



HAL
open science

Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles

Emmanuel Biabiany, Didier Bernard, Vincent Pagé, Helene Paugam-Moisy

► To cite this version:

Emmanuel Biabiany, Didier Bernard, Vincent Pagé, Helene Paugam-Moisy. Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles. *Computers & Geosciences*, 2020, 145, pp.104612. 10.1016/j.cageo.2020.104612 . hal-03058865

HAL Id: hal-03058865

<https://hal.science/hal-03058865v1>

Submitted on 17 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

Emmanuel Biabiany^{a,b,*}, Didier C. Bernard^{a,**}, Vincent Page^{b,2} and H el ene Paugam-Moisy^{b,3}

^aLaboratoire de Recherche en G eosciences et  nergie, Universit  des Antilles, Guadeloupe

^bLaboratoire de Math matique, Informatique et Applications, Universit  des Antilles, Guadeloupe

ARTICLE INFO

Keywords:

Machine Learning
Data processing
Clustering
Image Processing
Tropical Rainfall
Lesser Antilles

ABSTRACT

To expand our knowledge of the climate in the Lesser Antilles, we attempted to identify the spatio-temporal configurations of daily weather. We noticed certain pitfalls that can lead to poor results when using clustering algorithms and have proposed some steps towards the solution. These advancements might prove interesting for climate informatics, as well as for many applications that cluster physical fields. We illustrated the pitfalls with a dataset of cumulative rainfall from NASA's Tropical Rainfall Measuring Mission for the period 2000 to 2014. First, the pitfall is the lack of numerical evaluation of the clusters found by the algorithms, which prevents the comparison of algorithms. We used silhouette index for this evaluation and to demonstrate other problems. Second, algorithms like K-means cluster the points around their barycentre. For many physical fields, this barycentre is trivial, which may lead to poor performances. Third, the L2 norm used in conventional clustering methods, such as K-means and hierarchical agglomerative clustering, focus on the exact location of fields, which leads to poor evaluations of similarity between fields. We replaced it by a similarity measure called the expert distance (ED) that compares the histograms of four zones, based on the symmetrised Kullback–Leibler divergence. It integrates the properties of the observed physical parameter and climate knowledge. With these improvements, the results revealed five clusters with high indexes. The algorithms now discriminate the daily scenarios favourably, thereby providing more physical meaning to the resulting clusters. The interpretation of these clusters as weather types is discussed.

1. Introduction

The automatic identification of spatio-temporal structures that are indicative of the atmosphere's low frequency variability has long been known as an important research objective (Michelangeli et al. (1995); Burlando (2009); Kaufman and Roussew (1990); Ayrault et al. (1995); Ghil and Robertson (2002); Jury and Malmgren (2012)). Called weather types (WTs), these structures can be recurrent, nearly stationary (Michelangeli et al. (1995); Brunet and Vautard (1996)), and have their own identity (Vautard (1990)). The approach used in all these studies (except for Jury and Malmgren (2012) that will be discussed later), is as follows: for a selected physical parameter, a day is characterised by a spatial field that can be visualised as a map. The days that constitute the database are processed by a clustering method that creates clusters of similar days (Tian et al. (2014); Hadzimejlic et al. (2013); Monteleoni et al. (2013)). The average (named centroid) of each cluster of days is considered to be a WT. The two methods of clustering commonly

*Corresponding author

**Principal corresponding author

emmanuel.biabiany@univ-antilles.fr (E. Biabiany); didier.bernard@univ-antilles.fr (D.C. Bernard);

vincent.page@univ-antilles.fr (V. Page); helene.paugam-moisy@univ-antilles.fr (H. Paugam-Moisy)

ORCID(s): 0000-0003-2934-5291 (E. Biabiany); 0000-0001-7531-3528 (D.C. Bernard)

¹Developement of main algorithm implementation, manuscript writing and reviewing.

²Conceptualization of the work, supervisory guidance, manuscript writing and reviewing.

³Supervisory guidance and manuscript reviewing.

39 used to process climate clustering data are the K-Means (KMS) and Hierarchical Agglomerative Clustering (HAC)
40 methods (Tian et al. (2014)).

41 All these articles share another common point, namely, the absence of a numeric evaluation of the clusters' quality.
42 The evaluation of the clusters' quality can be carried out with a multiplicity of considerations, ranging from the physical
43 plausibility of the centroids of the resulting clusters, intra class and extra class variance, and analysis of the temporal
44 distribution of the clusters' days to deduce the composition of the seasons in terms of WTs (Chadee and Clarke (2015)).
45 Some authors (Moron et al. (2016)) chose to establish correlations with other physical parameters to evaluate the
46 reliability of these clusters. All these considerations make sense when it comes to the physical analysis of the climate,
47 but such multiplicity hinders the comparisons among methods.

48 In contrast, we resort mainly to a single performance measure in this study (namely, the silhouette index, Rousseeuw
49 (1987)) and reveal the major flaws with traditional clustering methods. Specifically, our results indicate that they do
50 not produce relevant clusters. To confirm this assertion, we illustrated that these clusters are characterised by a low
51 homogeneity (producing clusters that contain very distinct situations) and low separation (producing clusters faintly
52 distinct from each other). When this point has been established, we analyse the origin of the problems with these
53 methods, which have two main causes.

54 First, the use of L2 as a similarity measure between two spatial fields raises questions. We show that the distance
55 L2, used to compare two days, does not show the following basic characteristic: the proximity of two days in the sense
56 of the retained distance must be equivalent to the proximity between them from a physical point of view. To overcome
57 this, we propose a pseudo-distance to cluster spatio-temporal rainfall fields. We provide a physical explanation for the
58 principles that led to the definition of this pseudo-distance and show how incorporating it will enable further realistic
59 results.

60 Second, the use of a field average of a cluster days (its centroid) as the main representative of a cluster (Jain (2008);
61 Gokila et al. (2016)) also raises questions. The rainfall dataset used in this study has many spatial discontinuities that
62 can probably explain the poor results obtained by traditional methods. All the previous researchers used algorithms
63 like KMS that aggregate days around centroids on more continuous datasets. The majority of the studies describe the
64 final centroids of the clusters as a WT. However, for the type of data used in this study, this is trivial because this
65 average may not be the representative of an existing physical scenario. The pseudo-distance that we propose improves
66 this phenomenon when clustering.

67
68 To describe the resulting clusters, we present the most representative days of each cluster (not an average), which
69 seem to be more relevant endpoint. These clusters are related to the WTs and circulation types (CTs) presented in the
70 previous study. We consider their intra- and interannual dynamics to understand their links with the seasons and the

71 interseasons. A specific analysis by island allowed us to understand the impact of these rainfall patterns on the Lesser
72 Antilles, and we also compared satellite observation data (used for clustering) with data from ground rainfall stations.

73
74 The data used and the methodology developed are presented in section 2. Section 3 provides an evaluation of the
75 results obtained for the study and a discussion of the important aspects of the results. It is divided into the following
76 two parts: at first, we present a numeric comparison of methods, discuss our choices of parameters, and evaluate the
77 quality of the resulting clusters to show that our method gives significantly better results than traditional methods;
78 then, we provide a physical analysis by relating the clusters produced by this method to the existing knowledge on
79 climatology in the Caribbean region. The last section is dedicated to the conclusions and perspectives.

80 **2. Material and methods**

81 **2.1. Datasets**

82 The study data consist of rainfall measured by the satellites of the Tropical Rainfall Measuring Mission (TRMM)
83 Huffman et al. (2007). These data were spatialised with a grid of 0.25° longitude \times 0.25° latitude. The geographic
84 zone ranges from -66.25 to $-20.25^\circ E$ and from 5 to $30^\circ N$. It includes the Lesser Antilles islands along with the
85 northeastern part of South America and a part of the Central Atlantic Ocean with the Cape Verde archipelago (cf Fig
86 1). Each day was thus represented by a field of 101×187 values, which were transformed into a vector of 18,887
87 components. The data covered the period from 2000 to 2014, representing a base of 5,415 days.

88 To determine recurring situations from the rainfall field, we used conventional clustering methods such as the HAC
89 or KMS (Monteleoni et al. (2013); Rokach and Maimom (2010); Parmar and Saket (2017)). Notably, in the climate
90 datasets, a field is a vector in a space with very large dimensions (18,887 pixels in our case), which made the search
91 for relevant information relatively complex. To assess the study area, surface rainfall data supplied by Meteo France
92 (Guadeloupe and Martinique) from 1979 to 2014 were used in the design of the expert distance (ED) (Section 2.4). To
93 complete the analysis and interpretation of the clustering results, atmospheric circulation data at 850hPa from ERA-5
94 collected by radiosondes from *Wyoming Weather* from 1979 to 2014, and tropical storm and hurricane tracking data
95 from 2000 to 2014 indexed on *Unisys Weather* were used (Section 3).

96 **2.2. Clustering performance measures**

97 As pointed out in Section 2.4, current publications in the field lack an important element, namely, they do not
98 provide reliable, quantitative performance measures that indicate the quality of the resulting clusters. For example,
99 in the article by Chadee and Clarke (2015), only the centroids of the clusters were evaluated and the clusters were
100 considered to be correct because the centres were plausible and physically interpretable. However, the dispersion of

101 the clusters was not evaluated.

102 The absence of such quantitative quality measurements is particularly problematic when it comes in comparing the
103 clustering methods. It is necessary to implement a way to measure the quality of the clusters, even if quality is recog-
104 nised in the literature as a challenging problem, primarily in the cases where it is difficult to form hypotheses on the
105 shape of the clusters searched for Tibshirani et al. (2001); Halkidi et al. (2001); Lallich and Lenca (2015).

106 In the absence of ground truth information, here is a list of some internal quality measurements that can be used with
107 the clusters: the silhouette index (Rousseeuw (1987)), Calinski–Harabasz index (Calinski and Harabasz (1974)), and
108 Davies–Bouldin index (Davies and Bouldin (1979)). All of them combine the following two essential qualities expected
109 for clusters representing different physical situations: compactness (homogeneity) as well as distance (separation). We
110 selected the silhouette index because it has the following benefits:

- 111 • It provides a quality measure for each day within its cluster, for each cluster, and for the method through all its
112 resulting clusters,
- 113 • It is easy to interpret in terms of the pertinence of clusters.

114 Let us assume a clustering algorithm \mathcal{M}_k that divides a dataset into k clusters. At the end of the clustering process,
115 each point i of the data is assigned to a specific cluster C_i . The silhouette index $S(i)$ for each i is obtained as follows:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (1)$$

116 where $a(i)$ is the average of the distances between i and the other elements in its cluster and $b(i)$ is the minimum of
117 the averages of the computed distances between i and the elements of every other cluster. In equations 2 and 3, $d(i, j)$
118 expresses the distance used to compare two elements i and j , as shown below:

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j), \quad (2)$$

119 and

$$b_i = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j), \quad (3)$$

120 where C_i is the cluster of i and $|C_i|$ the size of C_i . By default, this distance $d(i, j)$ is the L2 norm. In this work, we

121 will replace the L2 distance with ED (Equation 9). Two other criteria may be estimated. The first, which indicates the
 122 average quality of a cluster C_i , can be calculated by using the following formula:

$$Sc(C_i) = \frac{1}{|C_i|} \sum_{j \in C_i} S(j). \quad (4)$$

123 In order to evaluate all of the clusters obtained by applying a clustering method \mathcal{M}_k to our data, we can consider the
 124 average of each cluster's coefficients, namely:

$$Sa(\mathcal{M}_k) = \frac{1}{k} \sum_{i=1}^k Sc(C_i). \quad (5)$$

125 By definition, each coefficient is between -1 and 1 . The coefficient reflects the combined evaluation of the proximity
 126 of an element to the elements of its cluster and the distance of this element to all the other clusters.

