

Article

Intra-Storm Pattern Recognition through Fuzzy Clustering

Konstantinos Vantas  and Epaminondas Sidiropoulos  *

Department of Rural and Surveying Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; kon.vantas@gmail.com

* Correspondence: nontas@topo.auth.gr

Abstract: The identification and recognition of temporal rainfall patterns is important and useful not only for climatological studies, but mainly for supporting rainfall–runoff modeling and water resources management. Clustering techniques applied to rainfall data provide meaningful ways for producing concise and inclusive pattern classifications. In this paper, a timeseries of rainfall data coming from the Greek National Bank of Hydrological and Meteorological Information are delineated to independent rainstorms and subjected to cluster analysis, in order to identify and extract representative patterns. The computational process is a custom-developed, domain-specific algorithm that produces temporal rainfall patterns using common characteristics from the data via fuzzy clustering in which (a) every storm may belong to more than one cluster, allowing for some equivocation in the data, (b) the number of the clusters is not assumed known a priori but is determined solely from the data and, finally, (c) intra-storm and seasonal temporal distribution patterns are produced. Traditional classification methods include prior empirical knowledge, while the proposed method is fully unsupervised, not presupposing any external elements and giving results superior to the former.



Citation: Vantas, K.; Sidiropoulos, E. Intra-Storm Pattern Recognition through Fuzzy Clustering. *Hydrology* **2021**, *8*, 57. <https://doi.org/10.3390/hydrology8020057>

Academic Editor: Davide Luciano De Luca

Received: 28 February 2021

Accepted: 22 March 2021

Published: 25 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Knowledge of the temporal and spatial distribution of rainfall is essential both for climatological studies, especially regarding climate change, and for purposes of flood studies and water resources planning. Effective and illuminating studies of rainfall data are achieved through the detection of patterns or groupings. Stochastic precipitation models utilize Markov chains to simulate the occurrence of wet or dry days and consequently the daily precipitation depth [1–3]. Numerous models, extensively developed, are based on the concept of rectangular pulses point process (RPPP), which can be categorized into two types, the Bartlett–Lewis [4] and the Neyman–Scott [5,6] precipitation models. Both of them use the assumption that storms arrive according to a Poisson process, the most basic example of continued-time Markov chains, of which a concise description can be found in Onof et al. [7]. The application of RPPP methods concerns the fitting of a small set of parameters that define the distributions that follow different rainfall quantities, such as the rainstorm and rain-cell origins, durations and intensities, a multi-objective problem without analytical solution and, sometimes, erratic results [5,7–9].

A different family of models are the profile-based ones [10,11] that utilize the internal structure of rainstorms in their entirety and not as derived parameters of statistical distributions. These rainfall models employ the construction of Huff or mass curves [12], a probabilistic method for sub-daily precipitation records, in which rainstorm data are represented in the form of normalized time, versus normalized cumulative precipitation depth and classified by the quartile where the maximum intensity occurs. The use of mass curves offers several advantages: (a) adequate stability regarding the sample size of rainstorms [13]; (b) coarser hourly data giving nearly identical results to data with

finer time-step in the order of minutes [13]; (c) similarity in very long distances [12–14]; (d) the fact that they are not affected by elevation, type of storm, storm duration or storm precipitation depth [15].

Given that precipitation records contain both wet and dry periods, an important issue is the delineation or the extraction of individual rainstorm events from the contiguous recorded precipitation data. A systematic review [16] of different methods that are based on various rainstorm events characteristics (depth, intensity, generated runoff) reveals the necessity of selecting a criterion that defines objectively the rainless intervals, or minimum inter-event time, and not the arbitrary selection of constant time length. The latter is a common practice: Huff used 6 h to separate storms [12], Yu et al. used 2 h [17] and the calculation of rainfall erosivity in the universal soil loss equation (USLE) and its revisions [18] uses, also, 6 h. A different approach was proposed by Restrepo-Posada and Eagleson [19], commonly used today by many authors [13,20], in which “an empirical and inexact”, but easily applied, method was developed for the separation of precipitation timeseries, into statistically independent rainstorms.

Unsupervised learning methods, clustering in particular, are now employed for pattern recognition in hydrological data [21]. In general, clustering analysis is a popular exploratory task aimed at partitioning the content of databases into smaller groups on the basis of inherent similarity criteria [22]. Clustering yields meaningful patterns to be further utilized for understanding of processes or for simulations. Clustering methods are applied in many fields, motivated by the large amounts of accumulated data and the developed needs for better data management through grouping and detection of patterns. The unsupervised nature of clustering analysis is a definite advantage, but it raises additional difficulties regarding suitable choice of methods and metrics. The challenge of clustering also lies in adapting the method and its parameters to the nature of the specific application.

A variety of such unsupervised learning methods are reviewed by Sheikholeslami et al. [23], including descriptions of clustering approaches. Among these, the well-known k-means and k-medoids methods divide the data into distinct clusters, such that each element belongs to exactly one cluster. This type of clustering may be characterized as hard, crisp or non-fuzzy. On the contrary, in fuzzy clustering, each point-element is allowed to belong to more than one cluster, albeit with a varying degree of membership. The membership degree of an element is a number in the interval [0,1]. Points closer to the center of a cluster are assigned a higher membership degree than points further away. One of the most widely used fuzzy clustering algorithms is fuzzy c-means clustering. It is used frequently in pattern recognition, and a related review can be found in Nayak et al. [24]. A major problem in clustering is the determination of the optimal number of clusters, which, in real-world applications, is generally not known in advance. As a result, a number of methods have been developed for the determination of the optimal number of clusters. Many of them use the concept of relative cluster validity [25], where results from different clustering methods are compared using a predefined metric. A number of these methods can be found in Milligan et al. [26] and Charrad et al. [27].

