute rain rate variation P_{c1} (#14).

The 3 variables (D_e, R_{max}, R_d) selected using this data-driven approach are commonly used in the study of hydrological processes (Haile and al., 2011). Moreover it should be noted that the commonly used variable R_m, which is computed simply by dividing the Rain event depth (R_d) by the duration (D_e), was not selected by the algorithm. This result could be expected, since it is correlated with the latter variables, and the algorithm provides a parsimonious description. Concerning the Absolute rain rate variation (P_{c1}), this variable was proposed by Moussa and Bocquillon (1991). It tends to provide information on the structure of the events, more specifically related to smooth events with a small number of sharp peaks. In fact, this variable promotes low variations of RR_t because P_{C1} is in a certain sense a structure function of order c_1 of the variable RR_t (see #14 column 4 in Tab. 2), with a low value for the exponent ($c_1 = 0.5$). Finally, the standard deviation variable (σ_R), which is a second-order moment, is the most commonly used indicator to describe the variability of the precipitation rate within the rain event.

4. SOM learned with the five selected variables

A SOM is a topological map composed of neurons. In the present case, a neuron is a vector of dimension 23 containing the 23 variables defined in Tab. 2. Each neuron has 6 neighboring neurons. SOM is an unsupervised neural network trained by a competitive learning strategy that performs two tasks: vector quantization and vector projection. The SOM, which is different to k-means, uses the neighborhood interaction set to learn the topological structure hidden in the data. In addition, in order to achieve optimal referent vector (neuron) matching, its neighbors on the map are updated, leading to the generation of regions in which neurons located in the same neighborhood are very similar. The SOM can thus be considered as an algorithm that maps a high-dimensional data space onto a two-dimensional space called a map. A map can be used both to reduce the amount data by means of clustering, and to project the data in a nonlinear manner onto a regular grid (the map grid).

In the present study we used the toolbox developed by “the SOM Toolbox Team”, which is available at the following site: http://www.cis.hut.fi/somtoolbox/. A SOM with 8×8 = 64 neurons is considered here. This choice corresponds to a compromise, since a smaller map would not be able to distinguish fine details whereas, in view of the number of observations, and a larger map would not be meaningful.

After learning by the GA algorithm described in the previous section, the resulting map $M(\boldsymbol{x}^{\boldsymbol{Best}})$ can be used to assign to any event the best matching reference vector (neuron), in accordance with the 5 selected variables associated with the chromosome $\boldsymbol{x}^{\boldsymbol{Best}}$. The $M(\boldsymbol{x}^{\boldsymbol{Best}})$ map obtained with this procedure can be considered as an optimal representation of the initial data set.

Fig. 3 shows the distance matrix. For each neuron, the color indicates the mean distance between a neuron and its neighbors. The value at the center of each neuron represents the number of rain events of the learning data set, captured by the corresponding neuron. All neurons capture rain events and slightly more than half of these capture between 3 and 5 rain events, which is close to the value that would be obtained ($234 / 64 \cong 4$) if the rain events were uniformly distributed over the map.

Figure 3: Distance matrix for the $M(x^{Best})$ map: The colour of each neuron represents the average distance between itself and its neighbouring neurons. The value inside each neuron indicates the number of rain events that it has captured inside the learning dataset. The black line separates the neurons into 2 classes, using the Hierarchical Ascendant Classification (see section 4.1). The arrows represent the gradients of the variables R_{max}, σ_R and D_e

4.1 Projection of the selected and unlearned variables onto the SOM

The five variables D_e, σ_R, R_{max}, R_d and P_{c1} used for learning are referred to as 'selected' variables, whereas the remaining 18 variables are referred to as 'unlearned' variables. In order to study the relationship between these variables, Fig. 4 shows the projections for each of the variables in the $M(x^{Best})$ map obtained with the aforementioned GA selection algorithm. The variables are discussed individually, by considering their structure, as well as the relationships between them. We note that the map is well structured for the majority of variables. This advantageous structuration of most of the variables confirms the ability of the selected variables to summarize all of the significant characteristics of rain events. Only a small number of characteristics are not adequately represented. It should be noted that almost all variables are structured according to the first or second diagonal. Among these, one may consider an initial subset comprising variables that are more or less structured according to the first diagonal. This is the case for the unlearned variable D_d , as well as for the selected variables P_{c1} and D_e. A second subset comprising variables that are structured in approximate accordance with the second diagonal can be identified. This is the case for the unlearned variables $R_{m,r}$, $R_{m,,}$ $P_{C_{Ni}}$ which are very similar to the selected variable σ_R. The unlearned variables Q_3, P_{C3}, $P_{S,C}$ also belong to the second subset and have a structure close to that of the selected variable R_{max}.

The map can be related to the previously implemented PCA (Fig.1). As can be seen in Fig. 4, the variables P_{C3} (#16) and R_{max} (#11), which have a similar structure, also belong to the same PCA group, namely group G_1 (see section 2.3, Fig. 1.a). It is interesting to note that the variables $P_{S,C}$ (#20) and R_{max} (#11), which also have a similar structure, do not belong to the same PCA group (groups G_4 and G_2 respectively) and are uncorrelated (they are orthogonal in Fig.1a). This remark means that the topological map reveals a relationship that cannot be detected using PCA. As the Rain event depth (R_d) depends on both the duration and the intensity of the events, the corresponding map has a top-down structure. Two distinct situations thus occur:

- Those events which contribute the greatest quantities of water (Fig. 4., brown neuron at the bottom right of R_d) are among the longest (see corresponding neuron of D_e), but do not have an extremely high peak rain rate (see corresponding neuron of R_{max}) and are quite smooth (see corresponding neuron of P_{c1} and σ_R).
- Other events which contribute large amounts of water (but less than previously) (Fig. 4. red neuron at the bottom left of R_d) have short durations (see corresponding neuron of D_e), but are violent (see corresponding neuron of R_{max}) and are less smooth (see corresponding neuron of P_{c1} and σ_R). The latter case reflects situations that are typical of convective storms.

The resulting map confirms the dependence structure of the two hydrological variables R_d and D_e studied by Gargouri and Chebchoub (2010).

The variable IET_p (Previous IET): the map is not structured, reflecting the independence of the characteristics of a rain event with respect to the drought period preceding the event. This corroborates the results of several previous studies (Lavergnat and Gole, 1998, 2006; Akrour et al., 2015) dealing with rain support simulations. When studying temperate mid-latitudes for relatively short periods, these authors noticed that successive rain and no-rain periods are uncorrelated, such that a rain time series could be considered as an independently drawn, alternating series of rain events and periods without rain. This is equivalent to an inter-event time (IET) that does not characterize the rain events. The same effects are not necessarily observed at other locations, and under different climatological conditions. Brown et al. (1983) also investigated a possible correlation between IETp and the intra-event characteristics, and concluded that their data provided no evidence of this.

Since the variable β_L: β_L (Llasat, 2001) is considered to represent a measure of the convective nature of the rain, it makes sense that the three variables β_{L1}, β_{L2}, β_{L3} are structured similarly, with the peak rain rate variable R_{max}. This relationship is clearly visible on the maps.

Several other relationships, which are not described in detail here, can be observed. These include the correlation between the Normalized Absolute rain rate variation ($P_{C_{Ni}}$) and the standard deviation of the intensity (σ_R). We conclude that the combination of the five selected variables provides a relatively accurate summary of the information needed to describe the rain events. The poor structuring of some variables is justified by the independence of these variables with respect to the properties of the rain events; this is the case for the variable Dry Percentage in event D_d or the variable IET_P.

Figure 4: Projection of the $M(x^{Best})$ map according to the 23 variables. The red-framed variables are those selected by the GA algorithm. The last two variables D_m and N_0^* are defined in section 5

4.2 Representation of rain events on SOM

In an effort to provide additional information for validation of the map, we compared each of the 23 variables with their corresponding value given by the SOM, for the learning dataset and the test dataset. For each of the 311 events of the test dataset, the best matching unit of the SOM, i.e. the neuron that is the closest to the event, is determined with respect to the five selected variables. As an example, for each event Fig. 5 shows the current value of the unlearned variable β_{L3} as a function of the corresponding value given by the best matching unit of the event. A spread can be seen, in particular in the central zone, whereas the spread is relatively small for values located near to the edges (which are more numerous). A linear regression leads to a relatively good determination coefficient (R-square) (0.96 and 0.89 respectively, for the learning and test data sets). Table 4 lists the value of R-square for the 23 variables obtained with the learning and test data sets. As expected, the coefficient of determination of the variable IET_p is very poor (*0.31/0.26*), since this variable is not related to the 5 selected variables, and as a consequence cannot be well represented by the SOM (Fig. 4). The selected variables have good determination coefficients, with both the learning and the test datasets; this confirms the quality of the learning and the generalization ability of the SOM. The quality of the learning step is confirmed by the fact that the R-square values of the selected variables obtained on the test set are close to those obtained on the learning set. The R-squares corresponding to the unlearned variables obtained on the learning data set emphasize the ability of the selected variables to provide the information contained in the unlearned variables; in the case of the test data set it denotes the ability of the SOM to derive all event characteristics from the selected variables only.

Figure 5: the variable β_{L3} versus its corresponding value, given by the best matching unit: from the learning data set (circles), and on the test dataset (stars). The solid line corresponds to the first diagonal

Table 4: Coefficient of determination obtained on the learning and test data sets. The values with a dark grey background correspond to the 5 selected variables

4.3 Hierarchical clustering of rain events

We have shown that the distance matrix (Fig. 3) confirms the successful deployment of the map. Based on the distance between neurons, it appears that neurons can be grouped to obtain a limited number of classes, each with its own characteristics. In order to group the 64 neurons into a small number of classes, a hierarchical cluster analysis was carried out (Everitt, 1974).