127 $Sa(\mathcal{M}_k)$ provides an indication of the quality of the result of a clustering method. For more information, readers can
 128 refer to Rousseeuw (1987). Here are some commonly accepted reference values that we will use later:

- 129 • values greater than 0.20 indicate good performance (existence of relevant and well separated clusters),
- 130 • values less than 0.10 indicate the opposite,
- 131 • negative values indicate that many points are assigned to clusters that do not represent the best possible choice.

132 2.3. Problems related to the L2 distance

133 All the work we were able to review in this domain uses the same distance measurement to compare two fields,
 134 namely, the distance associated with the L2 norm. For the record, this distance between two vectors of daily data
 135 $D_1(v_1, v_2, \dots, v_n)$ and $D_2(v_1, v_2, \dots, v_n)$ is calculated as follows:

$$d_{L2}(D_1, D_2) = \sqrt{\sum_{i=1}^n (D_1(v_i) - D_2(v_i))^2}, \quad (6)$$

136 where $D_x(v_i)$ is the i -th value v of the vector of daily data D_x and n is the length of vectors of daily data. This
 137 distance seems to be partly responsible for the difficulties encountered by the clustering methods in the field of climate
 138 informatics Monteleoni et al. (2013). Figure 2(a) shows this property schematically, in which the distance L2 found
 139 between the reference (in black) and the two fields is the same. In addition, for fields such as rainfall, which show

140 extensive spatial and temporal irregularities, a day D_1 with a low spatial disparity relative to the reference (Fig 2(b))
141 will have a norm L2 equivalent to a day D_2 , red curve, with a bigger spatial disparity. Thirdly, when data are described
142 in a large vector space, a multitude of small fluctuations that are spatially spread across the field can be considered as
143 important as one single big and very localised fluctuation.

144 **2.4. Design of the expert distance**

145 In this section, as with other authors in different application fields before (Pandit and Gupta (2011); Parmar and
146 Saket (2017); Gibbs and Su (2002); Shraddha and Suchita (2011)), we propose an alternative similarity measure based
147 on the fact that the correction, even slight, of L2's biggest weaknesses should result in significant improvements. The
148 construction of the proposed measure will reduce the influence of the spatial location by a suitable subdivision and a
149 quantification.

150 **2.4.1. Partial management of spatialization**

151 For several computer vision applications, the image can be analysed at the patch level rather than at the individual
152 pixel level (Barnes et al. (2011); Dinu et al. (2012); Amelio and Pizzuti (2016)). Image patches contain contextual
153 information and have advantages in terms of computations and generalisations (Guo and Dyer (2007)). From this
154 point of view, the decomposition of an image into patches or zones that do not overlap provides a simple but effective
155 way of overcoming the curse of dimensionality (Xin (2017)).

156 In this study, we subdivided the rainfall field of the North Atlantic tropical zone into four patches (Fig 1) to limit
157 the computation time and demonstrate the value of the approach without seeking to optimise it. The definition of these
158 zones serves an important purpose, namely, to take the knowledge of field experts into consideration. Thus, each zone
159 corresponds to a specific and known centre of action. In Fig 1, A1 is the zone in which the cold surges originate;
160 A2 is the North Atlantic Subtropical High (NASH) zone; A3 is influenced by the continental zone and the arc of the
161 Lesser Antilles; and A4 is the zone of low pressure that is linked to the presence of the Intertropical Convergence Zone
162 (ITCZ). To compare two fields, we subdivided the fields according to these four zones and then compared each of them
163 with their correspondent.

164 **2.4.2. Comparison of the distribution of intensities**

165 Once each field had been spatially subdivided, we no longer had to pinpoint the exact location of the phenomena
166 in each zone. It seems relatively reasonable to ignore their position down to the exact mesh, and instead we looked at
167 the distribution of the rainfall intensities of the observed field, ignoring the notion of spatial location. In the absence
168 of knowledge of the form of the distribution of rainfall intensities, the Kullback–Leibler divergence appears to be a

169 judicious choice (Kwitt and Uhl (2008); Kullback and Leibler (1951); Walker et al. (2004)). It is expressed as follows:

$$D_{KL}(P, Q) = \sum_c P(c) \log \frac{P(c)}{Q(c)}, \quad (7)$$

170 where P and Q are two distributions of discrete probabilities and c is the index of possible values taken from each
 171 distribution. To obtain a metric for \mathbb{R}_+ that also has the symmetry property, we will use the symmetrised divergence
 172 of Kullback–Leibler, namely:

$$D_{KLS}(P, Q) = D_{KL}(P, Q) + D_{KL}(Q, P). \quad (8)$$

173 Although it is possible to define a measure of the Kullback–Leibler divergence adapted to continuous fields, we pre-
 174 ferred to quantify the data differently because estimating the probability densities of the intensities raises parametri-
 175 sation problems. Moreover, such quantification might help to reduce the effects of the small fluctuations noted in
 176 Section 2.3, even though edge effects around the boundaries of the selected intensity classes remain. The histogram
 177 bins were determined from the rainfall data collected in the area. We selected eight bins of possible intensities. The
 178 boundaries of these bins were selected from the rainfall data collected in the area such that the distribution of the bins
 179 was uniform (cf. Table 1).

180

181 The distinct intensity distributions obtained were then used to compute the Kullback–Leibler divergence in each zone.
 182 The average of the divergences by zone provides the distance between two days. We call this the expert distance, which
 183 is referred to as ED, and the quantity is defined by:

$$ED(D_1, D_2) = \frac{1}{n} \times \sum_{i=1}^n D_{KLS}(D_1 Z_i, D_2 Z_i), \quad (9)$$

184 where D_1 et D_2 are days and $D_x Z_i$ is the histogram of the zone with reference i . In our case, the number of zones n
 185 equals four as presented in Fig 1. All the operations listed above are summarised in Fig 3.

186 3. Results and discussion

187 In this section, we present the results and discuss the results at first from a computer science point of view. The
 188 last subsection will be dedicated to an analysis of the results for the specific application considered.

3.1. Performance measurements of clustering methods

We used algorithms such as HAC and KMS with which we separately linked L2 and ED. To reveal the possibilities of using these two distances, the evaluation of the quality of the clusters obtained by HAC-L2, KMS-L2, HAC-ED, and KMS-ED was performed by computing the silhouette coefficients (Equation 5). Figure 4 shows the evolution of the silhouette coefficient as a function of the number k of clusters.

The results showed several interesting points for the comparison between KMS and HAC, which are summarised below. Notably, the methods based on the HAC algorithm produced coefficients that were significantly lower than those obtained by KMS algorithms. Additionally,

- For HAC-ED, the silhouette coefficient approached 0 for $k > 3$, thus indicating the irrelevance of the clusters detected.
- HAC-L2 exhibited negative silhouette coefficients, thus indicating that the points were affected by suboptimal clusters.
- HAC-ED produced results superior to HAC-L2, although the performance of both remained very weak.

We can now eliminate the HAC algorithms and focus on KMS. This choice is in line with the literature Chadee and Clarke (2015), in which HAC has been used only to define the initial centroids of the KMS algorithm by ruling out any random selection of these centroids. Here,

- KMS-ED largely outperformed KMS-L2.
- KMS-ED exhibited values mainly over 0.2 for a significant number of clusters, thus indicating the presence of relevant structures within the data.
- KMS-L2 exhibited values under 0.1 for $k > 2$, thus indicating irrelevant clusters.
- Although the shape of the curves generally decreased with k , a slight inflection was observed around $k = 5$ in both cases.
- For this value of $k = 5$, KMS-L2 had a silhouette value of 0.08, whereas KMS-ED had a value of 0.26.

The selection of the optimal k was not as clear-cut as we had hoped, in spite of the noted inflection around $k = 5$ in both cases. This led us to choose this number of clusters for the following analyses. When considering large maritime areas with little land surface, such as the Caribbean region (Fig 1), according to authors of earlier works Vigaud and Robertson (2017); Moron et al. (2016); Chadee and Clarke (2015); Sáenz and Durán-Quesada (2015), the number of WTs obtained by using KMS or HAC can range from 7 to 11. Hence, five WTs seems a bit low in regard to the literature.

3.2. Visual inspection of the clusters found

For the present time, we have relied on the silhouette index for our analysis. Let us now show how the values of the silhouette index are relevant in terms of quality of the resulting clusters. Accordingly, we will carry out a detailed inspection of the clusters found by the KMS-L2 and KMS-ED methods with $k = 5$.

3.2.1. Illustration of irrelevant clusters

The silhouette index results for KMS-L2 were indicative of irrelevant clusters (< 0.1). We can conclude that the L2 norm does not evaluate the similarity of spatial patterns of rainfall data properly and this leads to the aggregation of different situations in the same cluster, as illustrated in Fig 5(a). For the same reason, the method separates situations that are indeed very similar into different clusters, as seen on Fig 5(b). We precisely introduced the new similarity measure ED to overcome this flaw.

3.2.2. Illustration of the irrelevance of centroids for L2

As stated before, the centroids have a major role in KMS algorithms, and some author rely on them to describe the Weather Types extracted by their algorithms. The idea behind this is that these averaged fields are representative of the clusters. Fig 6 show why we believe this idea should be discarded when dealing with discontinuous fields such as rainfalls. It shows, for a specific cluster found by KMS-L2, the centroid (Fig 6(a)), the nearest day to the centroid (Fig 6(b)) and a random day of the same cluster (Fig 6(c)).

This figure shows the following interesting aspects. The centroid does not look like any of the members of the cluster considered. We think that computing the average tends to create continuous artificial zones, which gives a bad spatial appreciation of the existing meteorological structures and leads to poor results for the clustering of spatio-temporal rainfall fields using L2. This problem can be partially avoided when using the ED as the centroid is now a distribution of rainfall in each region. These distributions are much more meaningful for describing rainfall situations. For the following analysis, we will consider, as a representative of the cluster, not the centroid but the most representative element of each cluster obtained. This is the element with the shortest distance to each cluster centroid (distance being L2 or ED in their respective cases).

3.2.3. Separation and homogeneity of the clusters

As stated several times in this article, we prefer to rely on a single performance measure than on a multitude of considerations. We need, however, to prove that this performance measure is a good indicator of these considerations. This section is dedicated to the separation and homogeneity of the clusters.

3.2.3.1. Separation

To assess the **separation** of the clusters, Figs 7 and 8 show respectively the representative (see above) rainfall

248 fields of the clusters obtained by the KMS-L2 and KMS-ED methods for $k = 5$. These figures also show, for each
 249 representative field, the wind field of the corresponding day. This wind field has not been used during the clustering.
 250 It will be used in the physical analysis of the weather types extracted by our method, later in this article. Those
 251 representative fields (both rainfall and wind) should be as different as possible.

252 Figure 7 clearly shows that for KMS-L2, clusters C2 and C3 in fact depict situations that are quite similar with
 253 rainfall in the southeast. C1 and C5 are also similar and depict low rainfall across the space and an atmospheric
 254 circulation originating mainly from the northeast. The exact spatial location of these rainfall situations is the only
 255 thing that distinguishes them, thus illustrating precisely the limits of L2.

