Lightning stroke clustering into cloud-to-ground lightning flashes

Author: Eloi Dalmasso Blanch
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Abstract: Cloud-to-ground stroke data have been analysed for the years 2010-2012 for the Catalan region. These strokes have been grouped together into lightning flashes using a hierarchical clustering method, with two different approaches to the clustering radius, as the objective of the study was to evaluate different clustering methods. The results for the two approaches were compared with the results from a step-by-step method, giving us differences of 2% and 7% in the final number of lightning flashes. Using different values for the *lightning flash window* and *lightning spatial radius* the multiplicity values from the clustering methods results ranged from 1.6 to 2.0. These results can improve the reliability of studies where the number of lightning flashes is important such as in severe weather case studies or in climatological rainfall to lightning ratio analysis.

I. INTRODUCTION

The lightning detection network of the Catalan Meteorological Service (XDDE) has been providing data from thunderstorms for the past 10 years.

The XDDE started in 2003[1] with 3 SAFIR detectors equipped to analyse frequencies from 1Hz to 300MHz. These detectors use three types of sensors: a five dipole antenna, to detect very high frequency (VHF ranges from 50 to 300 MHz), often used to detect inter- and intracloud (IC) discharges; an electric field antenna, to detect low frequency (LF ranges from 300 Hz to 3 MHz), more appropriate to detect cloud-to-ground (CG) lightning flashes; and a GPS receiver, that provides a very precise measure of the universal time of each entry. By 2009 the network was updated to 4 LS8000 detectors. The new models used a better method, known as TOA (time-of-arrival), to calculate the landing position of CG flashes.

A lightning flash, either IC or CG, is an individual entity that may be formed by several strokes[2],[3],[4]. Those individual strokes are registered by lightning detection networks such as the XDDE and then grouped together into lightning flashes. The number of strokes that form a lightning flash is called the flash multiplicity.

Two field campaigns[5] (2004 and 2005) were launched to determine the detection efficiency of the XDDE. Those campaigns collected experimental data of thunderstorms and compared it with the data collected by the network. The experimental stroke detection efficiency, in evaluated thunderstorms, ranged from 82,9% to 98,7% and mean detection efficiency for 2004 and 2005 were 92,4% and 90,5%, respectively. On the other hand, out of the studied flashes from 2004, only about 30% of them agreed with the experimental multiplicity, while 44% completely disagreed, 9% presented both polarities and the rest were not detected. Similar results were obtained in the 2005 data (35% agreed) and the video recording (only 20%).

Those campaigns determined that, while the stroke data obtained by the XDDE were accurate, the way the network grouped the strokes in flashes was not. That meant that there was a need to improve the current clustering criteria.

The objective of this paper is to study different grouping methods of strokes into flashes.

II. METHODOLOGY

In this section we will describe the methods used to group the stroke data from the XDDE.

A. The XDDE data

The data processed was the one obtained by the XDDE during the years 2010, 2011 and 2012. It was split in files by months, each file having a different size, since the number of flashes varies depending on the seasons; about 50% of the yearly flashes occur between July and August.

The data was already modified, meaning that all the intracloud discharges were already erased form the database. The rest of strokes were already classified between the first return-stroke from a multiple flash (or single-stroke flashes) and the return-strokes that came after the first one. This grouping was done with a commercial software developed by the XDDE manufacturer and no specific details on the grouping method are available. Each stroke had four characteristics associated:

Time: The time when the CG flash hit the ground. It was split in different entries, from months to nanoseconds.

Position: Latitude and longitude of the calculated location (in decimal degrees).

Current: Estimated peak current (in kilo Amperes).

Stroke: Numeral of the position that the stroke holds in the flash.

B. Grouping parameters

To determine if a set of strokes belong to the same flash certain parameters are used to compare them.

Flash Time Window: Maximum time that a lighting flash can last. The XDDE system uses 1 second.

Flash Spatial Radius: Not all the strokes in a CG flash fall in the same spot. This parameter marks the maximum distance between two impact points. The XDDE uses 10 km.

Flash Inter Stroke Time Window: Maximum time between two consecutive strokes. The XDDE states that two consecutive strokes can be separated a maximum of 500 ms.

Maximum Multiplicity per flash: Number of strokes that form a flash. The XDDE detected a maximum of 24 strokes for one lightning flash.