Only the five selected variables were used for the classification, and a Euclidian distance was selected for the hierarchical algorithm. Fig. 6 shows the resulting dendrogram, applied to the 64 neurons.

Figure 6 Dendrogram obtained from the Hierarchical Cluster Analysis of the 64 neurons in the SOM. The horizontal dashed line represents the threshold between the two classes.

Depending on the physical processes involved, experts tend to separate rain events into two different classes: stratiform and convective events. Although this classification is relatively crude, since stratiform and convective events can sometimes exist inside the same rain event, it is very commonly used. Concerning the times series, most authors use a very simple scheme to distinguish between stratiform and convective rain types. For reasons of simplicity, rain classification is sometimes defined using the instantaneous rain rate and the standard deviation estimated over consecutive samples. As an example, Bringi et al. (2003) defined stratiform rain samples when the standard deviation of the rain rate, taken over five consecutive 2-min samples, is less than 1.5 mm.h^{-1}, the convective rain samples are defined for a rain rate greater than or equal to 5 mm.h^{-1}, and the standard deviation of the rain rate over five consecutive 2-min samples is greater than 1.5 mm.h^{-1}.
Firstly, we separate the dendrogram into two classes. The first class contains 51 neurons and 79% of the observations, whereas the second class contains 13 neurons and 21% of the observations. The solid black line in Fig. 3 corresponds to the dividing line between these two classes. The first class, containing the greatest number of neurons, is in most cases characterized by relatively low rain rates. This can be seen by examining the structure of the map, according to the mean rain rate variable (R_m). Moreover, analysis of the standard deviation (small values of σ_R), absolute rain rates P_c (high values of P_{c1} , and low values *of* P_{c3}) shows that this class is more or less characterised by quiet, homogeneous events. Our analysis of event durations (D_e) shows that this class contains both short and long durations, but is dominated by the latter. These characterizations are relatively well matched to a description involving stratiform and stable precipitations, which are often the consequence of the slow, large-scale uprising of a large mass of moist air which then condenses uniformly.
The second group is characterized by a smaller number of neurons. This corresponds to the higher values of the mean rain rates (R_m) and peak rain rates (R_{max}). The variables σ_R, P_c have the opposite values with respect to those of the previous group. Most of the event durations (D_e) in this group are short, with the exception of neuron #64 (bottom right on the maps). This group fits well with the definition of convective events resulting from the rapid rise of air masses loaded with moisture, for buoyancy. This convective moist air can lead to the development of cumulus clouds up to an altitude in excess of 10 km, and to heavy rain.

Our analysis of the structure of the variables $\beta_{L1}, \beta_{L2}, \beta_{L3}$ in Fig. 4 confirms the previous interpretation of the two groups. These three variables, which are representative of convective rain, have high values for the neurons belonging to this group.

Figs. 7.a and 7.b show the neurons in the R_m, $\beta_{L3}$and P_{C2} subspace. These 3 variables were not used in the learning step. Nevertheless, the two classes are well separated, although an overlap does occur in Fig 7a due to neuron #64 (bottom right on the map, Fig. 4). We checked although it belongs to the convective class, this neuron nevertheless has some characteristics of the stratiform class.

Figure 7: Representation of the neurons in the Rm, $\beta_{L3}$and R_m, and P_{C2} subspaces. The stars represent neurons from group 1 (stratiform), and the squares correspond to neurons from group 2 (convective). Dashed lines indicate the neuron #64

The hypothesis that the two categories of precipitation events corresponding to different dynamic regimes can be identified solely on the basis of hydrometeorological variables is in agreement with the findings of Molini et al. (2011). These authors have shown that there is a strong agreement between the hydro-meteorological classification (based on the duration and extent of events from rain gauge network data), and dynamic classifications (the convective adjustment time-scale identified to distinguish between equilibrium and non-equilibrium convection derived from ECMWF analysis). We conclude that this unsupervised automatic clustering, based on the five selected variables, makes it possible to correctly implement a classification with these two well-known classes (stratiform and convective). It should be noted that, unlike other classifications described in the literature, this was established without making use of *a priori* information, since it is produced by an unsupervised process.

4.4 Classification of events into several classes

From the stratiform and convective classification described above, it is interesting to refine the two classes into a set of subclasses. The synoptic rainfall associated with mid-latitude depressions provides an example of stratiform precipitation, which forms in depressions in the vicinity of warm and cold fronts. The very light type of rainfall (drizzle) associated with stratus or stratocumulus is included in the class of stratiform precipitation. This can occur under anticyclonic conditions, or in the warm region of a depression. The associated rain depths (R_d) are minimal, and usually have no hydrological impact other than superficial wetting. In order to identify relevant subclasses, our classification was broken down into a number of unknown subclasses, such that $n > 2$.

An important step in hierarchical clustering is the selection of an optimal number of partitions (n_{opt}) in the dataset (Grazioli et al., 2015). Many indices can be used to evaluate each partition, from the point of view of data similarity only. Most of these evaluate the scattering inside each cluster, with respect to the distance between clusters, and assign relatively favourable scores to partitions with compact and well-separated clusters. Although different indices were tested, these did not provide the same number of subclasses (between 2 and 32 with the indices tested in this study). It should be noted that these did not take the physical meaning of each class into account. Finally, we chose $n_{opt} = 5$, since higher values led to classes with the same physical sense. The new classification based on the use of five subclasses is shown in Fig. 8.

Figure 8: Hierarchical Clustering of the map into five subclasses. The colours represent the subclass numbers: Subclass 1: Dark blue, Subclass 2 : blue, Subclass 3 : Green, Subclass 4 : Orange, Subclass 5 : Red

From these five subclasses, two belong to the stratiform class and the other three belong to the convective class. In the learning dataset, the first subclass represents 12% of all events, and respectively 68%, 1.2%, 6.8% and 12% for subclasses 2 to 5. The characteristics of these five subclasses are summarized below and in Tab. 5. The five selected variables are remarkably heterogeneous between classes, meaning the accuracy of these variables for clustering*:*

- Subclass 1 (drizzle and very light rain): the main feature of this class is the very low mean value (R_m) and standard deviation σ_R of rain rate events, in addition to the features of the superclass. The mean rain rate events lie in the range [0, 0.5] mm.h^{-1} with a mean value of 0.36 mm.h^{-1} and σ_R in the range [0, 3] mm.h^{-1}, with a mean value equal to 0.1 mm.h^{-1}. Although this event has a significant duration, the corresponding subclass, which corresponds to drizzle, involves only small quantities of water. It can also be noted that a low value of β_{L3} is a good indicator (< 0.01) for drizzle.

- Subclass 2 ("normal" events): this is a relatively broad class containing 68 % of all events, with a mean event rain rate (R_m) in the range [0.5 , 6] mm.h^{-1} and a mean value of 1.48 mm.h^{-1}. The standard deviation σ_R lies in the range [1, 10] mm.h^{-1} with a mean value of 2. This subclass is characterised by a significant relative variation of some parameters (D_e, R_m, P_{c1} for instance), together with dry periods (D$_d$), which may be sufficiently long.

The three remaining subclasses correspond to convective classes of events, which are characterized by a strong temporal heterogeneity and significant intensities. Depending on the depth of rain events, this convective class is subdivided into:

- Subclass 3: containing relatively long events (D_e) with high values for the rain event depth (R_d) variable and P_{c1}. This class represents events with a very small likelihood of occurrence (1.2%).

- Subclass 4: containing relatively short events (D_e) with peak rain rate R_{max} > 50 mm.h^{-1}, in addition to strong heterogeneities (σ_R, P_{C2} and P_{C3} are high) and large values for the convective indicator (β_{L3}).

- Subclass 5: the events in this subclass are characterised by relatively low values for the rain event depth (R_d). This is due to the short duration of the events (D_e). The variables σ_R and P_{C3} remain high. Another feature of this subclass is that it includes continuous events only, with no short, embedded dry periods (low values of D_d in Fig. 4 and Tab. 5).

Table 5: Summary of rain event subclasses computed with the learning dataset.

To conclude this section, this new classification allows the conventional definition for stratiform events to be refined. The convective classification can be subdivided into five different subclasses, each of which is homogeneous. This classification is obtained for mid-latitude climates. As the dataset used in this study is representative of only one specific region and topography (i.e. the temperate climate encountered in the Ile de France region, France), its analysis cannot reveal information related to different processes, i.e. those which are not sampled in the dataset. Such processes could lead to the identification of additional specific clusters of events. In particular, there are no orographic rainfall events or oceanic observations. The final step in this study involves assessing whether the homogeneous character of each class is preserved at the microphysics scale, and attempting to identify any relationships between the information present at the scale of both the microphysics and the macrophysics of these events (hydrological information).

5. Microphysical point of view

Our study of the microphysical properties of rain is based on a comprehensive analysis of its drop size distribution *N(D)*, corresponding to the number of raindrops per unit volume and per interval of diameter *D*. The shape of *N(D)* reflects the microphysical processes involved. The identification of various features of the drop size distribution, as well as the type of precipitation, is very useful for many applications. As an example, this information is used in the calculation of heating profiles in the precipitation parameterization of atmospheric models, to gain a more detailed understanding of microphysical processes, as well as for the development of rain retrieval algorithms applied to remote sensing observations. The microphysical characteristics of rainfall act as hidden variables that affect the relationship between microwave remote sensing measurements and the volume of water in a rainfall event (Ulaby, 1981; Iguchi, 2009). It can thus be very useful to use conventional rain gauges to determine the microphysical characteristics of rainfall events, thereby improving the quality of active or passive remote sensing observations, and the spatial properties of rainfall events in particular.