256 The results obtained for KMS-ED, shown in Fig 8, highlight, conversely, very diverse clusters. Each case shows
 257 centres that are more or less active.

258 3.2.3.2. Homogeneity

259 We also wanted to evaluate the **homogeneity** (the internal variability) of the clusters obtained for each method.
 260 Of these five clusters, for the sake of conciseness, we will present only one cluster per method. For each method, we
 261 selected the cluster whose $Sc(C_i)$ value (Equation 4) was the closest to the overall value of the corresponding method
 262 $Sa(\mathcal{M}_k)$. These were the C2 cluster for the KMS-L2 method and the C4 cluster for the KMS-ED method. The distance
 263 of each of the days composing a cluster to their representative was computed, and elements were sorted in ascending
 264 order. We extracted six quantiles from this ranking, to evaluate how each cluster varied internally. Fig 9 shows the
 265 quantiles obtained for KMS-L2. Fig 10 does the same for KMS-DE.

266 In the case of the L2 distance (Fig 9), the first three fields were relatively similar to the representative element, whereas
 267 the last two were different. In fact, on the central part and the northwestern edge, we found rainfall zones that were
 268 well localised. When grouping the results by the L2 distance, we observed a cluster of situations that were certainly
 269 similar in terms of numbers but very distinct from a physical point of view.

270
 271 In the case of ED (Fig 10), the elements of cluster C4 had a certain physical constancy although they differed slightly
 272 in their visual appearance. This cluster consisted of low rainfall fields.

273 3.2.3.3. Conclusion on the intrinsic quality of clusters

274 The visual inspection confirmed the results observed by using only the silhouette index.

- 275 • KMS-L2 produced clusters of relative relevance, as some clusters representative were similar. Moreover, days
 276 among a cluster could be very different from a physical point of view.
- 277 • KMS-L2 produced more relevant clusters, which were reflective of diverse situations. Among a cluster, days

278 seemed similar.

279 From the point of view of the physical separation of the clusters, the introduction of the ED has clearly enabled
280 the acquisition of clusters with a much higher relevance. In addition to the better results obtained by KMS-ED, this
281 facet tends to reinforce that clusters with a high silhouette index are more relevant when it comes to their physical
282 significance.

283 3.3. Physical analysis

284 As stated in the introduction, we wanted to pinpoint some pitfalls encountered when clustering daily fields in
285 geoscience applications. We attempted to highlight them and present some partial solutions to these problems in the
286 previous sections. This section will focus on the remarkable results obtained by our simple method as soon as these
287 problems are dealt with. Our major goal was to extract typical geospatial structures of meteorological data, focussing
288 only on rainfall. Some authors, like Jury and Malmgren (2012), have used several fields such as the surface temperature,
289 sea-level pressure, and U zonal wind to extract dominant modes with the help of a principal component analysis (PCA).
290 While such a combined analysis of different fields is something we aim to do in the future, our article focusses on only
291 one field (rainfall) to illustrate the general method we are proposing.

292 In our method, we chose to aggregate spatial areas in a controlled way (by using regions) to add as much physical
293 knowledge as we could acquire in the design of the extraction of WTs. An interesting aspect of our study comes despite
294 one of its limitations, namely, we treated each day independently and clustered days by taking no account of the time
295 dependency between them, apart for the one found in the data themselves (this is a common approach).

296 We propose a physical analysis of the clusters obtained by looking at their intra- and interannual evolutions on a
297 large scale. Then, we will continue the study of the meteorological comparability of daily precipitation at the local
298 scale, including, in this case, surface observations.

299 3.3.1. Large scales

300 3.3.1.1. Annual variability and trends

301 It is thus very interesting to observe the temporal distribution of the fields contained in each cluster to highlight
302 possible annual and semi-annual cycles, as in Chadee and Clarke (2015). The frequencies were derived by applying
303 the cluster analysis technique KMS-L2 and KMS-ED to the daily atmospheric precipitation patterns defined over the
304 domain indicated in Fig 1. We also show the results of Mann–Kendall’s nonparametric trend test (Yue et al. (2002)),
305 which was performed on the annual frequencies of each group to test the null hypothesis of no trend at the 95%
306 confidence level (5% significance level). Radiosonde parameters at different levels, such as specific humidity, wind
307 direction, and intensity, were used to characterise the clusters. The days in each of the clusters would thus compare
308 to several configurations in terms of outgoing longwave radiation (OLR), sea surface temperature (SST), and wind

309 circulation. We will try, as far as possible, to link them to the clusters found in previous study (Chadee and Clarke
310 (2015); Moron et al. (2016); Martinez et al. (2019)).

311

312 For KMS-L2, the results are shown in Fig 11. On the left, we show the annual variability of the monthly frequencies
313 of days making up each cluster, whereas the right side shows the inter-year change. Many of them have a similar
314 annual distribution. Moreover, no significant trends over the 15 years of TRMM observations were found. This result
315 confirms that the L2 distance is still not very appropriate for classifying daily rainfall fields, as it weakly reflects the
316 seasonality effects for the TRMM daily rainfall.

317

318 For KMS-ED, Fig 12, the monthly distributions obtained were clearly separated because they better integrated regional
319 physical and climatic knowledge. With this slight improvement, the results revealed five clusters ($k=5$) with high
320 indexes ($Sa(M_k) > 0.2$), which can be characterised as follows :

- 321 • C1, represents 14.8% of the days analysed, with days spread over the whole 12 months, but data remained rather
322 grouped over the first part of the year. The maximum was centred on the months of April and May. This cluster
323 was associated with CTs 1 to 5 found in Chadee and Clarke (2015) and WTs 1 to 3, 7, and 8 found in Moron
324 et al. (2016). There were no significant trends for the annual frequencies.
- 325 • C2 and C3, 12.5 and 10.4% respectively were close to each other because they were localised at the end of
326 the year. However, for C2, the maximum matched with September–October, whereas for C3, this occurred in
327 November. Decreases in the annual frequencies were found for C2 (p -value=0.02 and $r^2=0.439$), whereas there
328 were no significant trends for C3. These clusters were associated with WTs 4,5,6, and 8 in Moron et al. (2016).
- 329 • C4 was the most prevalent type, in which it accounted for 32% of the days. The elements of this one were mainly
330 distributed from December to April with a peak in February. It represents the cluster of the dry season. The
331 annual distribution showed a marked “dip” between July and November (Fig 12). Figure 13 shows observations
332 of the upper atmosphere at station 78897-TFFR Le Raizet. The daily precipitation fields of C4 showed a thin
333 wet layer, which dried out at 800 hPa. In the lower levels, the trade winds were from the northeast, whereas for
334 the levels below 500 hPa, the flow was predominantly from the northwest. In this case, the drivers, which are
335 the NASH and the SST, limited the intensity of the observed rainfall (Davis et al. (1997); Dunion (2011); Moron
336 et al. (2015)). The NASH and the North American High are connected together and provide strong diverging
337 trade winds and subsidence in the Caribbean. The ITZC migrates southwards favouring the development of the
338 southern flank of NASH (Martinez et al. (2019)). A detailed analysis of C4 indicated the presence of heavy
339 precipitation. These conditions were related to cold fronts that managed to reach the low latitudes of the Lesser

Antilles (in January, April, and May). Indeed, at the interface between the two air masses, the trade winds carry a flow of warm, humid equatorial air, which intensifies convection and the development of active frontal cloud bands producing heavy rainfall Beucher (2010). The evolution of the number of days per cluster over the 15 years of TRMM observations showed a positive correlation, meaning an increase of C4 (p -value <0.022 , $r^2=0.342$). This cluster can be associated with CTs 1 to 5 found in Chadee and Clarke (2015) and WTs 1 to 3, 7, and 8 found in Moron et al. (2016).

- C5, centred on July–August, was well separated from the dry season and contained the last third of the daily rainfall analysed. It represents the cluster of the rainy season with the highest contribution to the annual rainfall. For C5, unlike C4, the trade winds were established on a column of atmospheric air, which was higher and wetter (Fig 13). At this time of year, the NASH is moving northeast, which allows the more humid trade winds to flow to the east Caribbean due to the longer course over the Atlantic Ocean on its southeast flank. The Atlantic ITCZ reaches its northern most extent in the Lesser Antilles and begins its southward migration at this time (Martinez et al. (2019)). This period corresponds to the increase in convection with increased moisture advection in the Caribbean, which is favourable to higher daily precipitation (Jury and Malmgren (2012); Herrera and Ault (2017)). There was also a significant positive trend (p -value $=0.028$, $r^2=0.321$) which reflects an increase of C5 between 2000 and 2014 (Fig 12). C5 can be related to CTs 6–7 and WTs 4–6 found in Chadee and Clarke (2015) and Moron et al. (2016) respectively. Conflicting trade winds (shear lines), caused in this period by changes in wind direction and wind speed, passing from the northeast sector to the southeast sector resulted in heavy cloud cover and heavy rainfall. In addition, the close passages of the eastern waves triggered powerful updrafts with large areas of surface convergence favouring heavier rainfall on the eastern Antilles (Beucher (2010)).

Among the five retained clusters, C4 and C5 represent 62% of the sample, and these are the representative clusters of the bi-modal cycle. The remaining third was divided equally between C1, C2, and C3. Among the latter, some of the data may represent temporal transitions between seasons. These transitions may highlight differences between the northern and southern Caribbean. C2, C4, and C5 showed significant trends. The drying trend found for C4 also has been found by other authors over longer periods and prior to these TRMM observations (Jury (2009); Jury and Malmgren (2012); Jury and Bernard (2019)). The decrease in the number of C2 days over the last decade, combined with the increase in the number of C5 days, indicates a potential shift or even a tightening of the rainy season (Fig 12). For TRMM daily rainfall measured during these two periods, the increase in convection at the beginning and end of summer can be attributed to the oceanic and continental monsoon annual cycles (Jury and Malmgren (2012)). The latter has a positive influence on early summer rainfall but negative influence on late summer rainfall.

370 **3.3.1.2. Links to hurricanes season**

371 In our study regions, hurricanes are a major concern for resident populations and thus, it was of interest to analyse the
372 distribution of hurricanes in clusters.

373 Tropical storms (TS) and hurricanes (H) recorded at the National Oceanic and Atmospheric Administration (NOAA)
374 site from 2000 to 2014 helped us to attribute the different proportions of these types of atmospheric hazards to the clus-
375 ters retained in KMS-ED. This work had not been done by Vigaud and Robertson (2017); Moron et al. (2016); Sáenz
376 and Durán-Quesada (2015); Jury (2009); Jury and Malmgren (2012). The proportions are presented in Table 2. C2,
377 C3, and C5 contained almost all of these hazards. The highest number of days with TS or H were found for C5, fol-
378 lowed by clusters C2 and C3. These clusters included around 80% of the H and more than 90% of the TS. In relation
379 to the size of the clusters, $P_{C_x}(H)$ in Table 2 results show a proportion twice as large (13%) of H in C2 as that of C5
380 (5.8%) and C3 (6.9%). C1 and C4 contained few hazards. However, in the previous subsection, these clusters have
381 been identified as dry day clusters, but we did observe the presence of a few TS or H days. This seems to have been
382 due to the fact that the NOAA database covers the central Atlantic Ocean and the whole Caribbean region. These
383 phenomena are at the limit of our domain (Fig 1).