Polarity: Negative CG flashes are the most common occurrence, 80% against 20% of positive polarity. Flashes with both polarities or positive flashes with more than one return-stroke are very rare events according to literature[6]. That means that the 9% of bipolar flashes from the XDDE must be an error, probably caused by interferences with the electric field antenna or IC discharges registered as CG flashes.

Experimental data that renders these parameters obsolete already exist. Flashes that lasted 2 seconds or strokes from the same flash that have fallen 16 km apart. Even if they are not common events it would be a mistake not to take them into account.

The new clustering program will use the following parameters:

Flash Time Window: It will range from 0.6 to 2 seconds.

Flash Spatial Radius: It will range from 6 km to 20 km.

Flash Inter Stroke Time Window: This is the only original parameter that has not been proven wrong, so it will not be modified.

Maximum Multiplicity per flash: It will be completely eliminated, the other parameters are enough to limit each CG flash.

Polarity: Since most of the positive strokes seem to be mistakes, it was decided to erase all positive strokes.

C. Hierarchical clustering

Hierarchical clustering[7] separates data into groups whose identities are not known in advance.

This procedure begins by considering that the data set consists of n groups containing one observation each. The first step is to find the two groups that are closest and combine them into a new cluster. It follows by finding again the two groups that are closest and merge them in a larger one. The process is repeated until the last two groups are merged together and all the data set is contained in one group.

The clusters dendrogram is a tree diagram that records all the clustering process. It records every merging and the distance both groups were. An example of a dendrogram can be seen in Fig.(1). Neither the first nor the last steps of the clustering are of any use to us. But, using the dendrogram, the clustering can be cut at any point, by the number of groups wanted or by the maximum radius of those groups.

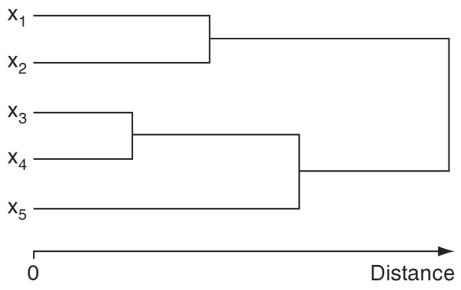

FIG. 1: Example of a dendrogram. The points 3 and 4, being the closest together, are the first to merge. And, as the distance increases, the groups start pairing with the ones they have closer.

In this case, the groups have to be cut by a given distance or time, so a 3-dimensional clustering is set, using latitude, longitude and universal time as variables. But another problem arises, since time and space have different dimensions, no single parameter can be used to analyse both of them at the same time. This is why a more general alternative is used. A radius approach called the Karl Pearson distance, which is a weighted version of the Euclidean distance:

$$d_{i,j} = \left[\sum_{k=1}^{K} w_k (x_{j,k} - x_{i,k})^2 \right]^{1/2} . \tag{1}$$

There are different weights for each variable to warrant that the distance is equivalent in any direction. In this case it can be thought as a normalization of the variables:

$$w_k = \frac{1}{P_k^2} \tag{2}$$

where P_k can be the *flash time window* or the *flash spatial radius*, depending if working with time or distance, making the cluster limiting distance, to belong to a CG flash, the same in any direction.

There is only one possible distance between two points but things change when operating with a group of points, picking which distance will be used to decide the merging is called *linkage*. The last step is to decide which linkage works better for lightning flash clustering.

The *flash time window* and the *flash spatial radius* are absolute limits. Meaning, that if a stroke is paired with another at the maximum distance, it can not also be paired with a third stroke at the maximum distance in another opposite direction, because the two extremes would be surpassing the maximum distance.

The *complete-linkage* works by always picking the maximum distance between points from different groups.

Taking two groups G_1 and G_2 the distance using the *complete-linkage* would be

$$d_{G_1,G_2} = max(d_{i,j}) \tag{3}$$

where $i \in G_1$ and $j \in G_2$.

D. Problem solving

The first problem is that the cluster analysis does not include the *Flash Inter Stroke Time Window* parameter. The solution was to examine every stroke cluster coming out, and cut those separated by more than 0.5 s into two different lightning flashes.