In hydrology, machine learning and clustering methods are being applied with increasing frequency, but, apparently, not to their full potential yet. More specifically, cluster analysis has been applied for the identification of hydrologically homogeneous regions, as noted in several literature publications: (a) hierarchical clustering on monthly rainfall erosivity density [21]; (b) fuzzy c-means clustering on annual precipitation [28] and annual maximum intensities [29]; self-organizing maps on monthly precipitation [30], while less work has appeared in terms of temporal pattern investigations applying: (a) the AL algorithm on rainstorms mass curves [31]; (b) self-organizing maps on design hyetographs; (c) k-means, also, on rainstorms mass curves [32]. On the other hand, rainfall mass curves have been utilized, as previously reported, in the stochastic generation of rainfall events [10,11,33,34], also as design storms [14,35–37], for the simulation of floods [38,39] and in changes in storm properties due to climate change [40].

In view of the above considerations, this paper aims to investigate the presence of intra-storm temporal patterns using timeseries of rainfall data coming from the Greek National Bank of Hydrological and Meteorological Information. Prior to clustering analysis, a more precise statistical analysis is applied in order to define a seasonal model to aid in the extraction of statistically independent rainstorm events, which has never been presented for Greece. The computational process that follows is a custom-developed, domain-specific algorithm that produces temporal rainfall patterns using common characteristics from the data via fuzzy clustering in which (a) every storm may belong to more than one cluster, allowing for some equivocation in the data, (b) the number of the clusters is not assumed known *a priori*, but is determined solely from the data and, finally, (c) intra-storm and seasonal temporal distribution patterns are produced. The optimal temporal rainfall distribution curves presented recently by the authors employed hierarchical clustering on principal components and utilized data from a specific water region of Greece [41]. Additionally, a preliminary research that utilizes self-organizing maps and crisp clustering has been presented by the authors in a recent European Geosciences Union General Assembly [42]. The present paper (a) adopts the more general approach of fuzzy clustering, which, to our knowledge, has not been applied so far to rainstorm timeseries; (b) the data utilized extend country wide; (c) the advantages of mass curves mentioned in the literature for different parts of the word are materialized for the case of Greece, regarding regionalization and the effect of elevation. Moreover, the overall result of clustering is verified and compared with the established Huff's classification, which has also never been presented for the country, via visualization effected through non-linear projection and topographic maps that are created using emergent self-organizing maps, a method that can reveal patterns in high-dimensional datasets.

2. Materials and Methods

The methodology that was applied in the study is presented in Figure 1 in the form of a flowchart. Precipitation data with a time step of 30 min were imported; rainstorms were extracted from the dataset and preprocessed in order to have an appropriate form as input to clustering analysis. A cluster tendency assessment was applied, so as to examine if clustering is meaningful, and the optimal number of clusters were determined by a custom method that utilizes fuzzy c-means. Additionally, the Huff's curves were compiled for the same dataset. The resulted patterns were visualized using a nonlinear projection of the data in two dimensions, and finally, their characteristics were examined in terms of internal and seasonal structure.

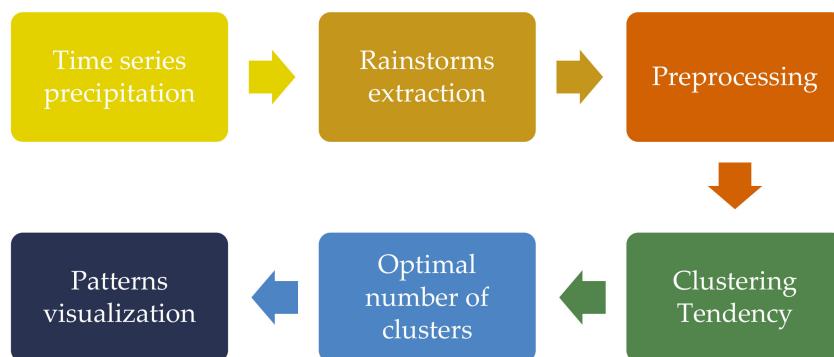


Figure 1. Flowchart of the applied methodology.

2.1. Data Acquisition and Processing

The data used in the analysis were taken from the Greek National Bank of Hydrological and Meteorological Information [43] for 108 meteorological stations across Greece (Figure 2). The timeseries comprised a total of 2926 years of 30 min records for the time period from 1953 to 1997, with a mean length of 26.6 years per station. The timeseries were

checked for consistency and errors as follows: (a) in records of repetitive values near zero (i.e., ≤ 0.01 mm), these values were set to zero and (b) records of aggregated values, where the time step was larger than the reported, were removed. The pluviograph data coverage was 43.2% on average. The missing values in the timeseries are not random points but contiguous data related to entire months or years.

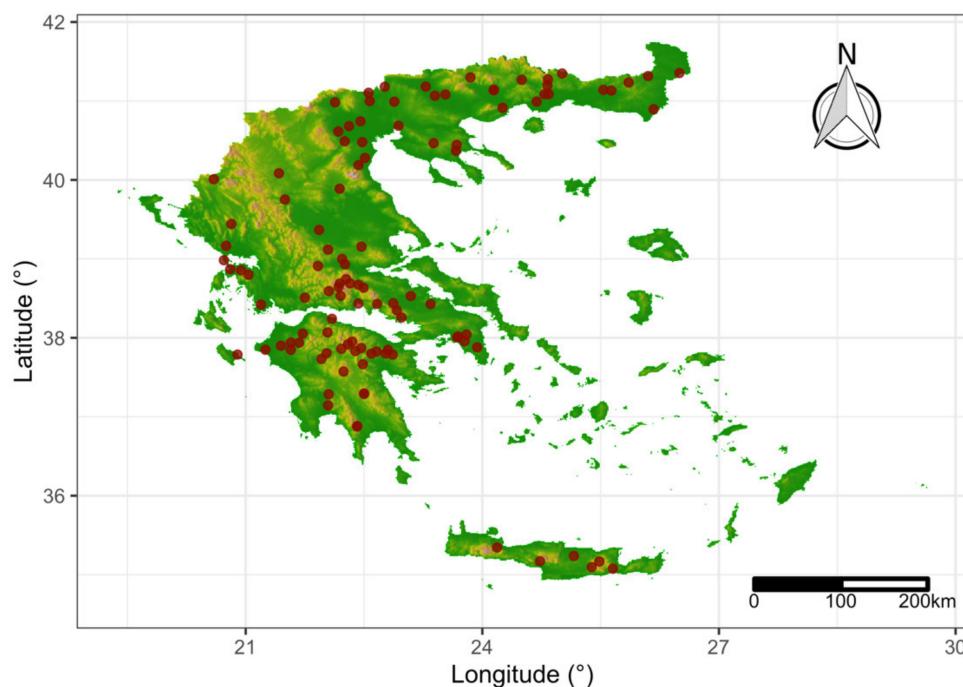


Figure 2. Station locations in Greece used in the analysis obtained from the Greek National Bank of Hydrological and Meteorological Information.