384 **3.3.2. Local scales: Lesser Antilles of the east Caribbean**

385 In this paragraph we focus on the TRMM rainfall observed over the mountainous and flat islands of eastern
386 Caribbean. Magnitude of rainfall across the Caribbean are affected by localized mechanism like orographic lifting
387 or diurnal heating. They can enhance, to a lesser extent, convection to produce very different rainfalls between the
388 windward and leeward parts in these islands (Jury and Bernard (2019)). In Figure 14, we have calculated the mean
389 cumulative rainfall over all the meshes (occupying at least 20%) of the land surface, mean of spatial sum (MSS). The
390 average per pixel of rainfall is represented by the mean of spatial mean (MSM). We also recorded the percentage of
391 zeros included in each case (DWR).

392
393 These quantitative measures regrouped C1 and C4 (dry days), and C2 and C5 (rainy days), while C3 remained the
394 intermediate cluster. The MSM results showed very different orders of magnitude, i.e. 5 to 7.5 mm/day for C2 and
395 C5, 2 to 3.5 mm/day for C3, and 1.04 to 2.07 mm/day for C1 and C4. The MSS values of the islands of Dominica,
396 Guadeloupe, and Martinique were the highest. Mountain ranges of these islands are real imposing barriers to the trade
397 winds, which leads to orographic ascent of the humid air mass and precipitation. C1 and C4 had the highest DWRs on
398 average, respectively between 34 and 60%, thus reflecting the presence of many days without precipitation in these clus-
399 ters. For the flat islands, the values found were generally high but we believe there was bias due to the low surface area.

400

401 The histogram in Fig 15 compares TRMM to weather stations observations in Guadeloupe and Martinique. The
402 TRMM data had overestimated pixels without rain and underestimated high rainfall. These results were found by Jury
403 and Bernard (2019) for the cloudy peaks of this region. Biases reported here (e.g. low precipitation) could be related
404 more to the peculiarities of the micro climate in steep topography than to the performance of model physics and data
405 assimilation. However, it should be noted that rainfall between 8.7 to 16.4 *mm/day* is relatively well measured by the
406 TRMM data.

407
408 Considering that the class 8.7–16.4 *mm/day* represents a moderate rainy day (well detected by TRMM), the Fig 16
409 shows the intra-annual (*y*-axis) frequency distribution for the analysed period (*x*-axis). The intra-annual evolution of
410 this class over time showed good agreement between the local and regional scales. For the dry (C4) and wet (C5)
411 seasons, the moderate rainfall class was almost evenly distributed with low frequencies. In the other three clusters,
412 peaks with higher frequencies appeared but the distribution was more scattered (dispersed). From 2002 to 2011 C1,
413 C2, and C3 showed punctually isolated maximums of moderate rainfall frequencies, which were observed from April
414 to December. C1 seemed to be the cluster of the April–May inter-season, with its dry to rainy transition, and C3
415 appeared to be the cluster of the reverse transition during November–December.

416 **4. Conclusion and perspectives**

417 **4.1. Conclusion**

418 In this study, we used a similarity measure more suitable for the analysis of climate data for clustering tasks based
419 on the Kullback–Leibler measure. This involved the design of a new metric, named the expert distance (ED), in which
420 some knowledge of the meteorological structures derived from observational studies was considered, and we proposed
421 a novel and unique strategy to strengthen the expected physical relevance. The unit data clustered were the daily rainfall
422 data obtained by TRMM measurements on either side of the Lesser Antilles (the Atlantic Ocean and Caribbean Sea),
423 and this was conducted to obtain the clusters of days with similar climate profiles.

424 We applied unsupervised classification methods (KMS or HAC) to discover global trends within the clusters. The
425 analysis we performed by using these two methods encountered the following difficulties. First, as rainfall totals were
426 widely scattered, “near” fields in the sense of L2 were rare; the L2 norm tended to agglomerate fields with a common
427 spatial structure, although they were otherwise different. Additionally, the centroid used by KMS did not always
428 represent a realistic physical situation.

429 Although the L2 norm is frequently used in climate data clustering, the results obtained herein were unsatisfactory,
430 as the L2 standard did not effectively quantify the physical similarity between configurations. To better quantify the
431 similarity of the daily rainfall, a spatial subdivision of fields was used. Once these patches were defined, we imple-

432 mented a pseudo-distance based on a relaxation of the precision of the spatial location of the fields, and we compiled
433 histograms and compared them by using the Kullback–Leibler divergence. Similar concerns have already emerged in
434 the field of retrieval image research, and we developed a subdivision technique suitable to our field corresponding to
435 known precipitation drivers.

436 In addition, we used the silhouette index to evaluate the quality, coherence, and separation distance between the
437 resulting clusters. We completed this analysis with a visual inspection of the clusters observed, which confirmed that
438 expert deviation allows us to obtain improved results than the L2 standard when applied to this type of meteorological
439 data. Five clusters, resulting from our best algorithm, were evaluated by experts.

440 Two main classes accounted for 62% of the daily totals, in almost similar proportions. For these, the monthly
441 distribution and spatio-temporal averages were representative of the two predominant seasons at these latitudes, namely,
442 the dry season and rainy season. These were focussed respectively on the months of February and August–September.
443 Moreover, the interannual evolution revealed a significant increase in the number of constituent elements of these two
444 clusters during the 15 years of measurements. There was a decrease in daily rainfall (drying), which was spread over
445 the first four months of the year. Additionally, the rainy season seemed to be earlier as it moved towards the month of
446 July. The spatio-temporal averages of the daily totals were 5–6 times greater for the mountainous islands.

447 Our study confirms the trends deduced by other bibliographical works. It also highlights certain small differences
448 that have remained unattainable by classical clustering methods. The results obtained led to finer analysis details
449 of the spatio-temporal dynamics of daily precipitation. Importantly, the analysed series was short and these results
450 may evolve in accordance with increases in the observation period and improvements in the technologies used in the
451 observation of rainfall by satellites, while remaining subjacent to the evolution of the climate.

452 The last third of the daily totals was rather representative of the transitions between the two main seasons. Two of
453 them showed a more spread out intra-annual distribution of dry days, i.e. 14.8% of the total, followed by an average
454 decrease in very rainy days in September–October. These results confirmed the drying trend in this region. With
455 variable positioning and monthly durations, transitions were generally very difficult to identify. Only one of the clusters,
456 10.4% of the total, was representative of the transition from the rainy to dry season.

457 The new metric used represents a concrete step forward in the analysis of possible climatic and seasonal trends
458 based on spatio-temporal data of daily rainfall. It offers good potential for identifying the dynamics of transitions,
459 which often herald the possible severity of the two main seasons.

460
461 This study was based on a collaboration between physicists and computer scientists interested in the methods of data
462 mining from large databases obtained either from satellite observations or computed as in the case of meteorological
463 reanalyses.

464 The extracted clusters, from a climate physicist point of view, were characterised by relevance, if not originality.
465 However, we believe that this method adds more finesse and accuracy in the analysis of the fields for a reasonable
466 number of clusters.

467 **4.2. Perspectives**

468 We presented an approach that should be widely relevant to the present scope of Climate Informatics, as our
469 observations remain valid for any application that uses the clustering of fields. However, as our goal was the extraction
470 of weather types, we focus on these perspectives in future.

471 At first, we could do the exact same study for other single parameters, such as the wind or temperature, with the
472 benefit that we can knowingly compare the quality of the clusters found for each parameter, utilizing silhouette index.
473 Hence, it is possible to select the parameters that give better clusters and probably show that some parameters are
474 irrelevant in the study of weather types in the Lesser Antilles.

475 Interestingly, we might, with small adjustments, combine parameters to extract clusters of days by relying on a
476 diverse and more complete information. We can evaluate such data with a couple or several parameters to obtain
477 potentially better clusters.

478 Furthermore, some parameters (for example, wind) might explain a phenomenon (for example, rainfall). Such an
479 analysis is now possible by clustering the days according to a parameter (wind) and evaluating the quality of the cluster
480 according to the explained phenomenon (rainfall). Employing silhouette index, we should be able to select the best
481 parameter for a given phenomenon.

482 **Computer code availability**

483 The algorithm presented in this work and developed by Emmanuel Biabiany, is called "dePrecitTRMM", imple-
484 mented in GNU Octave (an alternative open-source of Matlab), and hosted at [https://github.com/ebiabiany/expert-](https://github.com/ebiabiany/expert-distance-precipitation)
485 [distance-precipitation](https://github.com/ebiabiany/expert-distance-precipitation) with all the dependencies. Please note that the whole computation (including data pre-formatting,
486 clustering and silhouette index calculation) takes about 48 hours, on a standard computer with an 8-core processor and
487 32 GB RAM.

488 **Acknowledgements**

489 This research was supported by the Claude Emmanuel Blandin Foundation's research and FEDER&FSE 2014-2020
490 project intitled Changement Climatique et Conséquences sur les Antilles Françaises (<https://c3af.univ-montp3.fr/>).

491 **Data availability**

492 Satellite measured precipitation daily products are available on NASA's TRMM project website ([https://pmm.nasa](https://pmm.nasa.gov/data-access/downloads/trmm)
 493 [.gov/data-access/downloads/trmm](https://pmm.nasa.gov/data-access/downloads/trmm)). ERA5 hourly data on pressure levels for wind components are available on Copernicus Climate webservice (<https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels>).