A general expression for the drop size distribution defined by Testud et al. (2001) is commonly used in the literature. This allows a distinction to be made between the stable shape function *f* and the variability induced by rain. This variability is represented by two microphysical parameters, namely the mass-weighted volume diameter (D_m) and the parameter N_0^*. In some studies, the term N_w is used rather than N_0^*. Not all authors use exactly the same units, in particular Bringi et al. (2003) and Suh et al. (2016) use mm^{-1}m^{-3} for the units of N_w, rather than the unit m^{-4} which is used in this study for N_0^*.

$$N(D) = N_0^* f(\frac{D}{D_m}) \qquad [\mathrm{m}^{-4}] \qquad (5)$$

where D_m and N_0^* are defined as:

$$D_m = \frac{M_4}{M_3} \quad [\text{mm}]\ , \quad N_0^* = \frac{4^4}{\Gamma(4)}\frac{M_3^5}{M_4^4} \quad [\text{m}^{-4}] \qquad (6)$$

and M_i is ith-order moment of the drop size distribution *N(D)*:

$$M_i = \int_0^{+\infty} N(D)D^i dD \qquad (7)$$

Rain samples are usually analysed by computing the microphysical parameters (D_m and N_0^*) for each rain sample obtained over a given time scale. In the present study, *N(D)* is obtained by considering the entire raindrop collection corresponding to each rain event of (variable) duration D_e. This approach leads to one pair (D_m, N_0^*) of microphysics variables per rain event, whereas most other authors rely on values computed over a fixed time scale.

Projections of the learned map, according to D_m and N_0^*, are shown in Fig. 4 (bottom right). It can be seen that the two maps are well structured, and that these two parameters have opposite influences on the map projection. Although these two microphysical parameters were not learned, the relationship between them is clearly accounted for by the information used to structure the map (the 5 selected variables). Moreover, the existence of a relationship between the microphysical and macrophysical features of the rainfall is also confirmed in this figure, since both of the macrophysical variables used to learn the SOM, i.e. σ_R and R_{max}, have patterns similar to those revealed on the D_m map.

Many authors, including Atlas et al., 1999; Bringi et al., 2003; Marzuki et al., 2013; Suh et al. 2016, have endeavoured to associate specific microphysical properties with each type of precipitation (convective or stratiform). In view of the maps shown in Fig. 4 and the convective/stratiform classification developed in section 4.3, we are able to confirm that precipitation events classified as stratiform express small values for D_m and large values for N_0^*. In the case of the convective class, the opposite trend is observed (i.e. larger values for D_m and smaller values for N_0^*). Similar observations have been reported by Testud et al. (2001). It can also be noticed that the two microphysical variables are relatively homogeneous in the convective class, whereas in the stratiform class they are characterised by a higher level of variability.
In order to improve our analysis of the microphysical information embedded in the dataset, we analysed the relationship between the two microphysical parameters using the reference vectors (neurons) from the map, which include information related to the original rain events.

Fig. 9 shows the variable D_m as a function of N_0^* for the 64 neurons on the map. This relationship is indicated through the use of distinct markers to identify the five subclasses defined in section 4.4, thus facilitating the discussion of the microphysics associated with stratiform and convective rain. The two solid lines show the linear regressions computed for these two classes.

In the case of the stratiform subclasses (1 and 2) a clear relationship can be observed between the two variables. The microphysics characteristics of these two subclasses are clearly distinct. Indeed, subclass 1 (drizzle and light rain) has the smallest D_m and the highest N_0^*, and varies over just a small range. Conversely, as in the case of the macrophysical variables (see section 4.4), the microphysical characteristics of subclass 2 (normal events) are considerably more heterogeneous. Knowledge of D_m makes it straightforward to identify the corresponding subclass. As a consequence, an event with D_m lying in the range [0.5, 1] millimetre belongs to subclass 1. Similarly, it is very likely that an event with D_m lying in the range [1, 1.7] millimetre belongs to subclass 2.

For the convective events (subclasses 3, 4, 5), small differences can be noticed with respect to N_0^*. In the range [1.7 , 2.5] mm, two neurons belonging to subclass 4 are close to a neuron belonging to subclass 5, and therefore have similar microphysics. Although they are located far from all other subclass 2 neurons, three isolated neurons belonging to subclass 2 (stratiform) can be noted. These are characterised by relatively strong values of D_m (2 mm) and low values of N_0^*. The corresponding events are a mixture of stratiform and convective rain. A typical case is given by convective rain associated with strong rain rates occurring at the beginning of an event, whereas the remainder of the event is stratiform with low rain rates and small variations.

Figure 9: Microphysical variable N_0^* versus $\mathbf{D_m}$ for the five rainy event subclasses. The three neurons corresponding to mixed events are circled. The dashed lines indicate the limits defined by $\mathbf{D_m}$ > 1.66 and Log(N_0^*)>6.15

Following our classification, Fig. 9 indicates that there are real relationships between the macrophysical and microphysical variables. Nevertheless, knowledge of the variables (D_m, N_0^*) does not allow the correct subclass to be determined in all cases.

Researchers who study microphysical features and their association with specific types of precipitation use simple schemes, based on rain rate estimations over a fixed period of integration (a few minutes), in order to separate stratiform and convective rain types. They also use these simple schemes to label *D_m* and N_0^* as stratiform or convective (Testud et al., 2001). This approach is significantly different to the method presented here, which assumes that all of the samples in a given event belong to the same class. Our values for N_0^* and D_m are thus computed for the time scale of a given event, rather than for a fixed integration time. Thus, although in the present study a good agreement is found for the range of values covered by D_m, those determined for N_0^* do not cover the same range as in the case of the previously cited studies.
Many previous authors have observed that the drop size distribution is closely related to processes controlling rainfall development mechanisms. In the case of stratiform rainfall, the residence time of the drops is relatively long, and the raindrops grow by the accretion mechanism. In convective rainfall, raindrops grow by the collision–coalescence mechanism, associated with relatively strong vertical wind speeds. Numerous studies have been published concerning the variability of N_0^* and D_m: Bringi et al. (2003) studied rain samples from diverse climates and analysed their variability in stratiform and convective rainfall; Marzuki et al. (2013) investigated the variability of the raindrop size distribution through a network of Parsivel disdrometers in Indonesia; and Suh et al. (2016) investigated the raindrop size distribution in Korea using a POSS disdrometer. In the case of stratiform rain, all of these authors observe that N_0^* and D_m are nearly log-linearly related, with a negative slope. This is consistent with the trend shown in Fig. 9 for the two stratiform subclasses (1 and 2). Even the three distinct neurons, which are isolated from the others, appear to be governed by the same relationship.
Marzuki et al. (2013) noted that during convective rain the increase in value of N_0^* with decreasing D_m is nearly log-linear, with a flatter slope. In the present case, the dependence is also log-linear, with a slope that is slightly flatter for convective events than for stratiform events. In the aforementioned studies, the data was aggregated over time, campaign or site, on the basis of a criterion computed over a fixed period of time. We believe that this process is weakly suited to determining the properties of convective events, as a consequence of their strong variability and shorter characteristic time. In this study we were able to retrieve the log-linear relationship between N_0^* and D_m without having to learn it directly.
When applying our algorithm to the various macroscopic properties by rain event, we also take into account the variability of rain within an individual rain event. Fig. 9 clearly shows that the spreading of parameters N_0^* and D_m inside each subclass has the same magnitude as the distance between subclasses. This remark confirms the hypothesis of Tapiador et al. (2010): the intra-event variability can exceed the inter-event variability, due to events arising from different precipitation systems. It is thus preferable to examine the properties of events with a more general approach, rather than using individual samples to study the distinction between stratiform and convective processes. The three isolated neurons in subclass 2 described above (circled in Fig. 9) have the same properties as the other events of their subclass (i.e. the same slope for the log-linear relationship between N_0^* and D_m). This example confirms the ability of our methodology to preserve the macroscopic information needed to cluster rain events, thus allowing the intra-event variability as well as microphysical information to be (partially) retrieved.
Suh et al. (2016) also compare log(N_0^*) and D_m pdf, for the case of stratiform and convective samples over a 4-year period. On the basis of the D_m pdf of both stratiform and convective classes, they compute a threshold value for D_m, such that when D_m >

1.66 mm the rainfall samples are mainly convective, and when D_m < 1.66 mm they are mainly stratiform. This finding is consistent with the results of Atlas et al. (1999), who also found a threshold value for D_m, distinguishing between convective and stratiform rainfall. In Fig. 9, it can be seen that this threshold is confirmed (vertical solid line), with D_m smaller than 1.6 mm corresponding to stratiform events, whereas higher values correspond to mainly convective events. When we consider the events analysed in the present study, there are also three neurons corresponding to a “mixed event” beyond this threshold.

Suh et al. (2016) show in Fig. 4c of their study that the pdf for convective rainfall is higher than that corresponding to stratiform rainfall, when $\log(\mathrm{N}_0^*) > 6.2$ ($\mathrm{N_w} = 3.2$ in their figure). As described above, by considering the data corresponding to rain events, rather than to samples recorded over fixed periods of time, our range of values for N_0^* is smaller than that used in other publications. In addition, $\log(\mathrm{N}_0^*) < 6.15$ for all neurons labelled as convective in our study, which is very close to the value of 6.2 determined by Suh et al. (2016).