Greece, given its geographic characteristics, the distribution of sea and land and the complex and rich relief, has a mosaic of different micro-climates and regional variations, all in the Mediterranean context. Precipitation varies from its maximum values during winter to a minimum during summer. The highest values are observed westwards and on the mountain range of Pindos and its expansion on Peloponnesus (two to three times higher). This fact reveals the importance of relief on the distribution of rainfall over the country. Furthermore, higher precipitation depth is connected to the movement of Mediterranean depressions that follow a characteristic path from west to east. Northerly, at the valleys of central Macedonia, smaller amounts of precipitation are recorded, due to prevailing dry winds. At the northeastern part, at Thrace, higher values are observed also due to relief. Finally, eastern parts of Greece are drier. During summer months when convective activity prevails and over northern Greece, it produces higher precipitation amounts than in the drier southern parts.

2.2. Extraction of Rainstorms

A Poisson process hypothesis is assumed for the separation of the precipitation timeseries to statistically independent rainstorm events, where:

- The time intervals of rainstorms that come from the same month are distributed exponentially.
- The rainstorms are separated by a minimum critical time duration of no precipitation (**CD**) or inter-event time [19].
- There is a seasonal pattern for **CD** that is assumed to have constant monthly values.

- The probability density function is:

$$f(t_a) = \omega \cdot e^{-\omega \cdot t_a}, \quad t_a \geq 0 \quad (1)$$

where ω is the average storm arrival rate and

$$t_a = t_r + t_b \quad (2)$$

where t_a is the storm interarrival time, t_r is the storm duration, and t_b is the dry time between storms.

The estimation of CD that separates rainstorms is based on an iterative procedure of statistical tests (Algorithm 1). Inter-month data per station are used to ensure homogeneity and: (a) a test value of cd is used, coming from a predefined vector of values **CD**, (b) a vector T_a of t_a values is computed for each cd , (c) ω values for T_a are estimated by means of the maximum likelihood estimation method, (d) the goodness-of-fit between T_a and the exponential distribution is estimated via parametric bootstrapping of s samples that utilizes the Kolmogorov–Smirnov test [44], and (e) the cd value with the maximum p -value from the empirical non-parametric distribution is selected. In Algorithm 1, a threshold of 50 t_a values was imposed empirically prior to statistical testing.

Algorithm 1: Temporal model of CD

Input: Stations' precipitation time series P_i where $i = 1, \dots, k$; Critical durations test vector $CD = [120, 180, \dots, 1800]$ (min); Number of samples that are drawn for parametric bootstrapping $s = 50,000$;

```

1  for station  $i \leftarrow 1$  to  $k$  do
2    for month  $m \leftarrow 1$  to 12 do
3      for  $cd$  in  $CD$  do
4        Compute the vector of interarrival times  $t_a$  using inter-month data and  $n = \text{length}(t_a)$ ;
5        if  $n \geq 50$  then
6          Estimate the average storm arrival rate  $\hat{\omega}$  from  $t_a$  using Maximum Likelihood
          Estimation;
7          Obtain the Kolmogorov–Smirnov's  $p$ -value for the original sample  $t_a$  and the
          estimated distribution;
8          Generate  $s$  samples of size  $n$  from the estimated distribution;
9          For each sample compute the one-sample Kolmogorov–Smirnov's  $p$ -value using the
          estimated distribution as theoretical;
10         Use the empirical non-parametric distribution of  $p$ -values to obtain the  $p$ -value for
          the original sample  $t_a$ ;
11      Get minimum dry period duration  $MDPD_{i,m}$  from  $CD[\max(p-value)]$ ;

```

Result: Monthly values of CD

A first version of the above process was presented in previous publications of the authors [41,45]. The basic hypothesis concerns the application of the probability density function (1). A recent related publication [20] adopted the same exponential distribution to estimate inter-event times in a region of China, but using the “inexact” Restrepo-Posada and Eagleson approach.

2.3. Preprocessing

Prior to clustering analysis, preprocessing of data is necessary, in order to transform them into standardized uniform representations that will facilitate the recognition of the common features. More specifically, each one of the storms is a vector of different length, so that a method must be applied to transform the dataset to one with a fixed number of features that represent time. Thus, the storms were scaled and interpolated to unitless

form in which (a) the time expresses the ratio of the rainstorm duration and (b) the rainfall expresses, also, the ratio of total rainstorm depth:

$$p'_i = \frac{p_i}{\sum_{i=1}^n p_i} \quad (3)$$

$$t'_i = \frac{t_i}{\sum_{i=1}^n t_i} \quad (4)$$

where p_i and t_i are the precipitation depth and time for the timestep i , respectively, and p'_i and t'_i their scaled values.

In the sequel, since the scaled rainstorm vectors have variable length, linear interpolation was applied to compute the unitless rainfall for every 1% of unitless time values. Finally, a matrix of unitless rainstorms \mathbf{U} was produced with the values of unitless rainfall, in which every row represents a storm and each column the unitless time values. In the clustering analysis, use was made only of the rainstorms with cumulative rainfall greater than 12.7 mm (i.e., with the exception of light rain) that were no single points, due to the timestep of 30 min, and also erosive by definition [18], as in other similar publications [12–14].