495 **References**

- 496 Amelio, A., Pizzuti, C., 2016. A patch-based measure for image dissimilarity. *Neurocomputing* 171, 362–378. doi:[https://doi.org/10.1016/](https://doi.org/10.1016/j.neucom.2015.06.044)
 497 [j.neucom.2015.06.044](https://doi.org/10.1016/j.neucom.2015.06.044).
- 498 Ayrault, F., Lalauette, F., Joly, A., Loo, C., 1995. North atlantic ultra high frequency variability. *Tellus A* 47, 671–696. doi:10.1034/j.1600-0870.1995.00112.x.
- 499 Barnes, C., Goldman, D.B., Shechtman, E., Finkelstein, A., 2011. The patchmatch randomized matching algorithm for image manipulation. *Commun. ACM* 54, 103–110. doi:10.1145/2018396.2018421.
- 500 Beucher, F., 2010. *Météorologie tropicale: des alizés au cyclone*. Number vol. 2 in *Cours et manuels - Direction de la météorologie, La Documentation Française*. URL: <https://books.google.gp/books?id=PNhDYgEACAAJ>.
- 501 Brunet, G., Vautard, R., 1996. Empirical normal modes versus empirical orthogonal functions for statistical prediction. *Journal of the Atmospheric Sciences* 53, 3468–3489. doi:10.1175/1520-0469(1996)053<3468:ENMVED>2.0.CO;2.
- 502 Burlando, M., 2009. The synoptic-scale surface wind climate regimes of the mediterranean sea according to the cluster analysis of era-40 wind fields. *Theoretical and Applied Climatology* 96, 69–83. doi:10.1007/s00704-008-0033-5.
- 503 Caliski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1–27.
- 504 Chadee, X.T., Clarke, R.M., 2015. Daily near-surface large-scale atmospheric circulation patterns over the wider caribbean. *Climate Dynamics* 44, 2927–2946. doi:10.1007/s00382-015-2621-2.
- 505 Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 224227.
- 506 Davis, R.E., Hayden, B.P., Gay, D.A., Phillips, W.L., Jones, G.V., 1997. The north atlantic subtropical anticyclone. *Journal of Climate* 10, 728–744. doi:10.1175/1520-0442.
- 507 Dinu, L.P., Ionescu, R.T., Popescu, M., 2012. Local patch dissimilarity for images, in: Huang, T., Zeng, Z., Li, C., Leung, C.S. (Eds.), *Neural Information Processing, Springer Berlin Heidelberg, Berlin, Heidelberg*. pp. 117–126.
- 508 Dumion, J.P., 2011. Rewriting the climatology of the tropical north atlantic and caribbean sea atmosphere. *Journal of Climate* 24, 893–908. doi:10.1175/2010JCLI3496.1.
- 509 Ghil, M., Robertson, A.W., 2002. “waves” vs. “particles” in the atmosphere’s phase space: A pathway to long-range forecasting? *Proceedings of the National Academy of Sciences* 99, 2493–2500. doi:10.1073/pnas.012580899.
- 510 Gibbs, A.L., Su, F.E., 2002. On choosing and bounding probability metrics. *International Statistical Review* 70, 419–435. doi:10.1111/j.1751-5823.2002.tb00178.x.
- 511 Gokila, S., Ananda Kumar, K., Bharathi, A., 2016. Different versions of k-mean clustering in complete set of numerical data points. *International Journal of Scientific Engineering and Applied Science* 2.
- 512 Guo, G., Dyer, C.R., 2007. Patch-based image correlation with rapid filtering, in: *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6. doi:10.1109/CVPR.2007.383373.
- 513 Hadzimejlic, N., Donko, D., Hadzimejlic, N., 2013. Climate data analysis using clustering data mining techniques. *Latest Trends in Applied Informatics and Computing*.
- 514 Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* 17, 107–145.
- 515 Herrera, D., Ault, T., 2017. Insights from a new high-resolution drought atlas for the caribbean spanning 1950–2016. *Journal of Climate* 30, 7801–7825. URL: <https://doi.org/10.1175/JCLI-D-16-0838.1>, doi:10.1175/JCLI-D-16-0838.1.
- 516 Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Wolff, D.B., Adler, R.F., Gu, G., Hong, Y., Bowman, K.P., Stocker, E.F., 2007. The trmm multisatellite precipitation analysis (tmpr): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology* 8, 38–55. doi:10.1175/JHM560.1.
- 517 Jain, A.K., 2008. Data clustering: 50 years beyond k-means, in: *International Conference on Pattern Recognition (ICPR)*, ICPR.
- 518 Jury, M.R., 2009. An intercomparison of observational, reanalysis, satellite, and coupled model data on mean rainfall in the caribbean. *Journal of Hydrometeorology* 10, 413–430. doi:10.1175/2008JHM1054.1.
- 519 Jury, M.R., Bernard, D., 2019. Climate trends in the east antilles islands. *International Journal of Climatology* 40, 36–51. URL: <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6191>, doi:10.1002/joc.6191, [arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.6191](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.6191).
- 520 Jury, M.R., Malmgren, B.A., 2012. Joint modes of climate variability across the inter-americas. *International Journal of Climatology* 32, 1033–1046.
- 521 Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data - an introduction to cluster analysis. *Journal of Applied Meteorology*, 1131–1147.
- 522 Kullback, S., Leibler, R., 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.
- 523 Kwitt, R., Uhl, A., 2008. Image similarity measurement by kullback-leibler divergences between complex wavelet subband statistics for texture retrieval, in: *2008 15th IEEE International Conference on Image Processing*, pp. 933–936. doi:10.1109/ICIP.2008.4711909.
- 524 Lallich, S., Lenca, P., 2015. Indices de qualité en clustering, in: *Journée thématique : clustering et co-clustering, Société française de classification*, Issy Les Moulineaux, France.

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

- 547 Martinez, C., Goddard, L., Kushnir, Y., Ting, M., 2019. Seasonal climatology and dynamical mechanisms of rainfall in the caribbean. *Climate*
548 *Dynamics* doi:10.1007/s00382-019-04616-4.
- 549 Michelangeli, P.A., Vautard, R., Legras, B., 1995. Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences* 52,
550 1237–1256. doi:10.1175/1520-0469(1995)052<1237:WRRAS>2.0.CO;2.
- 551 Monteleoni, C., Schmidt, G.A., Alexander, F., Niculescu-Mizil, A., Steinhäuser, K., Tippett, M., Banerjee, A., Blumenthal, M.B., Ganguly, A.R.,
552 Smerdon, J.E., Tedesco, M., 2013. *Climate informatics*. Chapman and Hall/CRC. Data Mining and Knowledge Discovery Series, pp. 81–126.
- 553 Moron, V., Frelat, R., Jean-Jeune, P.K., Gaucherel, C., 2015. Interannual and intra-annual variability of rainfall in haiti (1905–2005). *Climate*
554 *Dynamics* 45, 915–932. doi:10.1007/s00382-014-2326-y.
- 555 Moron, V., Gouirand, I., Taylor, M., 2016. Weather types across the caribbean basin and their relationship with rainfall and sea surface temperature.
556 *Climate Dynamics* 47, 601–621. doi:10.1007/s00382-015-2858-9.
- 557 Pandit, S., Gupta, S., 2011. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer*
558 *Science* 2, 29–31.
- 559 Parmar, H., Saket, S., 2017. Overview of clustering algorithm for weather data. *IJARIE* , 2395–4396.
- 560 Rokach, L., Maimom, O.Z., 2010. *Clustering Methods*. Springer. chapter 15. pp. 321–352.
- 561 Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*
562 20, 5365. doi:10.1016/0377-0427(87)90125-7.
- 563 Shraddha, P., Suchita, G., 2011. Computer science and technology. computing. data processing. *International journal of research in computer*
564 *science* , 29–31.
- 565 Sáenz, F., Durán-Quesada, A.M., 2015. A climatology of low level wind regimes over central america using a weather type classification approach.
566 *Frontiers in Earth Science* 3, 15. doi:10.3389/feart.2015.00015.
- 567 Tian, W., Yuhui, Z., Yang, R., Ji, S., Wang, J., 2014. A survey on clustering based meteorological data mining. *International Journal of Grid and*
568 *Distributed Computing* 7, 229–240.
- 569 Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in data set via the gap statistic. *Royal Statistic Society* 5, 411–423.
- 570 Vautard, R., 1990. Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Monthly Weather Review* 118, 2056–
571 2081. doi:10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2.
- 572 Vigaud, N., Robertson, A., 2017. Convection regimes and tropical-midlatitude interactions over the intra-american seas from may to november.
573 *International Journal of Climatology* 37, 987–1000. doi:10.1002/joc.5051.
- 574 Walker, S., Damien, P., Lenk, P., 2004. On priors with a kullback-leibler property. *Journal of the American Statistical Association* 99, 404–408.
575 URL: <http://www.jstor.org/stable/27590396>.
- 576 Xin, L., 2017. *Perceptual Digital Imaging: Methods and Applications*. CRC Press. volume 1. chapter Patch-Based Image Processing: From Dictio-
577 nary Learning to Structural Clustering.
- 578 Yue, S., Pilon, P., Cavadias, G., 2002. Power of the mann-kendall and spearman's rho tests for detecting monotonic trends in hydrological se-
579 ries. *Journal of Hydrology* 259, 254 – 271. URL: <http://www.sciencedirect.com/science/article/pii/S0022169401005947>,
580 doi:[https://doi.org/10.1016/S0022-1694\(01\)00594-7](https://doi.org/10.1016/S0022-1694(01)00594-7).

Table 1

Boundaries of the histogram classes used to quantify daily rainfall data. These edges were determined from rainfall records of the study area.

Centiles (%)	0	0.35	0.5	0.7	0.8	0.9	0.95	0.99	1
Rainfall (mm)	0]0,1.2]]1.2,2.2]]2.2,5.2]]5.2,8.7]]8.7,16.4]]16.4,26.9]]26.9,59.2]]59.2,+∞[

Table 2

Descriptive statistics and probabilities: analysis of the distribution of hurricanes and tropical storms in the five clusters of the KMS-ED method. $P_{TS}(Cx)$ expresses the probability that a TS is in Cx , $P_H(Cx)$ expresses the probability that a H is in Cx , $P_{Cx}(TS)$ expresses the probability that Cx produces a TS, and $P_{Cx}(H)$ expresses the probability that Cx produces a H.

Clusters	TS	H	Cluster sizes	$P_{TS}(Cx)$	$P_H(Cx)$	$P_{Cx}(TS)$	$P_{Cx}(H)$
C1	1	12	799	0.013	0.049	0.001	0.015
C2	20	88	677	0.253	0.362	0.029	0.130
C3	11	39	567	0.139	0.160	0.019	0.069
C4	3	9	1749	0.038	0.037	0.002	0.005
C5	44	95	1623	0.557	0.391	0.027	0.058
Total	79	243	5415				

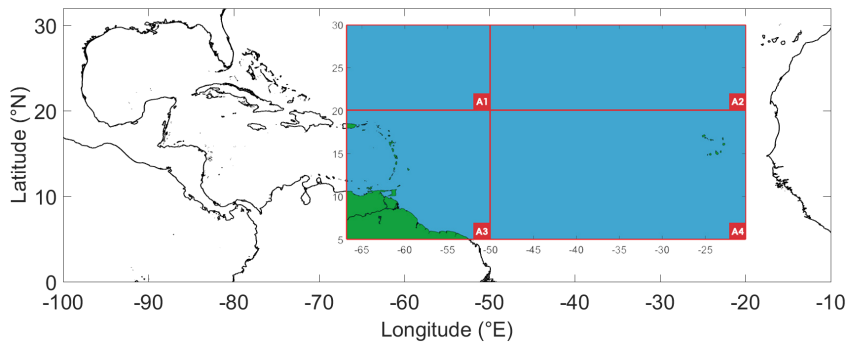


Figure 1: Area of interest. Land is in zone A3: Lesser Antilles with the northeasterly part of South America. Zones A1, A2, and A4 are predominantly sea: a part of the Central Atlantic Ocean and the Cape Verde archipelago. These four zones were used for the design of the expert distance (ED).

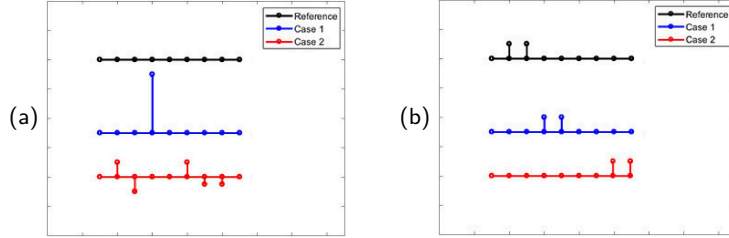


Figure 2: Representation of the characteristics of the L2 distance: (a) a strong and localised fluctuation (blue) produces the same L2 distance as a multitude of small variations (red) from the reference (black); (b) whether the spatial shift is low (blue) or high (red), it produces the same L2 distance from the reference (black).