The second, and more important problem, is that the clustering radius does not cover all the strokes belonging to the same flash. The weighted Euclidean distance method allows us to group strokes using both, the distance and the time parameters at the same time. But it is not a perfect method, because it links the two parameters, making it impossible for a flash to reach the maximum in both aspects at the same time. As can be seen in Fig.(2), clustering with a $(R_S = (w_k)^{1/2}P_k)$ radius will allow a lightning flash to last P_t seconds only if all strokes fall in the same spot. And, the farther two strokes fall, the shorter the clusters flash time window will be. It can also be seen that the point in (P_d,P_t) is out of the usual clustering radius R_S. This can be solved by using a new cluster radius:

$$R_L = \sqrt{w_t(P_t)^2 + w_d(P_d)^2} = \sqrt{2R_S^2} = \sqrt{2}. \tag{4}$$

This radius is the distance between two strokes placed the farthest possible from each other, maximum distance (P_d) and time (P_t). R_L is also depicted in Fig.(2) and it shows that it is not a perfect solution. This new radius accounts for those strokes separated by long distances and times at the same time, but it can also add some strokes that were not supposed to enter. If all the strokes fall in the same spot the new clustering can create lightning flashes that last longer than P_t and, for short flashes, it reaches distances longer than P_d.

E. Step-by-step method

Unfortunately there was not enough experimental data on thunderstorms for 2010-2012 to evaluate the clustering method results. Instead, a new method named step-by-step was created to be compared with the clustering method.

The step-by-step program does the same calculations, to group strokes into CG flashes, without relying on any clustering functions. It compares every stroke with the rest of the archive, one by one, trying to find any other strokes that met the conditions to belong to the same flash. The results produced by the step-by-step method will be useful to determine the accuracy of the clustering methods.

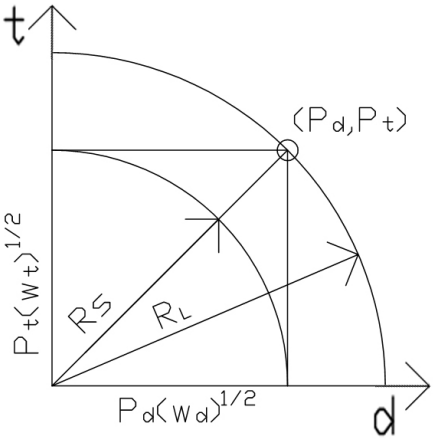

FIG. 2: Effect of the Karl Pearson distance ($R_S = (w_k)^{1/2}P_k$) used as a radius in space time clustering. Also effect of the proposed new clustering radius R_L. The **x** coordinate represents the **distance between strokes**. It is obtained using the longitude and latitude parameters. The **y** coordinate represents the **time between strokes**. Both directions are equal because they have been weighted with w_k.

III. RESULTS

In this section four different methods of stroke clustering are used. The original method from the XDDE, hierarchical clustering using both R_S and R_L and the step-by-step method.

A. Comparing methods

Both Fig.(3) and Fig.(4) compare the original groups from the XDDE with the ones obtained from the different developed methods, using a *flash time window* of 1 s and a *Flash Spatial Radius* of 10 km, to maintain the original values.

Fig.(3) and Fig.(4) show the expected comparative. While the three developed methods give us similar results, the original XDDE data is quite different. One of the reasons may be those large bipolar flashes that grouped more strokes together than the actual number.

The number of lightning produced by the step-by-step method is often between the two values obtained with the clustering methods. The short radius clustering cuts off those strokes that have values near the clustering parameters, creating more flashes, while the clustering with the longer radius groups more strokes than the ideal value, creating flashes with higher multiplicity.

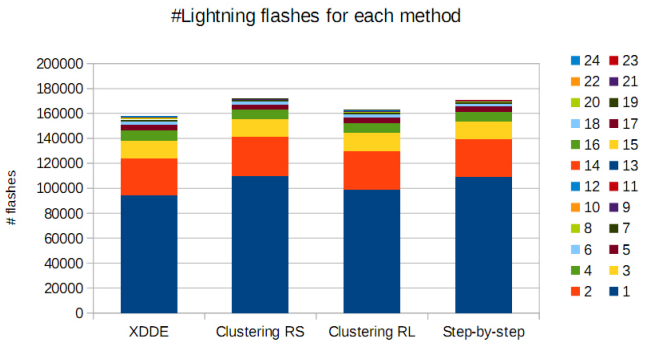

FIG. 3: Number of lightning flashes, separated by their multiplicity, accounted using several methods (using 1 s and 10 km).