2.4. Clustering Tendency

Before the application of a clustering algorithm, it is advisable to have a preliminary look into the dataset in order to detect any existing clustering tendency. This was done in the following ways: (a) a visual assessment of cluster tendency (VAT [46]) and (b) application of the Hopkins index, H [47].

The VAT method creates an image matrix that can be used to visually assess the cluster tendency. In the method, as it was applied, the pairwise dissimilarity values of the rainstorms were computed, and these values were reordered in the form of a matrix and displayed as an intensity image. In this image, the clusters are indicated as dark blocks of pixels along the diagonal [46]. As a dissimilarity measure, the Euclidean distance d was used

$$d(u, v) = \sqrt{\sum_{i=1}^{100} (u_i - v_i)^2} \quad (5)$$

where u and v are two row vectors from the unitless rainstorms \mathbf{U} matrix. The re-ordering was achieved applying agglomerative hierarchical clustering using the Ward's minimum variance criterion, an algorithm that minimizes the total within-cluster variance [48]. At the beginning of the algorithm, the number of the clusters is equal to the number of data points (all clusters contain a single point). At every step, the algorithm finds the pair of clusters that result after merging to the minimum increase in the total within-cluster variance, which is expressed as the sum of squared differences between the clusters' centers. Finally, all clusters are combined to one cluster that contains all the data using a hierarchical method [21].

The Hopkins index, H , can be used to test the null hypothesis of randomly and uniformly distributed data, generated by a Poisson point process and is calculated with

$$H = \frac{\sum_{j=1}^m u_j^d}{\sum_{j=1}^m w_j^d + \sum_{j=1}^m u_j^d} \quad (6)$$

where X is a collection of n data points that have d dimensions. A random sample from X without replacement with members x_i ($i = 1$ to m , $m \ll n$) is formed. Y is a set of uniformly random data points, also with d dimensions and members y_j ($j = 1$ to m), u_j in turn is the Euclidean distance from y_j to its nearest neighbor in X , and w_j is also the Euclidean distance from x_i to its nearest neighbor in X . A value of H close to one indicates that the data are highly clustered, 0.5 indicates randomly distributed data, and zero indicates regularly spaced data [21,25].

2.5. Fuzzy Clustering

The unitless rainstorm data were clustered by the fuzzy c-means (FCM) algorithm, which was developed by J. Dunn [49] and improved by J. Bezdek [50]. FCM aims to minimize the objective function

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^n \sum_{j=1}^c w_{ij}^m \|x_i - c_j\|^2 \quad (7)$$

where $w_{ij} \in [0, 1]$ is the degree or membership of item x_i from a set of n elements $X = x_1, \dots, x_n$ that belong to cluster c_j , C is the set of c cluster centers $C = c_1, \dots, c_c$, and $\|\dots\|$ denotes any norm that expresses similarity, such as the Euclidean distance.

From an iterative optimization procedure on (7), each element w_{ij} of the partition matrix W is equal to

$$w_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - c_k\|}{\|x_i - c_j\|} \right)^{\frac{2}{m-1}}} \quad (8)$$

where m is the “fuzzifier” that determines the level fuzziness with $m \in \mathbb{R}$ and $m \geq 1$, commonly set to two [51]. The larger the m value is, the fuzzier the membership values of the clustered data points are. The centers c_j of the clusters are the mean of all the elements weighted by their degrees:

$$c_j = \frac{\sum_{i=1}^n w_{ij}^m x_i}{\sum_{i=1}^n w_{ij}^m} \quad (9)$$

FCM stops when the maximum number of iterations given a priori is reached, or when the algorithm is unable to reduce the current value of the objective function further to a predefined, usually very small, value.

2.6. Optimal Number of Clusters

The most common and fundamental problem in clustering analysis is the determination of the number of clusters in an unlabeled dataset, as the one used in the analysis. As previously reported and due to the fact that most clustering algorithms, including FCM, require the number c of clusters to be known a priori, the next step, after answering the question of clustering tendency, has to do with the determination of c . The proposed method uses statistical testing (Algorithm 2). Variations of it have been presented by the authors in the context of different clustering algorithms, namely, (a) self-organizing maps [42] and (b) a dendrogram produced by hierarchical clustering on principal components [41].

A different approach has been presented [52], using k-means (crisp) clustering, which involves two different parameters: (a) an initial threshold that indicates similarity, through a Pearson correlation coefficient, between all pairs of the unitless cumulative rainstorms and (b) an additional one, based on the distances matrix of all pairs of the unitless cumulative rainstorms, which creates the initial seeds used in the k-means clustering algorithm. In other studies, the number of clusters was set without any estimation [32] or through a visual, ambiguous method involving the density maps derived from self-organizing maps [53]. Comparing the reported methodologies to the proposed one, Algorithm 2 presents the following advantages: (a) It is easier to implement, and it can be combined with any clustering algorithm that uses as a parameter the number of clusters directly or indirectly, as is for example the eps parameter in the DBSCAN algorithm [54]. (b) It utilizes hypothesis testing about the centers of the clusters, and (c) it has only one parameter to choose, the significance level of well-known statistical tests.

In Algorithm 2, firstly, the cumulative values of the unitless rainstorms matrix are computed, as that kind of transformation was found to help the computational procedures of clustering algorithms. At each step of the iteration, FCM is applied using a trial number of clusters, greater than two and smaller than a predefined maximum value n_{max} . Afterwards, the cluster centers that resulted are represented as cumulative rainfall distributions.

These centers, and for all possible pairs, are tested to find out whether they are drawn from the same distribution using the two-sample Kolmogorov–Smirnov test [55]. The test quantifies the distance between the two samples' empirical distribution functions and has been used in similar comparisons [56,57]. Due to the multiple statistical testing that arise, the resulted p -values are adjusted using the Benjamini and Hochberg method [58], which controls the false discovery rate. The algorithm stops when any p -value is greater than a predefined significance level α , and in that case, the clusters from the previous step are returned.