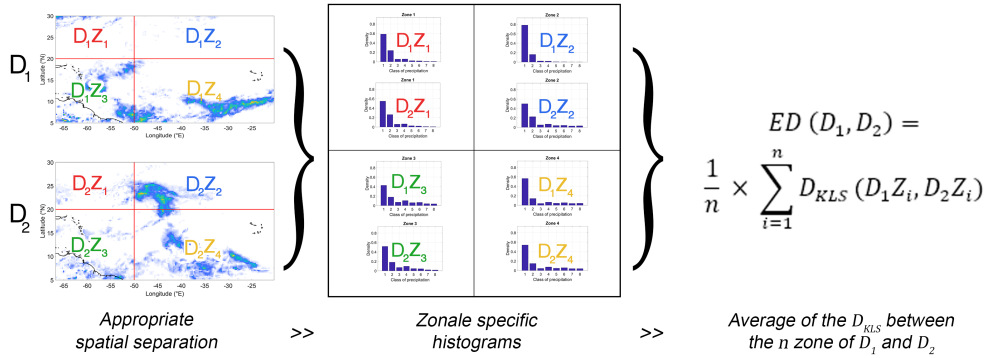


Figure 3: Schematic showing the computation process of the expert distance (ED) for two days D_1 and D_2 : zonal quantification using custom edges ($D_x Z_i$), the use of symmetrised Kullback–Leibler divergence (D_{KLS}) on each zone to obtain four values, and the computation of the average to obtain $ED(D_1, D_2)$.

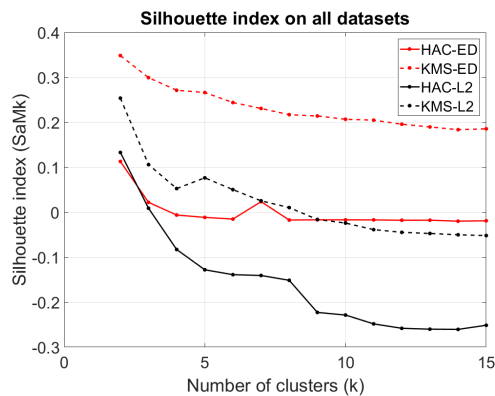


Figure 4: Diagram of the evolution of silhouette index ($Sa(M_k)$ defined in Equation 5) in the function of k ; the number of clusters—HAC (solid line), KMS (broken line)—using L2 (black) and ED (red).

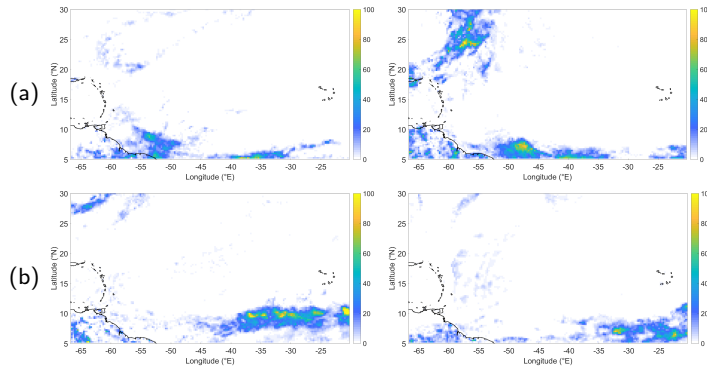


Figure 5: (a) Two days, for members of the same KMS-L2 cluster, but describing very different precipitation fields (TRMM), (b) Two days, for members of different KMS-L2 clusters, but very similar physically.

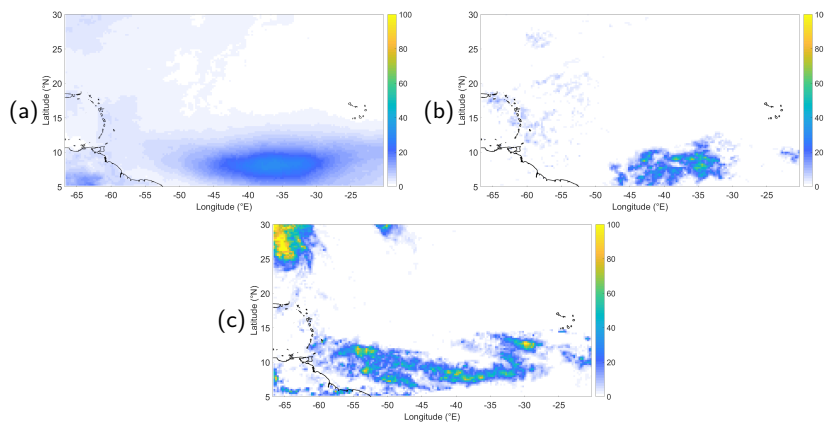


Figure 6: Example of the centroid of cluster C2 from *KMS-L2* computed by using the average (a) which is compared to the nearest element according to L2 (b) and another element of the cluster taken at random (c).

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

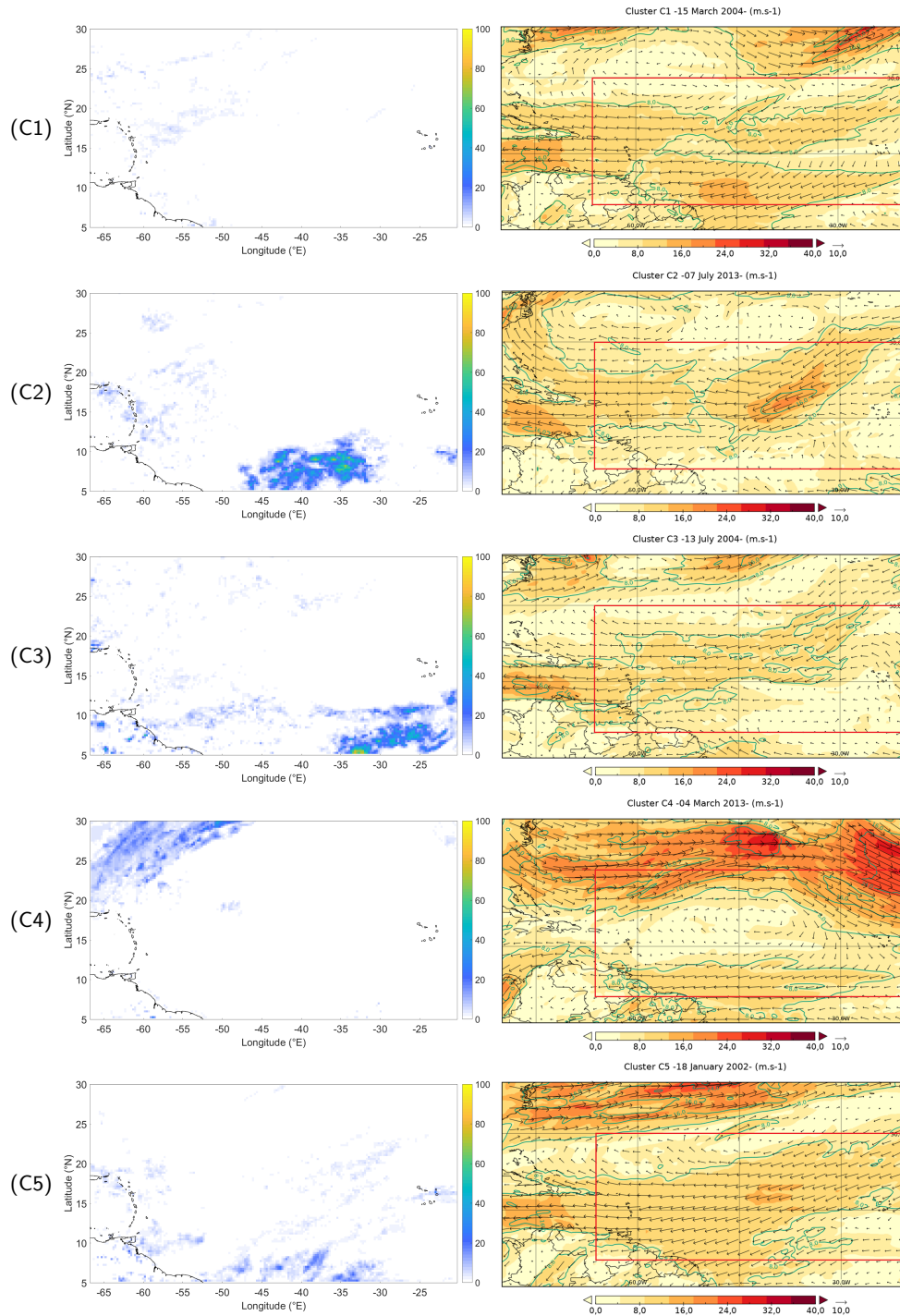


Figure 7: LEFT: Graph of representative elements from the KMS-L2 method, with $k = 5$, RIGHT: Graph of the wind direction and velocity from ERA-5 corresponding to the representative elements of the five clusters of the KMS-L2 method.

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

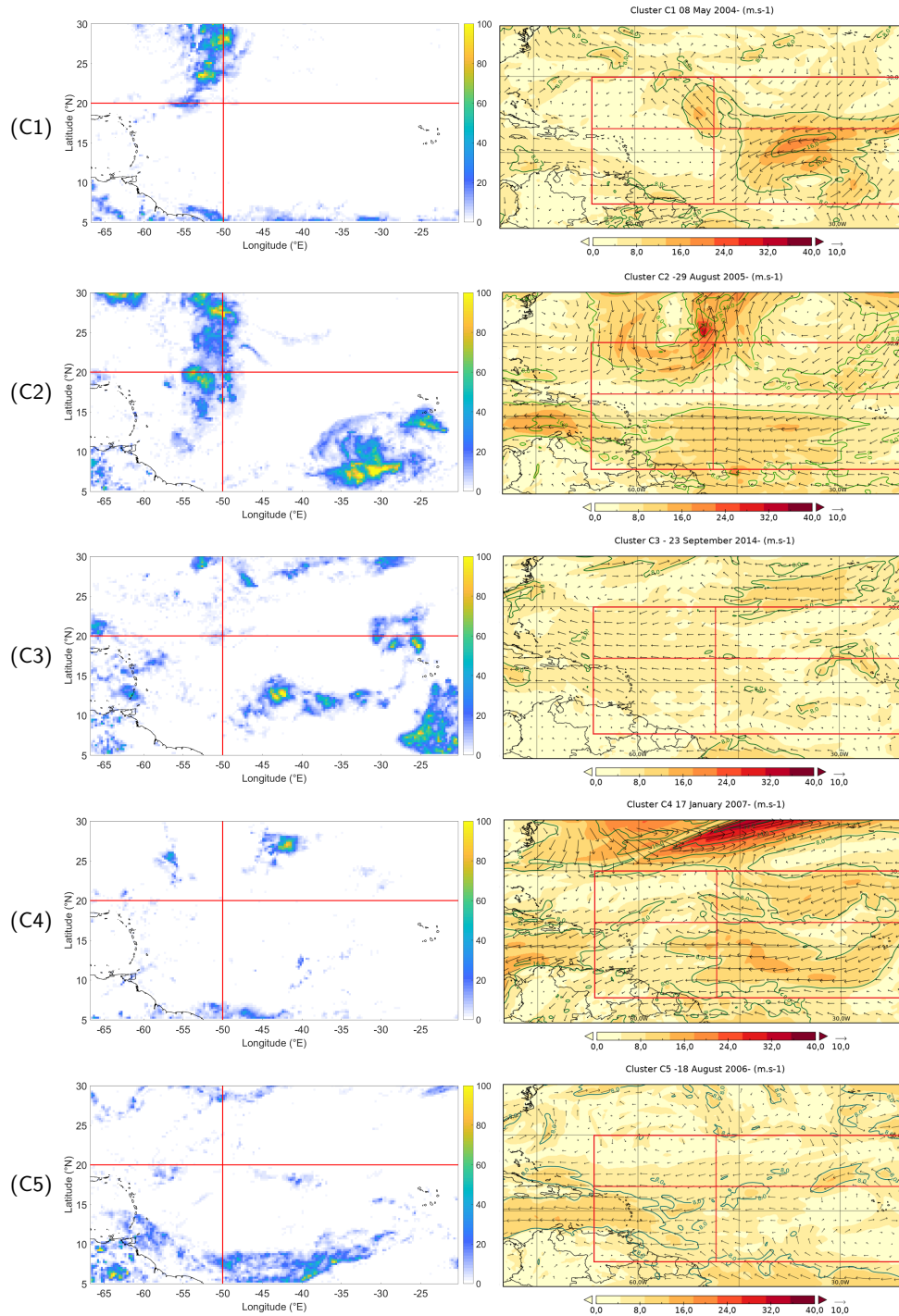


Figure 8: LEFT: Graph of representative elements of the clusters of TRMM rainfall from the KMS-ED method, with $k = 5$. RIGHT: Graph of the wind direction and velocity from ERA-5 corresponding to the representative elements of the five clusters of the KMS-ED method.