In view of the generally satisfactory retrieval of microphysical information from macrophysical parameters, we are of the opinion that the topological map successfully restores some of the information implicitly embedded in the dataset. It is thus interesting to note that the macrophysical parameters of rainfall are related to its microphysical properties. Firstly, the map collects similar events, whilst ensuring, through the minimization of topological errors, that the unfolding of the map is correct. A neuron is thus closer to its neighbours than to any other neuron on the map. This criterion ensures that the data space is optimally partitioned into connected subparts, such that the neurons on the map can be related to the underlying processes governing rainfall.

6. **Conclusion**

Although the definition of a ‘rain xc event’ is relatively subjective, this study underlines the advantages of using event analysis rather than sample analysis. This data-driven analysis of events shows that rain events exhibit coherent features. As a consequence of the discrete and intermittent nature of rainfall, some of the features commonly used to describe rain processes are inadequate, in particular when they defined for a fixed duration. Excessively long integration times (hours or days) can lead to the mixing of observations that correspond to distinct physical processes, and also to the mixing of rainy and clear air periods, within the same sample. An excessively short integration time (seconds, minutes) leads to noisy data, which is sensitive to the sensor’s characteristics (sensor area, detection threshold and noise). By analyzing entire rain events, rather than short individual samples of fixed duration, it is possible to clearly identify certain relationships between the different features of rain events, in particular the influence of the microphysical properties of rain on its macrophysical characteristics. This approach allows the intra-event variability caused by measurement uncertainties to be reduced, thus improving the accuracy with which physical processes can be identified.

Once an event has been clearly identified, it is possible to choose a small number of variables to describe it. We present a new data-driven approach, which can be used to select the most relevant variables for this characterization. This approach has generic properties and can be adapted to many multivariate applications. A genetic algorithm, when combined with Self-Organizing Map (SOM) clustering, can allow the unsupervised selection of an optimal subset of five macrophysical variables. This is achieved by minimizing a score function, which depends on the topology error of the SOM and the number of variables. This score provides a parsimonious description of the event, whilst preserving as much as possible the topology of the initial space.

Numerous variables derived mainly from rain rate recordings are used to describe precipitation in the context of rain time series studies, and a wide variety of topics of interest, including hydrology, meteorology, climate, and weather forecasting. The algorithm proposed in this study produces a subspace formed by only 5 of the 23 rain features described in the literature. We show that these five features can be selected by the algorithm in an unsupervised manner and, from the macrophysical point of view, can provide an adequate description of the main characteristics of rainfall events. These characteristics are: the event duration, the peak rain rate, the rain event depth, the standard deviation of the event rain rate, and the absolute rain rate variation of order 0.5.

In order to confirm the relevance of the five selected features, we analyze the corresponding SOM and are able to clearly reveal the presence of relationships between these features. This approach also reveals the independence of the inter-event time (IET_p) characteristic, and the weak dependence of the Dry percentage in event ($D_{d‰e}$) characteristic, thus confirming that a rain time series can be considered as an alternating series of independent rain events, interrupted by periods without rain. Hierarchical clustering allows the well-known separation between stratiform and convective events to be clearly identified. This dual classification is then refined into a set of five relatively homogeneous subclasses. The stratiform class is divided into 2 subclasses: a drizzle / very light rain subclass, and a normal event subclass. The convective class is divided into 3 subclasses, characterized by a strong temporal heterogeneity and significant rain rates.

As this research was based on the analysis of observations made in mid-latitude plains in France, the relevance of this classification remains to be confirmed through the analysis of datasets recorded in different climatic zones, and under different meteorological conditions, such as those encountered in mountainous or coastal areas. If the SOM described in the present study were learned with a more exhaustive dataset, a larger map would be produced, and this could reveal new types of rainfall behavior, which remained undetected in the current dataset. This point will be addressed in future studies.

The data-driven analysis of entire rain events (rather than the analysis of fixed-length samples) is relevant to the study of interactions between the macrophysical (based on the rain rate) and microphysical (based on raindrop) properties of rain. In the present study, several strong relationships were identified between these microphysical and macrophysical characteristics, and we show that some of the five subclasses identified in this analysis have specific microphysical characteristics. When a relationship between the microphysical and macrophysical properties of rain is identified, this can have many practical implications, especially for remote sensing. In the context of weather radar applications, the microphysical properties of rain are needed in order to estimate rain rates, through the use of the Z–R relationships. The estimation of microphysical rain characteristics, based on easily observable rain gauge measurements, could play a significant role in the development of the quantitative precipitation estimation (QPE).

References

Akrour, N., Chazottes, A., Verrier, S., Mallet, C., and Barthes, L.: Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution. Water Resources Research, 51(9), 7417–7435, 2015.

Atlas, D., Ulbrich, C.W., Marks, F. D., Amitai, E., and Williams, C. R.: Systematic variation of drop size and radar-rainfall relations, J. Geophys. Res.-Atmos., 104, 6155–6169, 1999.

Balme, M., Vischel, T., Lebel, T., Peugeot, C., and Galle, S. : Assessing the water balance in the Sahel: impact of small scale rainfall variability on runoff Part 1: rainfall variability analysis. Journal of Hydrology 331:336–348, 2006.

Bringi, V. N., Chandrasekar, V., Hubbert, J., Gorgucci, E., Randeu, W. L., and Schoenhuber, M.: Raindrop size distribution in different climatic regimes from disdrometer and dual-polarized radar analysis, J. Atmos. Sci., 60, 354–365, 2003.

Brown, B.G., Katz, R. W., and Murphy, A.H.: Statistical analysis of climatological data to characterize erosion potential: 1. Precipitation Events in Western Oregon. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University (1983).

Brown, B.G., Katz, R. W., and Murphy, A.H.: Statistical analysis of climatological data to characterize erosion potential: 4. Freezing events in eastern Oregon/Washington. Oregon Agricultural Experiment Station Spec. Rep. No. 689, Oregon State University, 1984.

Brown, B.G., Katz, R. W., and Murphy, A.H.: Exploratory Analysis of Precipitation events with Implications for Stochastic Modeling. Journal of Climate and Applied meteorology(57-67), 1985.

Cosgrove, C.M. and Garstang, M.: Simulation of rain events from rain-gauge measurements. International Journal of Climatology 15, 1021–1029, 1995.

Coutinho, J.V., Almeida, C. Das, N., Leal. A.M. F., Barbarosa, L. R.: Characterization of sub-daily rainfall properties in three rain gauges located in northeast Brazil. Evolving Water Resources Systems: Understanding, Predicting and Managing Water–Society Interactions Proceedings of ICAR 2014, Bologna, Italy, 345-350, 2014.

Daumas, F. : Méthodes de normalisation de données, Revue de statistique appliquée, 30(4), 23-38, 1982.

Driscoll, E. D., Palhegyi, G. E., Strecker, E. W., & Shelley, P. E. (1989). Analysis of storm events characteristics for selected rainfall gauges throughout the United States. *US Environmental Protection Agency, Washington, DC.*

Dunkerley, D.: Rain event properties in nature and in rainfall simulation experiments: a comparative review with recommendations for increasingly systematic study and reporting, Hydrological Processes, 22(22), 4415–4435, 2008a.

Dunkerley, D.: Identifying individual rain events from pluviograph records: a review with analysis of data from an Australian dryland site, Hydrological Processes, 22(26), 5024–5036, 2008.

Eagleson, P. S.: Dynamic Hydrology, McGraw-Hill, 1970.

Galmarini, S., Steyn, D. G., and Ainslie, B.: The scaling law relating world point-precipitation records to duration. Int. J. Climatol., 24, 533–546, 2004.

Everitt, B.: Cluster Analysis. London: Heinemann Educ. Books, 1974.

Delahaye, J.-Y., Barthès, L., Golé, P., Lavergnat J., and Vinson, J.P.: a dual beam spectropluviometer concept, Journal of Hydrology, 328(1-2), 110-120, 2006.

Gargouri, E., Chebchoub, A.: Modélisation de la structure de dépendance hauteur-durée d'événements pluvieux par la copule de Gumbel. Hydrological Sciences–Journal–des Sciences Hydrologiques, 53(4), 802-817, 2010.

Grazioli, J., Tuia, D., and Berne, A.: Hydrometeor classification from polarimetric radar measurements: a clustering approach, Atmos. Meas. Tech., 8, 149-170, 2015.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection (Kernel Machines Section), 3, 1157--1182, 2003.

Haile, A. T., Rientjes, T. H. M., Habib, E., Jetten, V., and Gebremichael, M.: Rain event properties at the source of the Blue Nile River, Hydrol. Earth Syst. Sci., 15, 1023-1034, 2011.

Iguchi, T., Kozu, T., Kwiatkowski, J., Meneghini, R., Awaka, J., and Okamoto, K.: Uncertainties in the rain profiling algorithm for the TRMM precipitation radar. J. Meteor. Soc. Japan, 87A, 1-30, 2009.

Holland, J. H.: Adaptation In Natural And Artificial Systems, University of Michigan Press, 1975.

Kohonen, T.: Self-organizing formation of topologically correct feature maps. Biological Cybernetics, 46, 59-69, 1982.

Kohonen, T. (2001). Self-Organizing Maps. Springer-Verlag, ISBN 3-540-67921-9, New York, Berlin, Heidelberg, 2001

Larsen, M. L. and Teves, J. B.: Identifying Individual Rain Events with a Dense Disdrometer Network, Advances in Meteorology, 2015, ID582782, 2015.

Lavergnat, J. and Golé, P.: A Stochastic Raindrop Time Distribution Model. Journal of Applied Meteorology, 37, 805-818, 1998.