Algorithm 2: Optimal number of clusters using FCM

Input: unitless rainstorms U ; maximum number of clusters to test n_{max} ; significance level $\alpha = 0.05$

- 1 compute the row-wise cumulative values U' of U
- 2 **for** $i \leftarrow 2$ to n_{max} and all p -values $< \alpha$ **do**
- 3 apply FCM on U' for $c = i$ and compute cluster centers C ;
- 4 for all pairs in C obtain the Kolmogorov–Smirnov two sample test, p -values;
- 5 adjust the obtained p -values using Benjamini and Hochberg method;

Result: optimal number of clusters c_{opt} and clustering results

2.7. Projecting Data Using Non-Linear Mapping

The non-linear method that was used to visualize the rainstorm data in a two-dimension scatterplot was the generalized U-Matrix [59]. This method uses the results from a dimensional reduction method, such as principal components analysis, or a non-linear method, such as t stochastic neighbor embedding (t-SNE, [60]), in conjunction with emergent self-organizing maps (ESOM [61]). This step is a necessity due to the Johnson–Lindenstrauss lemma [62] stating that the “low-dimensional similarities do not represent high-dimensional distances coercively” [63]. In other words, specific measures are taken to avoid the common mistake of assuming that, when the projected points are similar to each other after dimensional reduction, such is the case in the actual high-dimensional space. The generalized U-Matrix after its calculation by ESOM is visualized by a three-dimensional space called “topographic map with hypsometric tints” or colors on surface that represent elevation ranges [64]. The topographic map has no actual real borders and is toroidal, which means that the left border is connected to the right one and the top to the bottom.

3. Results and Discussion

3.1. Rainstorms Extraction

Empirically, a minimum set of 50 values per month and station were used in Algorithm 1 and some basic statistics; results about CD are presented in Table 1. Due to the dry summer period in Greece, these monthly CD values could be computed only in a small set of stations. A seasonal sinusoidal model (Figure 3) was developed to use in Greece:

$$f(CD) = \theta_1 \sin\left(\frac{2\pi}{12}m\right) + \theta_2 \sin\left(\frac{4\pi}{12}m\right) + \theta_3 \cos\left(\frac{2\pi}{12}m\right) + \theta_4 \cos\left(\frac{4\pi}{12}m\right) \quad (10)$$

where m is the month, and $\theta_1, \dots, \theta_4$ parameters that were estimated using linear regression.

Using the CD sinusoidal model, a set of 174,883 rainstorms were extracted from the dataset. With the exception of summer months and September, the model gives values around 6 h, close to the value selected empirically by Huff [12] and in rainfall erosivity calculations [65]. A subset of 26,678 rainstorms that met the criterion of minimum depth were preprocessed and used to compile the U matrix.

Table 1. Average statistical properties of monthly CD values for the stations. SD is an abbreviation for standard deviation and CV for coefficient of variation (the ratio of the standard deviation to the mean) and h for hours.

CD (h)	Min	Mean	Median	Max	SD	Skew	Kurtosis	CV
January	2	5.4	5	13	1.6	1.40	4.77	0.19
February	2	5.0	5	10	1.4	0.90	1.39	0.16
March	2	5.9	5	12	1.8	0.95	0.82	0.20
April	4	6.3	6	10	1.4	0.62	-0.20	0.16
May	4	6.8	6	12	1.9	0.91	0.49	0.24
June	4	8.2	8	13	2.1	0.15	-0.50	0.34
July	5	9.3	9	13	2.0	-0.17	-0.01	0.58
August	5	7.8	8	11	2.1	0.12	-1.73	0.70
September	6	9.1	9.5	11	1.6	-0.52	-1.03	0.45
October	2	7.4	7	13	1.9	0.36	0.78	0.25
November	2	6.7	6	11	1.6	0.25	0.17	0.19
December	2	5.2	5	11	1.4	0.85	2.00	0.16

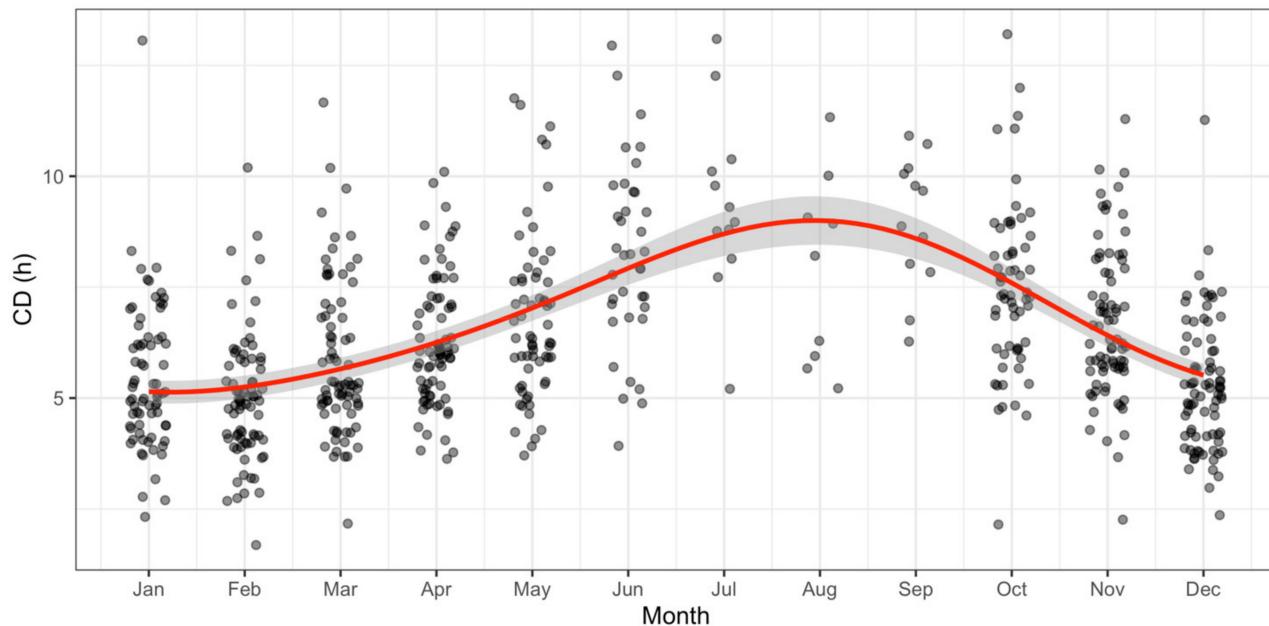


Figure 3. Red represents the monthly sinusoidal model of CD (Equation (11)). The grey area corresponds to the standard error of the model using linear regression. The grey dots represent the monthly CD values per station, as computed from Algorithm 1.