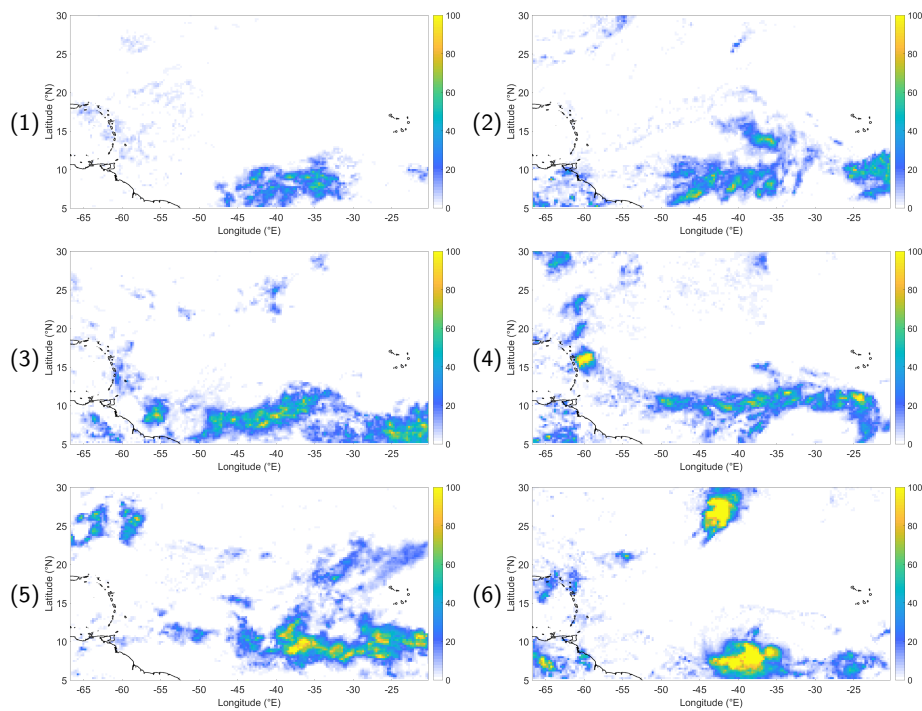


Figure 9: Internal variability of cluster (C2) from the KMS-L2 method. Six days of this cluster are presented in increasing order of the L2 distance from the representative features of the cluster: (1) representative element and (2,3,4,5,6) other elements taken from a regular interval in relation to their distance L2 with (1).

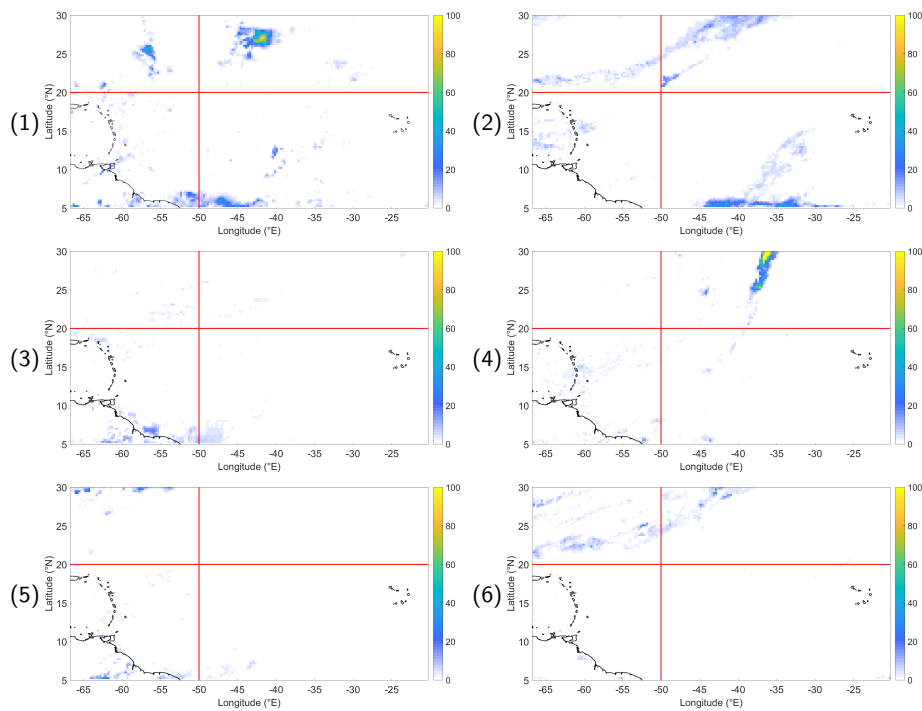


Figure 10: Internal variability of cluster (C4) from the *KMS-ED* method. Six days of this cluster are presented in increasing order of the L2 distance from the representative features of the cluster: (1) representative element and (2,3,4,5,6) other elements taken from a regular interval in relation to their ED with (1).

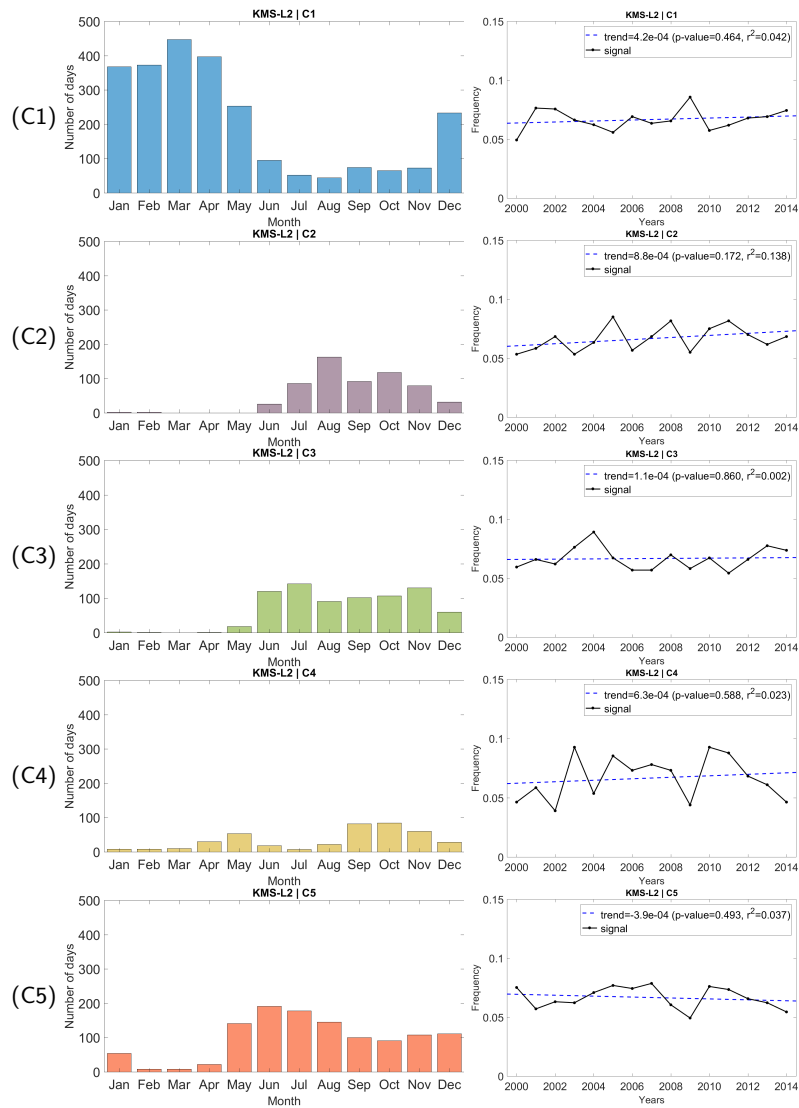


Figure 11: LEFT: Monthly distribution of clusters using the KMS-L2 method over the year for the period 2000 to 2014. RIGHT: Variation in the frequency of each of the clusters over the years (with a dashed line for trends and the validity indexes). p-values are not statistically significant.

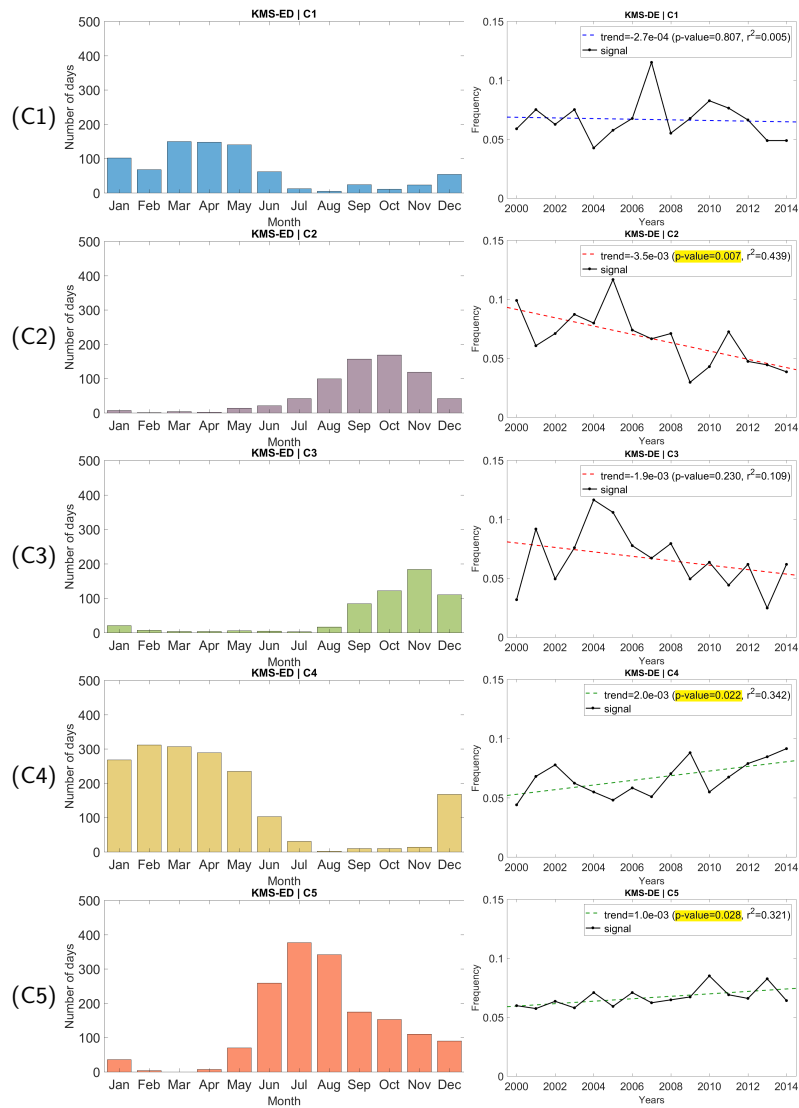


Figure 12: LEFT: Monthly distribution of clusters using the KMS-ED method over the year for the period 2000 to 2014. RIGHT: Variation in the frequency of each of the clusters over the years (with a dashed line for trends and the validity indexes). Highlighted p-values are statistically significant (<0.05).