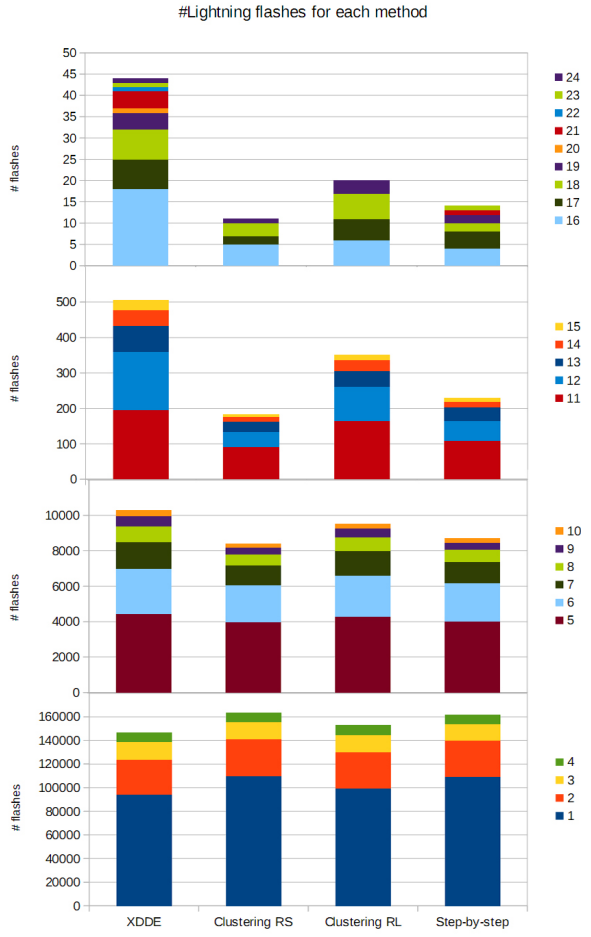

FIG. 4: Number of lightning flashes accounted using several methods (using 1 s and 10 km). Unlike Fig.(3) this plot separates the different multiplicities into sections to compare them separately.

This is why the calculations were made using both clustering radius, and, even if the results are a bit different, their behaviour is the same.

B. Clustering results

The parameter used to represent the results is the average flash multiplicity.

The flash multiplicity for a single cloud-to-ground flash is simply the number of return-strokes. However the average multiplicity for a group of CG flashes gives us certain information about its structure. If there are a lot of single-stroke lightning flashes the average multiplicity will drop, but if strokes are merged in big groups the multiplicity values will rise.

If multiplicity for one cloud-to ground flash is:

$$m = n_s$$

where n_s is the number of return-strokes of that lightning flash, then average multiplicity for N flashes will simply be:

$$M = \frac{1}{N}\sum_{k=1}^{N} m_k = \frac{1}{N}\sum_{k=1}^{N} n_{sk} = \frac{N_s}{N}. \qquad (5)$$

This means that the average flash multiplicity can be calculated with just the number of strokes and the final number of flashes for each set of parameters.

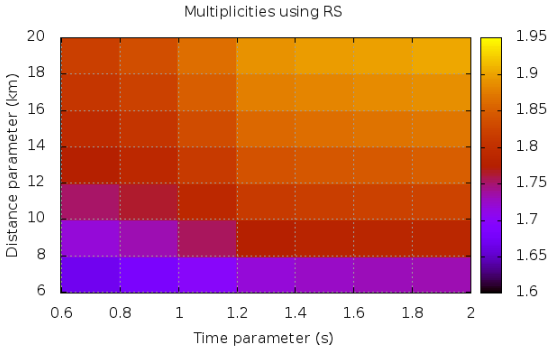

FIG. 5: Average flash multiplicity for each pair of parameters The calculations were made using the short radius R_S.

As can be seen in both Fig.(5) and Fig.(6) the average flash multiplicity grows with the clustering parameters. This responds to logic since, for a longer radius, more strokes will be allowed in a multiple stroke flash and multiplicity will increase. The same logic can also be applied comparing the two figures. R_L allows more strokes to come together, producing a bigger multiplicity for the same set of parameters than R_S.

Another important detail of the figures is that they are not symmetric, multiplicity grows faster vertically than horizontally. That means that the *flash spatial radius* has more impact on the CG flash building than the *flash time window*. And that flashes with impact points far from each other are more common than flashes that last long.