Lavergnat, J. and Golé, P.: A stochastic model of raindrop release: Application to the simulation of point rain observations, Journal of Hydrology, 328(1), 8-19, 2006.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K.: Performance evaluation of the self- organizing map for feature extraction, Journal of Geophysical Research, 111, C05018, doi:10.1029/2005JC003117, ISSN 0148-0227, 2006

Liu, Y. and Weisberg R.H.: A review of self-organizing map applications in meteorology and oceanography. In: Self-Organizing Maps-Applications and Novel Algorithm Design, 253-272, 2011.

Llasat, M.C.: An objective classification of rainfall events on the basis of their convective features. Application to rainfall intensity in the north east of Spain. International Journal of climatology, 21, 1385-1400, 2001.

Marzuki, M., Hashiguchi, H., Yamamoto, M. K., Mori, S., and Yamanaka, M. D.: Regional variability of raindrop size distribution over Indonesia, Ann. Geophys., 31, 1941-1948, 2013.

Molini, L., Parodi, A., Rebora, N., Craig, G. C.: Classifying severe rainfall events over Italy by hydrometeorological and dynamical criteria. Quarterly Journal of the Royal Meteorological Society, 137(654), 148-154, 2011.

de Montera, L., Barthes, L., and Mallet, C.: The effect of rain-no rain intermittency on the estimation of the Universal Multifractal model parameters, J. of Hydrometeorology, 10, pp. 493–506, 2009.

Moussa, R. and Bocquillon, C.: Caractérisation fractale d'une série chronologique d'intensité de pluie. Rencontres hydrologiques Franco-Romaines, 363-370, 1991.

Suh, S.-H., You, C.-H., and Lee, D.-I.: Climatological characteristics of raindrop size distributions in Busan, Republic of Korea, Hydrol. Earth Syst. Sci., 20, 193-207, 2016.

Tapiador, F. J., Checa, R., and de Castro, M.: An experiment to measure the spatial variability of rain drop size distribution using sixteen laser disdrometers. Geophysical Research Letters, 37(16), ID L16803, 2010.

Testud, J., S., Oury, P., Amayenc, and Black, R. A.: The concept of ''normalized'' distributions to describe raindrop spectra: A tool for cloud physics and cloud remote sensing, J. Appl. Meteor., 40, 1118–1140, 2001.

Ulaby, F. T., Moore, R. K., and Fung, A. K.: Microwave Remote Sensing: Fundamentals and Radiometry. Vol. I. Artech House, 321-327, 1981.

Uriarte, E. A. and Martín, F. D., Topology Preservation in SOM, World Academy of Science, Engineering and Technology. International Journal of Computer, Electrical, Automation, Control and Information Engineering 2, 9, 2008

Verrier, S., Barthès L., Mallet C.: Theoretical and empirical scale dependency of Z-R relationships: Evidence, impacts, and correction, Journal of Geophysical Research: Atmospheres, 118 (14), 7435-7449, 2013.

Vesanto, J. and Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks, Vol. 11, 586–600, ISSN 1045-9227, 2000

Tables

	Observation period	Availability (%)	Number of rain events
Learning data set	01/01/2013-12/31/2014	96.4%	234
Test data set	04/16/2008-01/31/2012	60%	311

Table 1: Observation periods, availability of DBS observations, and numbers of rain events for the learning and test datasets

Number (#)	Feature name	symbol	Formula	Normalisation
1	Event duration	D_e	$D_e = T_{end} - T_{begin} + 1$ [min] With T_{begin}: Event start time and T_{end}: Event end time	1
2	Mean event rain rate	R_m	$R_m = \frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h^{-1}]	2
3	Intra-dry duration	D_d	$D_d = \sum_{t=T_{begin}}^{t=T_{end}} I_t$ [min] With $I_t = \begin{cases} 1 & if\ RR_t = 0\ [\text{mm}\ h^{-1}] \\ 0 & else \end{cases}$	0
4	First quartile	Q_1	The 25th percentile [mm h^{-1}]	0
5	Median	Q_2	the 50th percentile [mm h^{-1}]	0
6	Third quartile	Q_3	The 75th percentile [mm h^{-1}]	2
7	Previous IET	IET_p	$IET_p = T_{begin}(current\ event) - T_{end}(previous\ event) + 1$ [min]	0
8	Mean rain rate over the rainy period	$R_{m,r}$	$R_{m,r} = \frac{1}{(D_e - D_d)}\sum_{t=T_{begin}}^{t=T_{end}} RR_t$ [mm h^{-1}]	3
9	Event Rain rate std.	σ_R	$\sigma_R = \sqrt{\frac{1}{D_e}\sum_{t=T_{begin}}^{t=T_{end}} (RR_t - R_m)^2}$ [mm h^{-1}]	2
10	Mode	M_0	M_0 =the most frequent RR$_t$	0
11	Rain rate peak	R_{max}	$R_{max} = \max(RR_t)$	2
12	Dry Percentage in event	$D_{d\%e}$	$D_{d\%e} = \frac{D_d}{D_e}$	5
13	Rain event depth	R_d	$R_d = R_m * D_e /60$ [mm]	0
14	Absolute rain rate variation of order c	P_{c1}	$P_{c_i} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert RR_{t+1} - RR_t \rvert^{c_i}$ For $c_i = 0.5, 1, 2$	6
15		P_{c2}		3
16		P_{c3}		2
17	Normalized Absolute rain rate variation of order c$_i$	$P_{C_{N1}}$	$P_{C_{Ni}} = \frac{P_{c_i}}{D_e}$ For $i = 1\,..3$	3
18		$P_{C_{N2}}$		2
19		$P_{C_{N3}}$		0
20	Absolute rain rate variation of order C and threshold S	$P_{S,C}$	$P_{S,C} = \sum_{t=T_{begin}}^{t=T_{end}-1} \lvert \max[(RR_{t+1} - S), 0] - \max[(RR_t - S), 0] \rvert^{C}$ With $s = 0.3\ and\ c = 2$	6
21	β_L parameter	β_{L1}	$\beta_{L_i} = \frac{\sum_{i=T_{begin}}^{T_{end}} RR_t\ \theta(RR_t - L_i)}{\sum_{i=T_{begin}}^{T_{end}} RR_t}$ For L$_i$ = $0.3\,, 1\,, 3\ mm\ h^{-1}$ With $\theta(RR_t - L_i)$ is the Heaviside function defined as $\theta(RR_t - L_i) = 1\ if\ RR_t \geq L_i$ $\theta(RR_t - L_i) = 0\ if\ RR_t < L_i$	5
22		β_{L2}		0
23		β_{L3}		0

Tablec2: The 23 variables identified in the literature, used for the characterization of rain events

Transformation number	Transformation name	Formula f(x)	Note & remark
0	Standardisation	$\frac{x - mean(x)}{std(x)}$	-
1	Power	x^n	$n = 0.05$
2	Boxcox	$\frac{(x^\gamma - 1)}{\gamma}$	$\gamma = -0.1$
3			$\gamma = -0.2$
4			$\gamma = -0.3$
5	Arc-sin of square	$\arcsin\sqrt{x}$	Data are between 0 and 1
6	decimal Logarithm	$\mathrm{Log}(x + c)$	$c = 0.1$

Table 3: Transformations used to normalize the variables listed in Table 2.

Variables	D_e	R_m	D_d	Q_1	Q_2	Q_3	IET_p	$R_{m,r}$	σ_R	M_0	R_{max}	$D_{d\%e}$
R^2 learninglearning data set	0.96	0.91	0.58	0.57	0.48	0.77	0.31	0.93	0.97	0.50	0.96	0.52
R^2 Test data set	0.93	0.84	0.55	0.57	0.50	0.74	0.26	0.84	0.86	0.54	0.82	0.50

Variables	R_d	P_{c1}	P_{c2}	P_{c3}	P_{cN1}	P_{cN2}	P_{cN3}	$P_{S,C}$	β_{L1}	β_{L2}	β_{L3}	-
R^2 learning data set	0.97	0.97	0.93	0.94	0.91	0.95	0.78	0.94	0.70	0.89	0.96	
R^2 Test data set	0.94	0.99	0.91	0.83	0.83	0.85	0.71	0.82	0.61	0.76	0.89	-

Table 4: Coefficient of determination obtained on the learning and test datasets. The values with a dark grey background correspond to the 5 selected variables

	Stratiform events		Convective events		
	Subclass 1	**Subclass 2**	**Subclass 3**	**Subclass 4**	**Subclass 5**
Variables	Mean	Mean	mean	mean	mean
$D_e(min)$	321	149	464	75	49
σ_R	0.36	2.01	3.62	11.7	9.64
$R_{max}(mm\ h^{-1})$	2.08	10	22	52.7	36.06
$R_d(mm)$	1.99	2.62	11.24	6.9	2.72
P_{c1}	75.7	64.5	193	78.2	40.94
$R_m(mm\ h^{-1})$	0.37	1.48	2.35	7.85	7.11
$D_d(min)$	80	31	75	11	1
β_{L3}	0.01	0.42	0.48	0.89	0.86

Table 5: Summary of the rain event subclasses computed with the learning dataset.