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

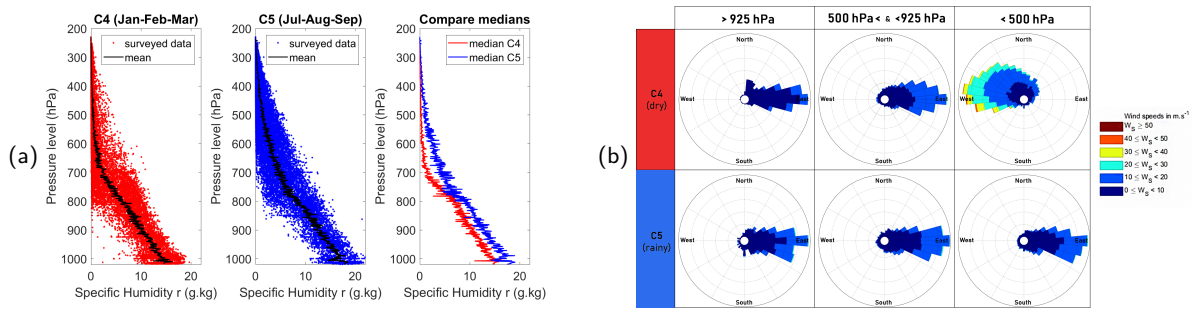
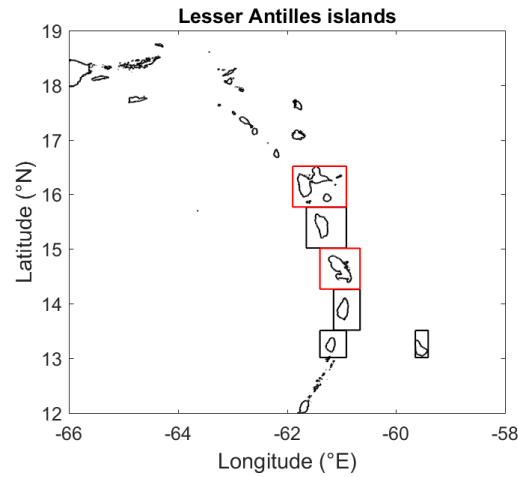


Figure 13: (a) Diagram of the evolution of humidity in the air layers as a function of the pressure level: the data collected by a radiosonde belonging to cluster C4 in the period January–February–March (in red) and that of cluster C5 in the period July–August–September (in blue), with their respective means (in black). (b) Diagram of the evolution of the wind direction and velocity in the air layers as a function of the pressure level: the data collected by a radiosonde belonging to cluster C4 (in red) and that of cluster C5 (in blue).



Clusters Infos/Islands	Guadeloupe	Dominica	Martinique	St-Lucia	Barbados	St-Vincent	Flat islands
C1	MSS [mm/day]	8.98	16.24	13.21	10.79	4.14	7.21
	MSM [mm/day]	1.50	1.80	1.47	1.80	2.07	1.80
	DWR [%]	37	39	34	43	57	50
C2	MSS [mm/day]	38.68	57.78	45.88	33.48	12.12	24.55
	MSM [mm/day]	6.45	6.42	5.10	5.58	6.06	6.14
	DWR [%]	22	24	18	26	42	37
C3	MSS [mm/day]	12.45	19.24	19.29	13.64	6.89	13.55
	MSM [mm/day]	2.07	2.14	2.14	2.27	3.44	3.39
	DWR [%]	34	31	24	33	49	50
C4	MSS [mm/day]	6.27	10.67	9.84	7.40	4.14	7.31
	MSM [mm/day]	1.04	1.19	1.09	1.23	2.01	1.82
	DWR [%]	38	36	31	40	57	55
C5	MSS [mm/day]	33.48	57.74	46.69	30.78	14.60	26.46
	MSM [mm/day]	5.58	6.19	5.19	5.13	7.30	6.61
	DWR [%]	24	21	14	19	41	35

Figure 14: UP: Lesser Antilles islands. DOWN: Detailed rainfall values measured by satellite for the Lesser Antilles islands (with **MSS**=Mean of spatial sum [mm/day], **MSM**=Mean of spatial mean[mm/day], and **DWR**=Percentage of days without rainfall [%]), for KMS-ED clusters (from C1 to C5).

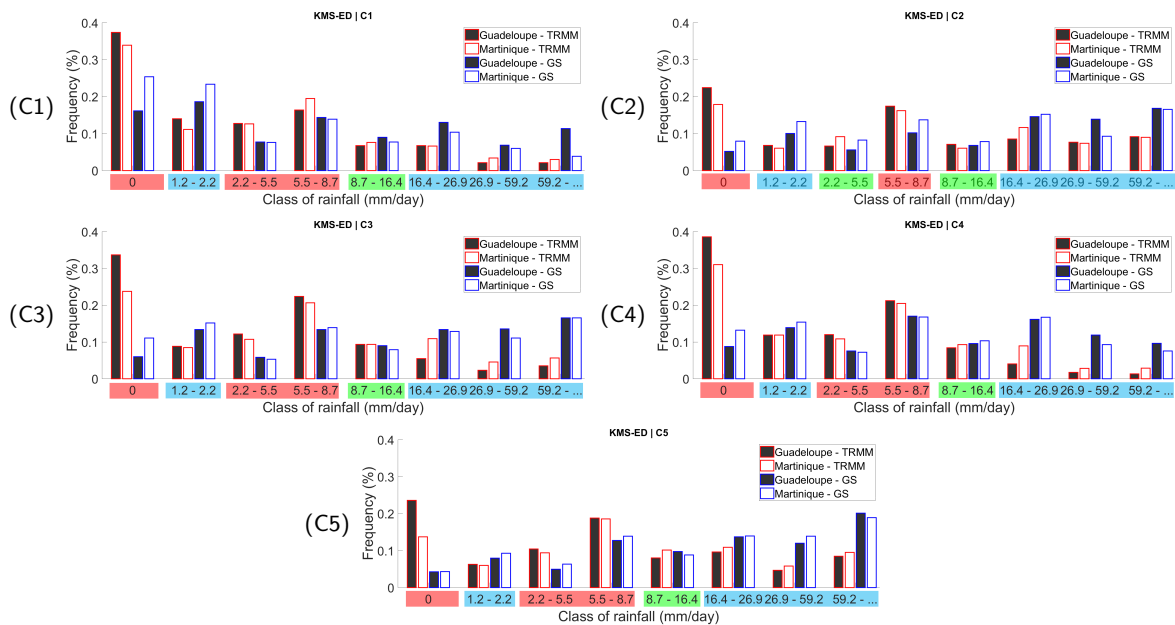


Figure 15: Distribution of the TRMM rainfall (red outline) compared to ground stations (GS) rainfall (blue outline) observed in Guadeloupe (in black) and Martinique (in white) for KMS-ED clusters (from C1 to C5). Classes that are overestimated by TRMM are highlighted in red, those that are underestimated are highlighted in blue, and when TRMM is nearly similar to GS, the classes are highlighted in green.

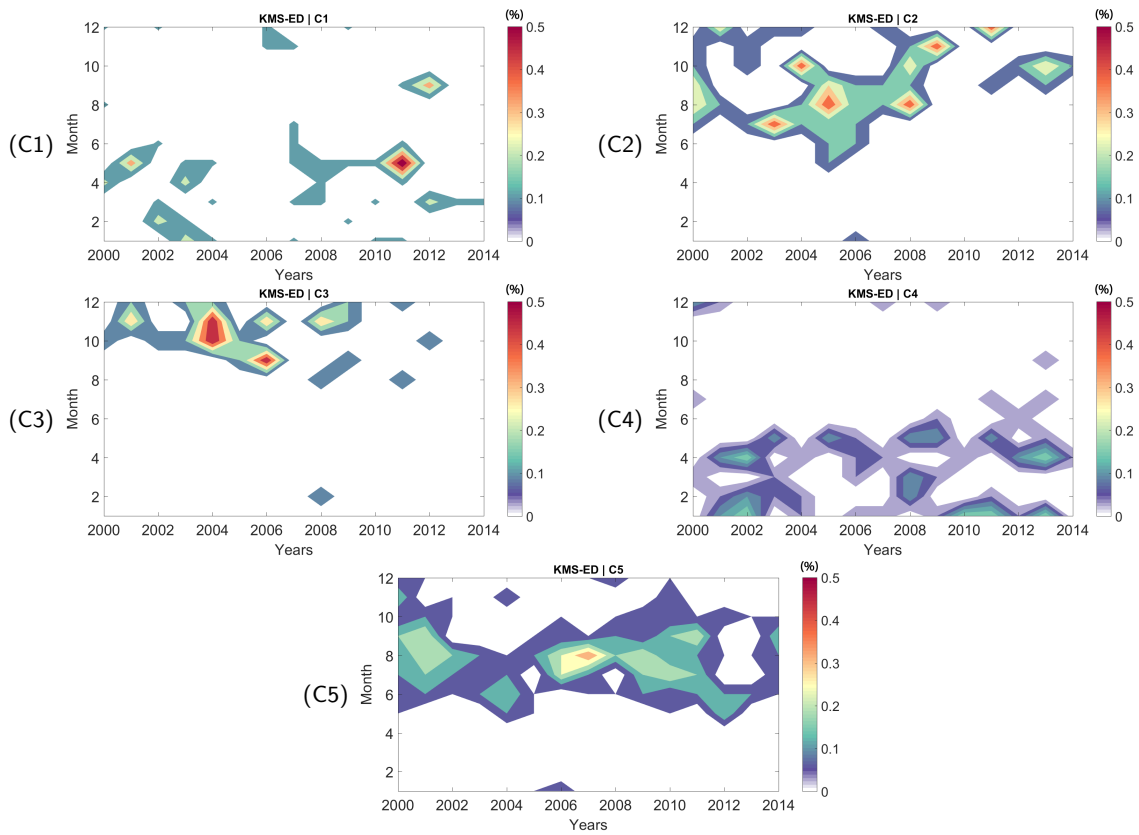
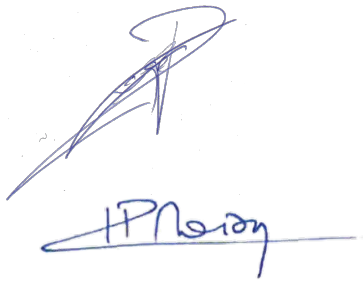


Figure 16: Variation in the intra-annual (y -axis) frequency distribution of moderate rainfall in Guadeloupe (8.7–16.4 mm/day) for the analysed period (x -axis) from 2000 to 2014 in the five different clusters of KMS-ED (from C1 to C5).

A handwritten signature in blue ink, consisting of a stylized, cursive script that appears to read "IP Nelson". The signature is written on a light blue grid background.

07/07/2020

09/07/2020

Graphical Abstract

Design of an expert distance metric for climate clustering: the case of rainfall in the Lesser Antilles

Emmanuel Biabiany, Didier C. Bernard, Vincent Page, H el ene Paugam-Moisy