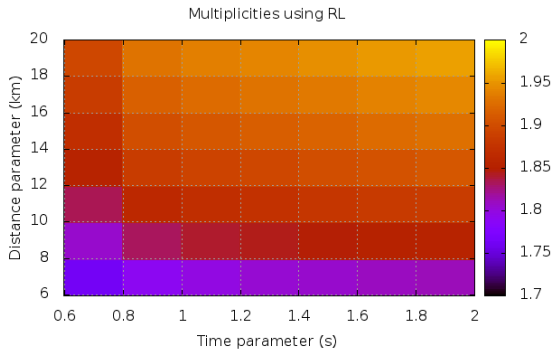

FIG. 6: Average flash multiplicity for each pair of parameters. The calculations were made using the long clustering radius R_L.

IV. CONCLUSIONS

- Cluster analysis has proven to be a reasonably accurate method to group strokes into lightning flashes. From the two radius R_S and R_L the first one provides with a better set of results for this range of parameters, because they get closer to the ones yielded from the step-by-step method when analysing big amounts of data. For three year data analysed (using 1 s and 10 km), there were differences with only 2% of strokes for R_S and 7% for R_L.

  Without experimental data to contrast with the results the better method can not be identified. But those programs will work with any future data from the XDDE. And, if there is a new field campaign to collect experimental data, it will help finding out which method provides the better results.

- Another example of the accuracy of this program are the multiplicity results. The range of average multiplicities from 1.6 to 2 was within the expected results. The similar results also indicate that clustering works fine with any combination of parameters.

  Until there is experimental testing, a specific combination can not be chosen. Probably even both radius could provide clustering results close to reality using different parameters.

- These results can help with future lightning related studies. For example the studies relating the number of lightning with precipitation[8] or in individual thunderstorm case studies[9]. It could also give more concrete directions to prevent lightning related accidents. Reducing the *flash spatial radius* to one or two kilometres, while keeping the *flash time window the same*, could give information not only of the amount of cloud-to-ground flashes but also of the amount of impact points.

Acknowledgments

The lightning data were obtained from the *Servei Meteorològic de Catalunya*. The picture shown in Fig.(1) was adapted from D. S. Wilks, (2006). I would like to thank my advisors Joan Bech (UB) and Nicolau Pineda (SMC) for guiding me during this project and providing me with research material. I also thank my parents for supporting me and for the valuable suggestions that improved this project.

[1] N. Pineda, X. Soler and E. Vilaclara, *Aproximació a la climatologia de llamps a Catalunya: anàlisi de les dades de l'SMC per al període 2004-2008*, Nota d'estudi del Servei Meteorolgic de Catalunya **73** (2011).

[2] J.M. Wallace and P.V. Hobbs, *Atmospheric Science: An Introductory Survey*, 2nd ed. (2006).

[3] R.E. Orville, G.R. Huffines, W.R. Burrows, R.L. Holle and K.L. Cummins, *The North American Lightning Detection Network (NALDN). First Results: 1998 - 2000*, Monthly Weather Review **130**: 2098- 2109 (2002).

[4] O. Pinto Jr., I.R.C.A. Pinto and K.P. Naccarato, *Maximum cloud-to-ground lightning flash densities observed by lightning location systems in the tropical region: A review*, Atmospheric Research **84**: 189 200 (2007).

[5] J. Montanyà, N. Pineda, V. March, A. Illa, D. Romero and G. Solà, *Experimental evaluation of the Catalan Lightning Detection Network*, 19th International Lightning Detection Conference (2006).

[6] V.A. Rakov and G.R. Huffines, *Return-Stroke Multiplicity of Negative Cloud-to-Ground Lightning Flashes*, Journal of applied meteorology **42**: 1455- 1462 (2003).

[7] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 2nd. ed. (2006).

[8] N. Pineda, T. Rigo, J. Bech and X. Soler, *Lightning and precipitation relationship in summer thunderstorms: Case studies in the North Western Mediterranean region*, Atmospheric Research **85**: 159 170 (2007).

[9] J. Bech, N.Pineda, T. Rigo and M. Aran, *Remote sensing analysis of a Mediterranean thundersnow and low-altitude heavy snowfall event*, Atmospheric Research **123**: 305 322 (2013)