Figures

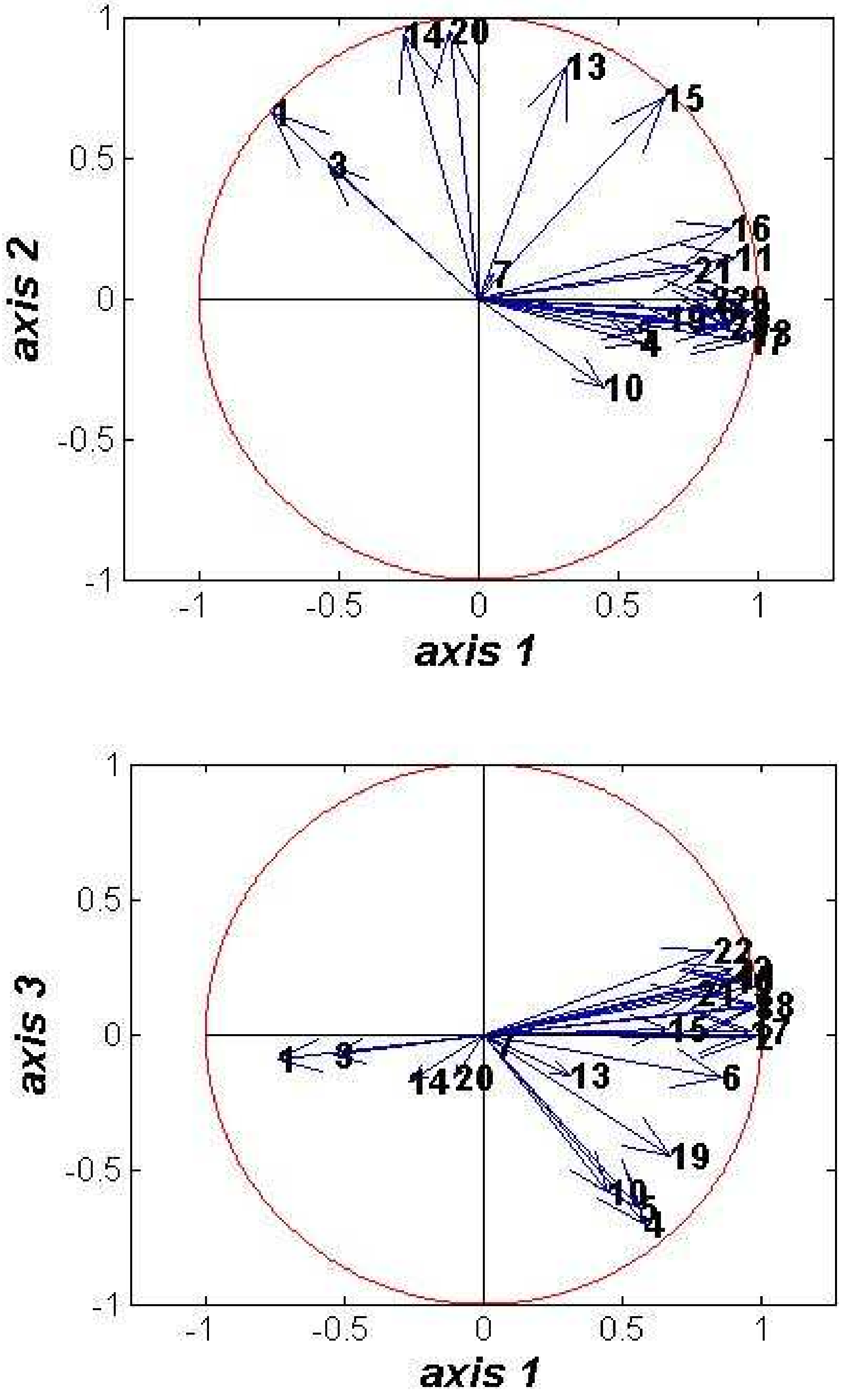


Figure 1: PCA on the learning data set based on the 23 variables described in Table 3. Left: correlation circle on axes 1 & 2. Right: correlation circle on axes 1 & 3. All of the variables are normalised according to the last column of Table 2.

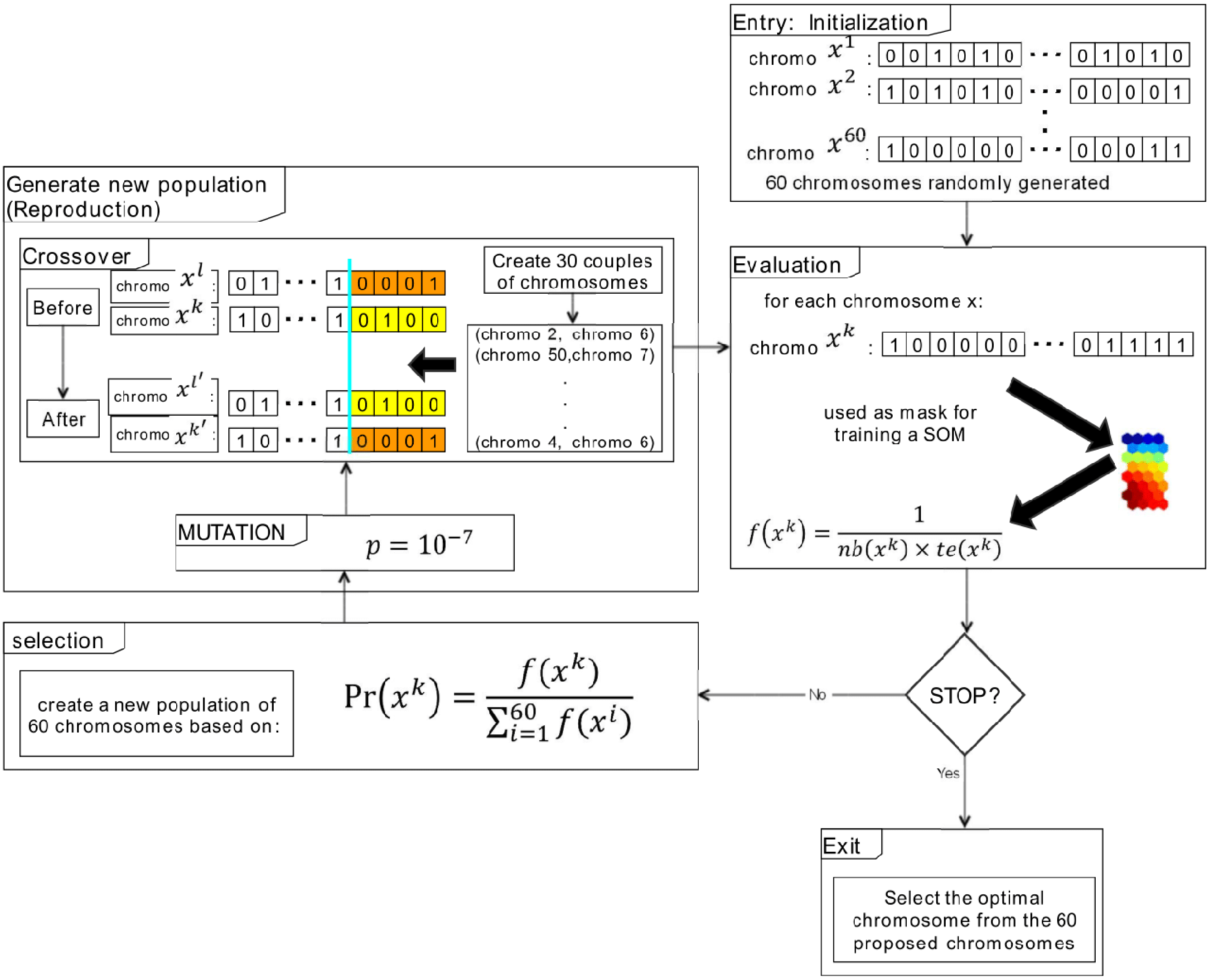

Entry: Initialization
chromo x^1 : 0 0 1 0 1 0 ··· 0 1 0 1 0
chromo x^2 : 1 0 1 0 1 0 ··· 0 0 0 0 1
chromo x^{60} : 1 0 0 0 0 0 ··· 0 0 0 1 1
60 chromosomes randomly generated
Evaluation
for each chromosome x:
chromo x^k : 1 0 0 0 0 0 ··· 0 1 1 1 1
used as mask for training a SOM
$f(x^k) = \frac{1}{nb(x^k) \times te(x^k)}$
STOP?
No
Yes
Exit
Select the optimal chromosome from the 60 proposed chromosomes
selection
create a new population of 60 chromosomes based on:
$\Pr(x^k) = \frac{f(x^k)}{\sum_{i=1}^{60} f(x^i)}$
Generate new population (Reproduction)
Crossover
Before
chromo x^l : 0 1 ··· 1 0 0 0 1
chromo x^k : 1 0 ··· 1 0 1 0 0
After
chromo $x^{l'}$: 0 1 ··· 1 0 1 0 0
chromo $x^{k'}$: 1 0 ··· 1 0 0 0 1
Create 30 couples of chromosomes
(chromo 2, chromo 6)
(chromo 50,chromo 7)
(chromo 4, chromo 6)
MUTATION
$p = 10^{-7}$

Figure 2: Diagram for the selection of variables based on a Genetic Algorithm associated with Kohonen Maps