3.2. Clustering Tendency

Due to computation issues, a random sample of 10% was used to compute the value of the Hopkins index with $H = 0.886$, so the null hypothesis of random data could safely be rejected. This result indicated that there was a physical meaning in the categorization of the rainstorms based on their internal structure. The VAT method (Figure 4) created an image matrix, where the clusters are indicated as at least four darker blocks along the diagonal. Given the results from these two methods, there is strong evidence that the dataset contains meaningful clusters and the next step can follow, to apply the proposed method and select the optimal number of clusters.

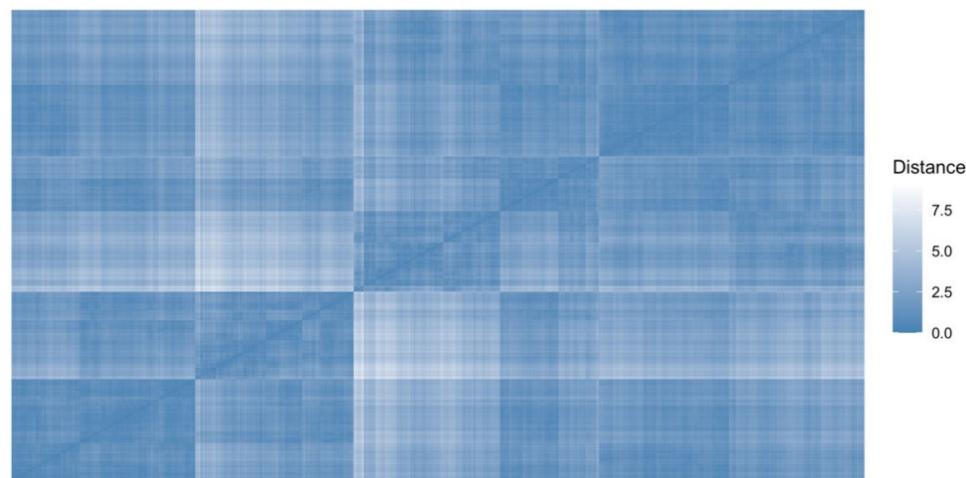


Figure 4. Image matrix created by the VAT method, in order to visually inspect the clustering tendency of data. Distances (dissimilarities) among unitless rainstorms are unitless. The rainstorms that belong to the same cluster are displayed in successive order.

3.3. Clustering Results and Visualization

After the application of Algorithm 2, the optimal number of clusters are proposed to be four, a suitable number comparing it with VAT results. In Figure 5, the matrix of unitless rainstorms \mathbf{U} is depicted, with its rows, the rainstorms, reordered by the cluster they belong to. This representation shows that the rainstorms belonging to clusters number one and four are more similar. The latter statement is strengthened by Figure 6 that illustrates the distribution of the membership of every unitless rainstorm that belongs to a cluster (that value must be >0.25 to be classified to a cluster). In the same figure, the mean values of the memberships for each one of the clusters are $\{0.67, 0.60, 0.59, 0.67\}$. The latter values are not normally distributed, and cluster one and four are left-skewed meaning less equivocality, in contrast to clusters two and three that are right-skewed.

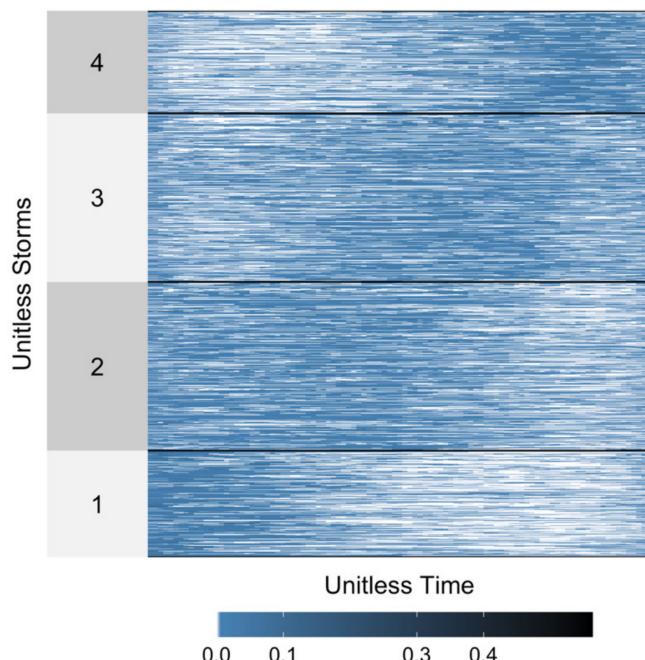


Figure 5. Unitless rainstorms values re-order by clustering results. The names of the clusters were used in order to match with Huff's classification. White symbolizes zero precipitation depth, and different blues depict the presence of unitless precipitation.