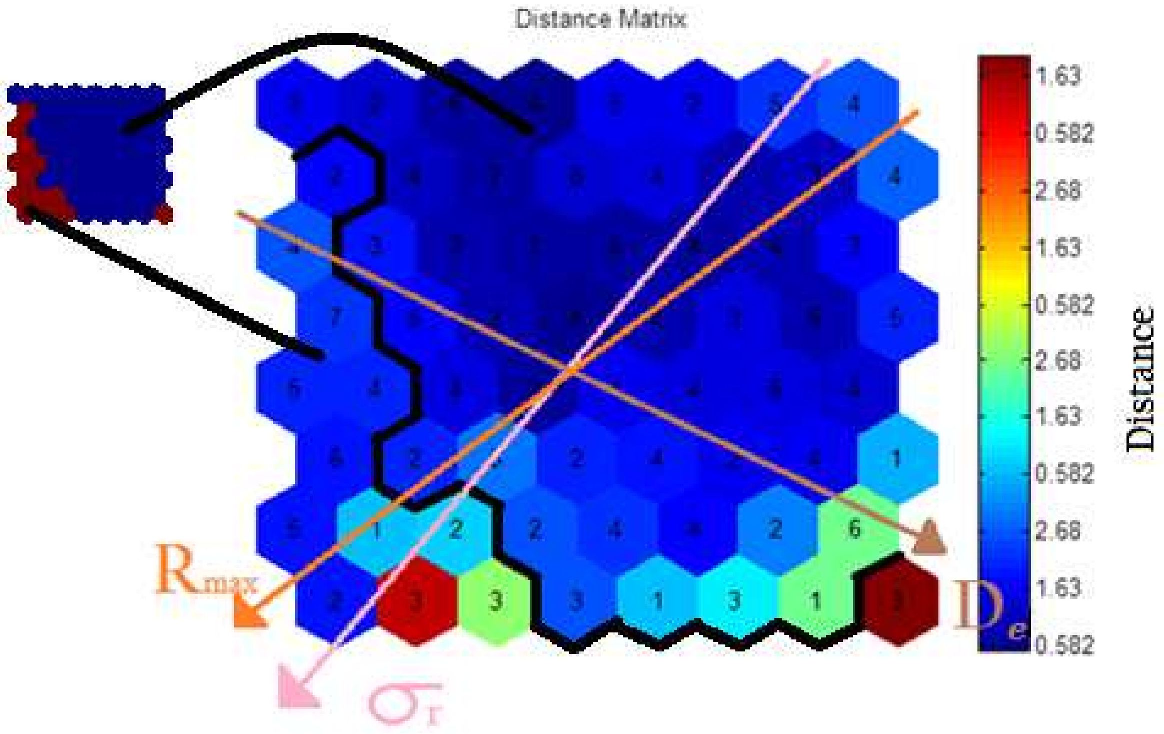


Figure 3: Distance matrix for the $M(x^{Best})$ map: The colour of each neuron represents the average distance between itself and its neighbouring neurons. The value inside each neuron indicates the number of rain events that it has captured inside the learning dataset. The black line separates the neurons into 2 classes, using the Hierarchical Ascendant Classification (see section 4.1). The arrows represent the gradients of the variables R$_{max}$, σ_R and D$_e$

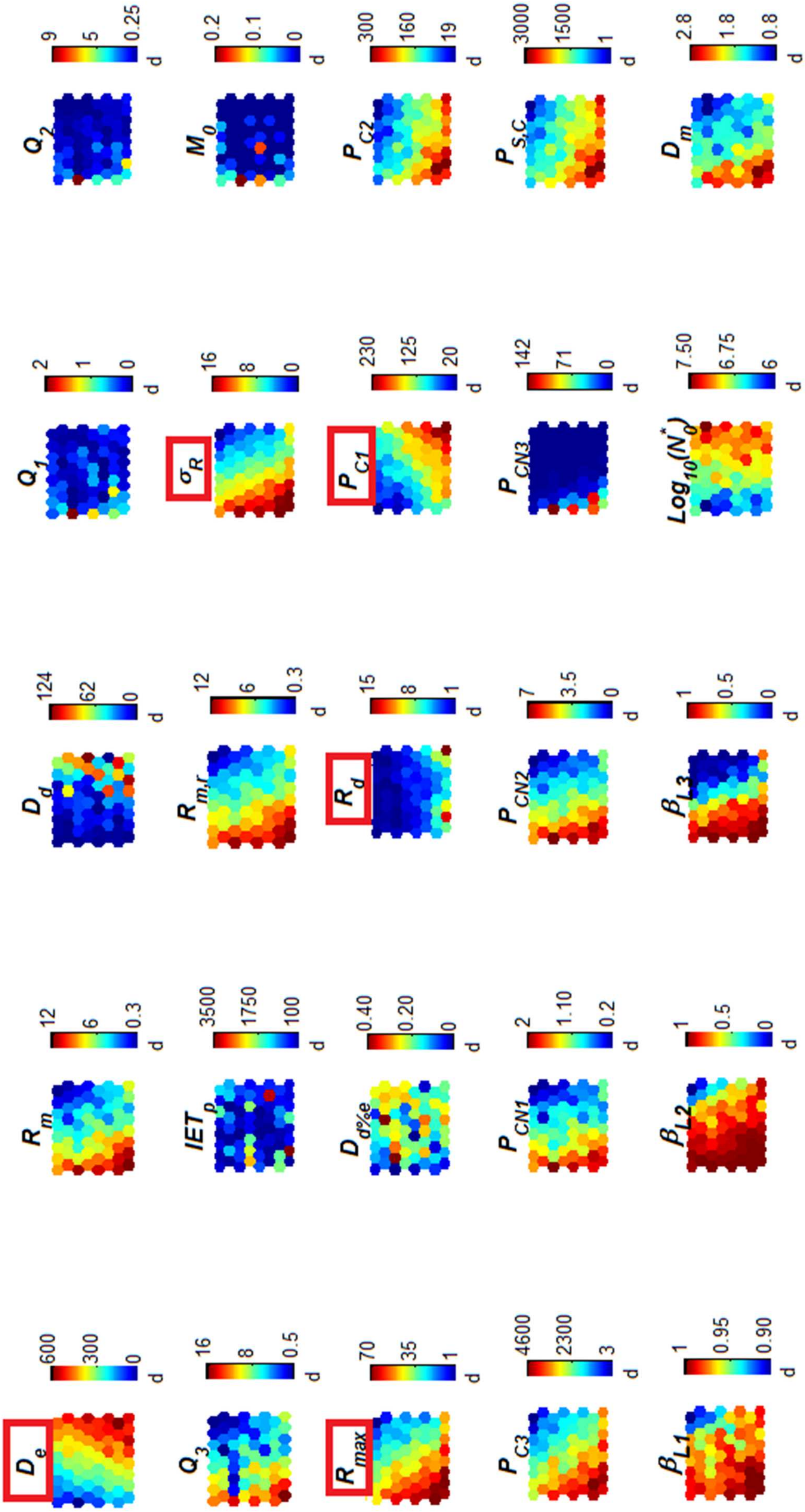


Figure 4: Projection of the $M(x^{Best})$ map according to the 23 variables. The red-framed variables are those selected by the GA algorithm. The last two variables D$_m$ and $N_0^*$are defined in section 5.

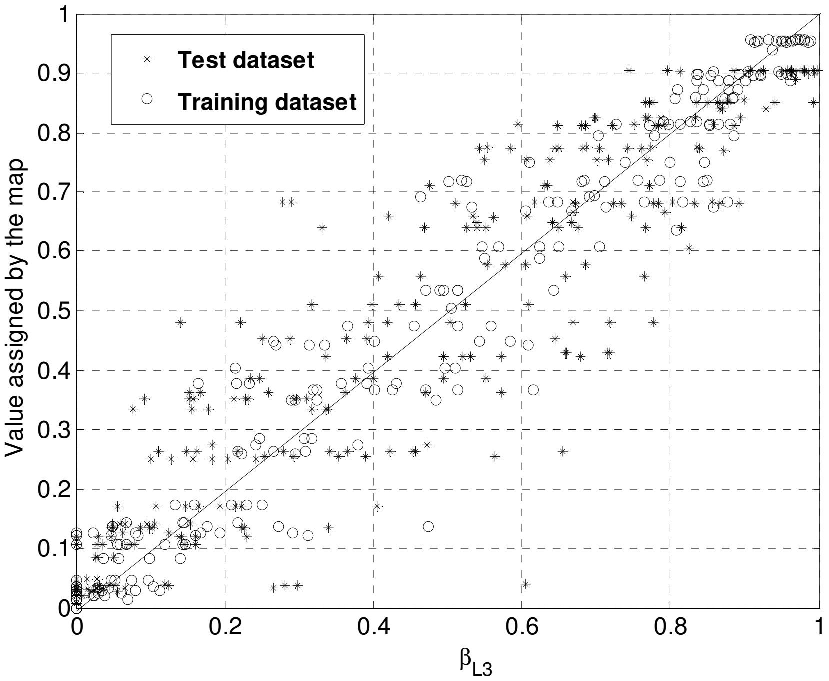


Figure 5: The variable β_{L3} versus its corresponding value, given by the best matching unit: from the learning dataset (circles), and the test dataset (stars). The solid line corresponds to the first diagonal.

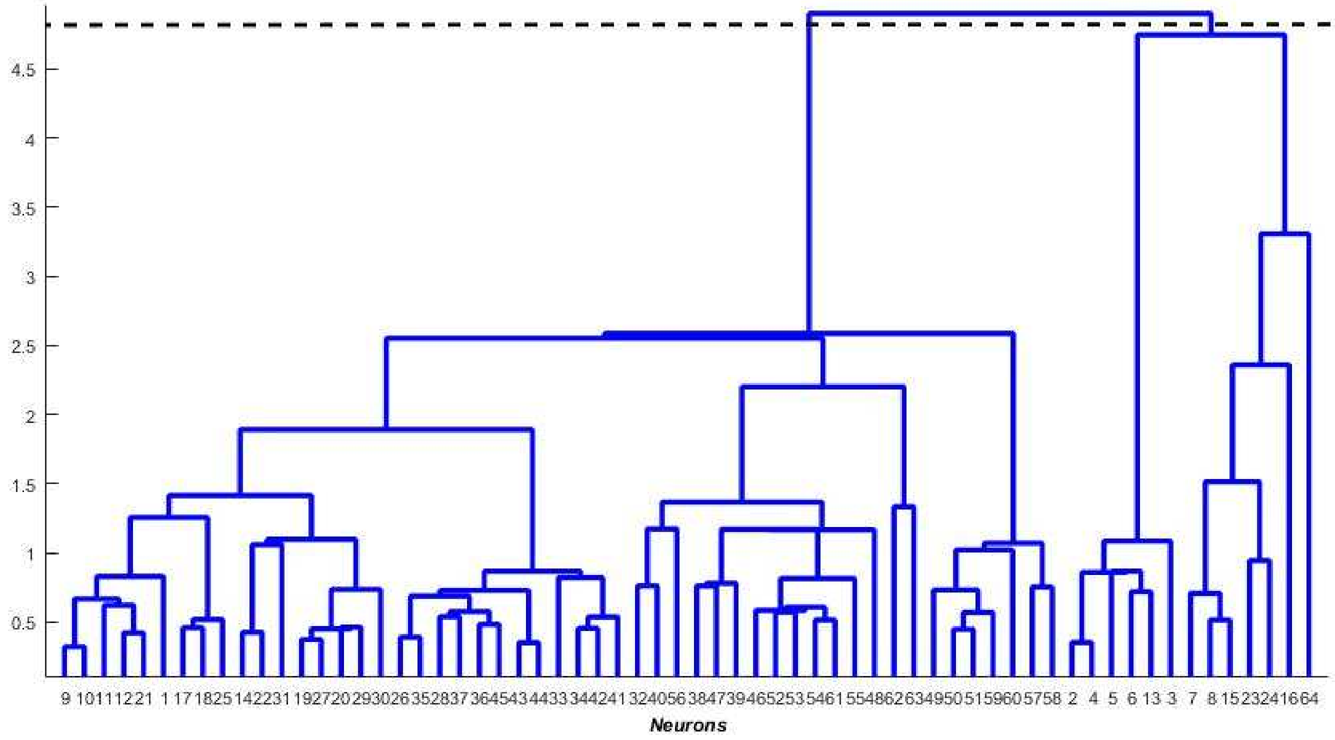


Figure 6 Dendrogram obtained from the Hierarchical Cluster Analysis of the 64 neurons in the SOM. The horizontal dashed line represents the threshold between the two classes.

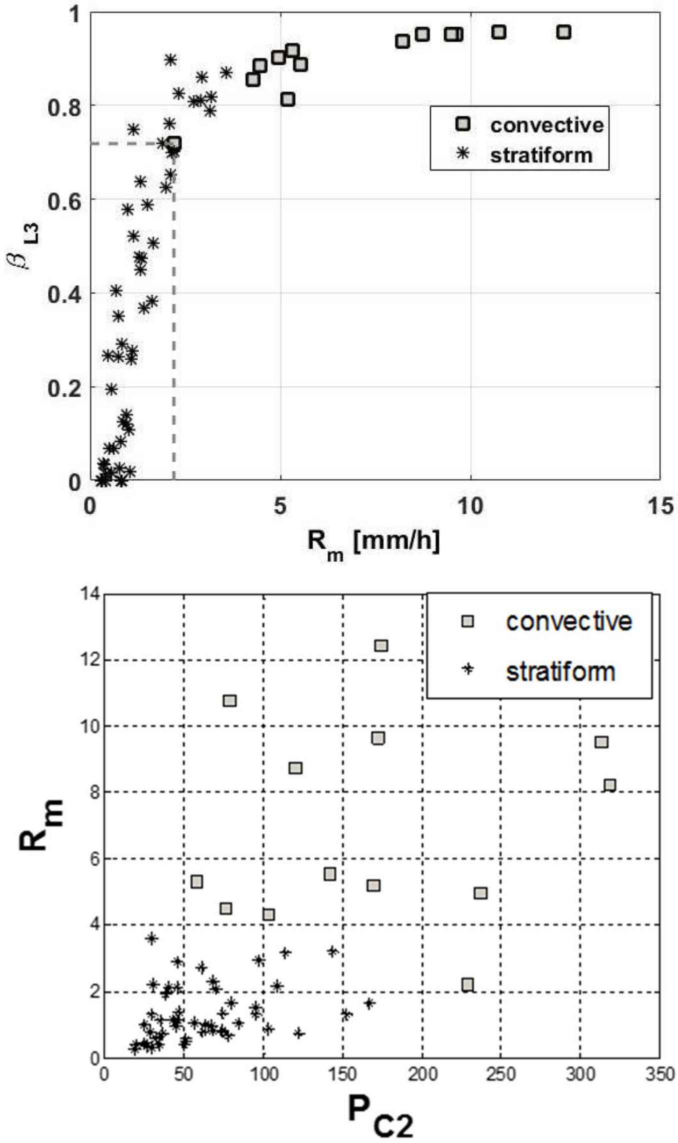


Figure 7: Representation of the neurons in the Rm, $\beta_{L3}$and R_m, P_{C2} sub-spaces. The stars represent neurons from group 1 (stratiform), and the squares correspond to neurons from group 2 (convective). Dashed lines indicate the neuron #64.

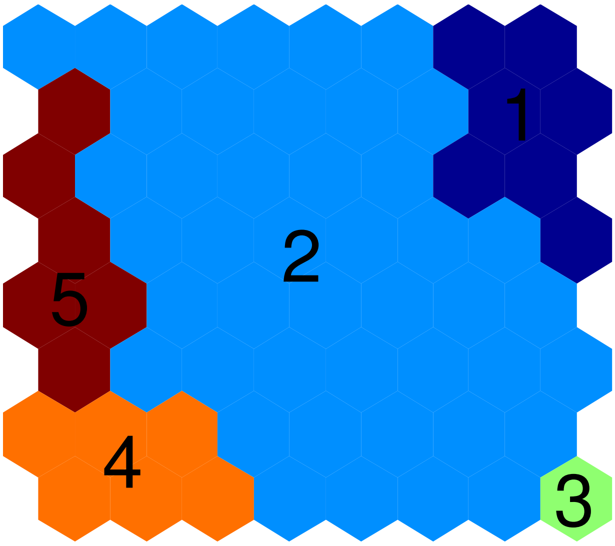
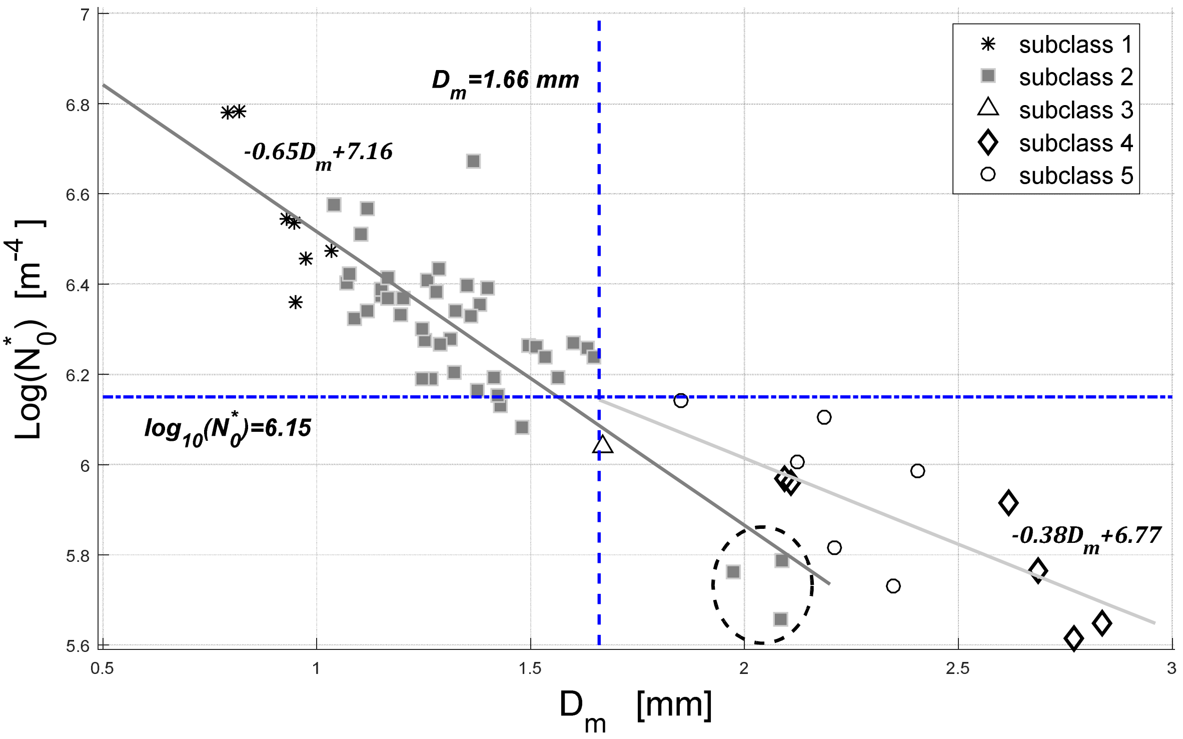


Figure 8: Hierarchical Clustering of the map into five subclasses. The colours represent the subclass numbers: Subclass 1: Dark blue, Subclass 2 : blue, Subclass 3 : Green, Subclass 4 : Orange, Subclass 5 : Red

5



5 *Figure 9: Microphysical variable N_0^* versus D_m for the five rainy event subclasses. The three neurons corresponding to mixed events are circled. The dashed lines correspond to borders D_m > 1.66 and Log(N_0^*)>6.15*