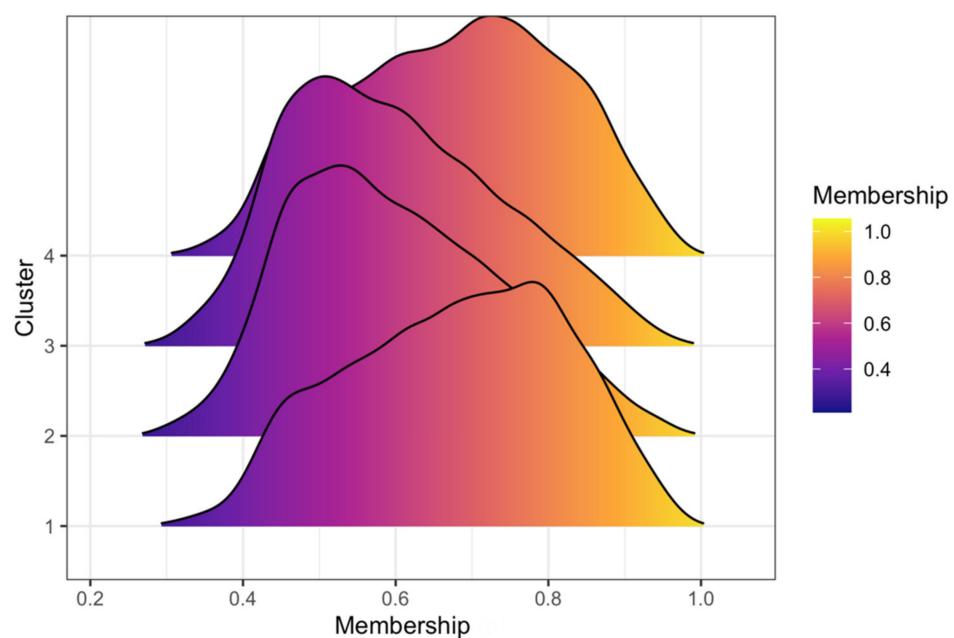


Figure 6. Distribution of the membership values from fuzzy c-means (FCM) of every unitless rainstorm that belongs to a cluster.

From Figure 5, initially, it seems that the selection of Huff about the grouping in four clusters has physical meaning, in contrast to criticism about being artificial [66].

The average values related to precipitation depth, duration and intensity per cluster are presented in Table 2. Cluster one has on average shorter duration rainstorms, with lower precipitation depth, but higher mean intensity. Clusters two and three have both rainstorms with similar statistics, with larger duration and precipitation depth than one and four. Cluster four has analogous statistics to one with the exception of intensity, which is lower.

Table 2. Average values of occurrence of clusters, duration and precipitation depth intensity of clusters' rainstorms.

	Cluster			
	1	2	3	4
Cluster Ratio (%)	19.5	30.86	30.84	18.79
Duration (h)	12.5	16	16.5	14.5
Precipitation depth (mm)	20.7	23.2	23.7	21.9
Intensity (mm/h)	1.82	1.66	1.66	1.71

Cluster number one has notably higher variance and maximum occurrence during spring and summer months that match convective activity over Greece (Figure 7). Clusters two and three have distributions similar to each other and inverse to the first one, related to the prevailing winter weather systems, having their minima during summer. Cluster four has a more uniform distribution with the exception of winter months, when it has its maximum. The latter cluster, with its relatively higher intensity values and, also, higher ratio of occurrence during the wetter winter months in Greece, can be utilized in hydrologic design in the construction of design storms.

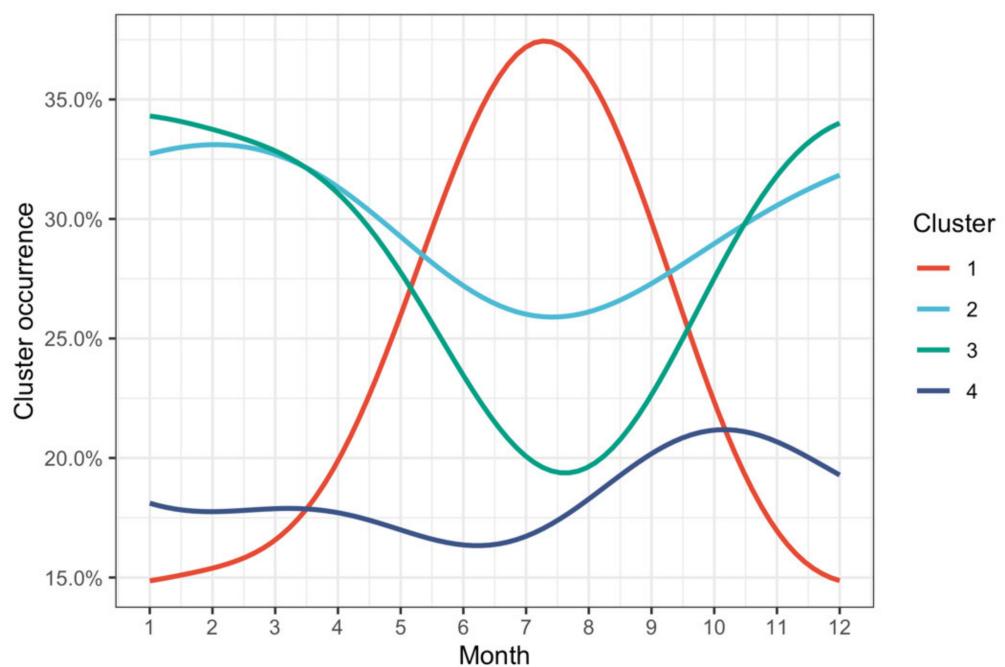


Figure 7. Seasonal variability of clusters' monthly occurrence.

In order to examine if elevation affects the distribution of rainstorms in the clusters, density plots were compiled (Figure 8) for each cluster. The distribution of the elevations of the stations related to rainstorms appears to be almost identical, despite the fact that altimetry is closely connected to the precipitation regime in Greece. This result about elevation is in accordance with the reported one for Huff's curves by Loukas et al. in Canada [15].

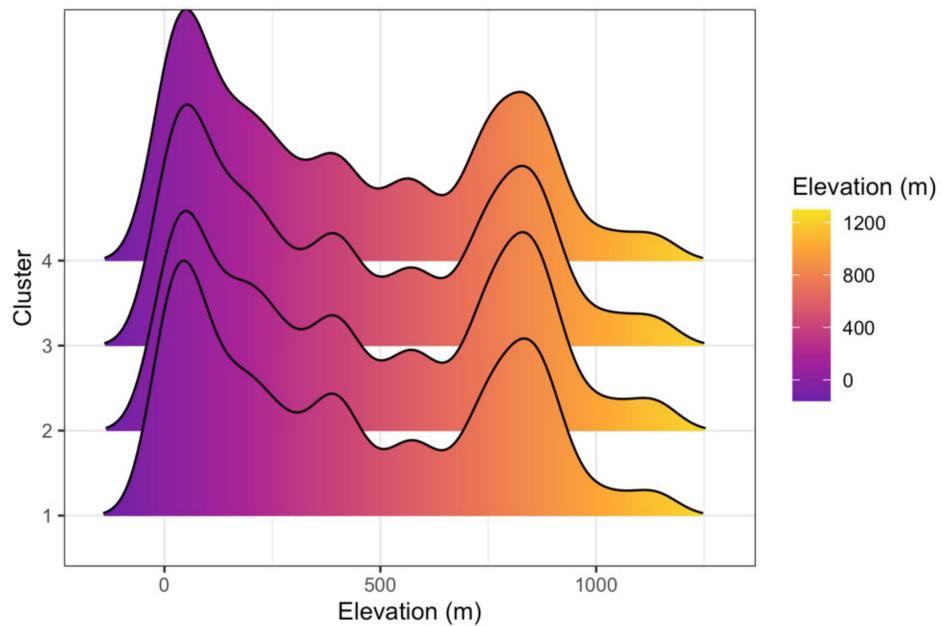


Figure 8. Distribution of elevation values from FCM of every unitless rainstorm given its cluster.

In order to examine the temporal patterns in every station using the results from FCM, correlation matrices were computed using Pearson's r coefficient [67]:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \quad (11)$$

where x_i is the average values of unitless rainstorms coming from a station x and cluster c , y_i , similarly, is the average values of unitless rainstorms coming from a station y and the same cluster c , $n = 100$ (the ordinates of unitless rainstorms), s_x and s_y are the standard deviations of x_i and y_i and, finally, \bar{x} and \bar{y} the mean values, respectively. The patterns per station and given cluster showed very high similarity with $r \geq 0.974$ and on average for all values $r = 0.999$. These results indicate that despite the rich topography and the variability of micro-climates in Greece, the clusters have regional stability in long distances.

In the sequel (Figure 9), the cumulative values of the centers of the four clusters are compared to the ones calculated using Huff's method of classification that is based on the quartile with the highest intensity. The two methods produce different results, while fuzzy clustering creates unitless cumulative curves that do not overlap. The four Huff curves were also tested, for all possible pairs, whether or not are drawn from the same distribution using the Kolmogorov–Smirnov test and three pairs failed to reject that hypothesis for both significance levels $\alpha = 5\%$ and $\alpha = 10\%$.

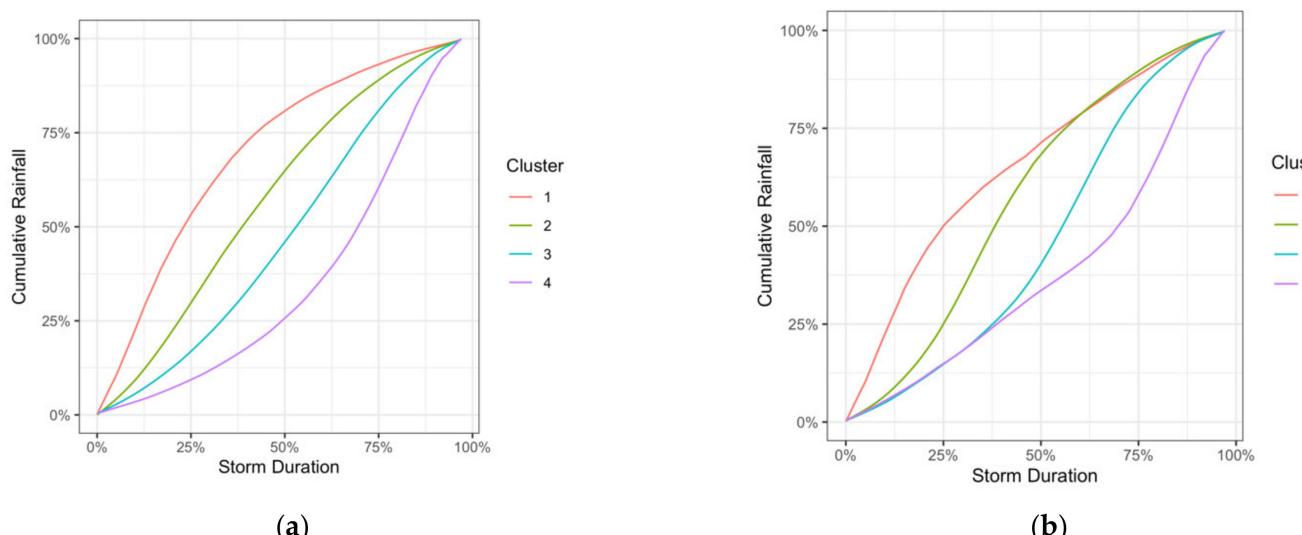


Figure 9. (a) Unitless cumulative values using FCM. (b) Unitless cumulative curves using Huff's classification.

Finally, the topographic maps utilizing the generalized U-matrix using ESOM, as described in Section 2.7, using both FCM and Huff's classification results were produced (Figure 10). In both maps, the unitless rainstorms with membership $>80\%$ are used to remove equivocation from the data. The visualization indicates the presence of four clusters and that Huff's method misclassifies the data. The topographic map produced valleys with blue and green color (indicating smaller distances of the points) that are separated by ridges of brown and white color (representing larger distances).

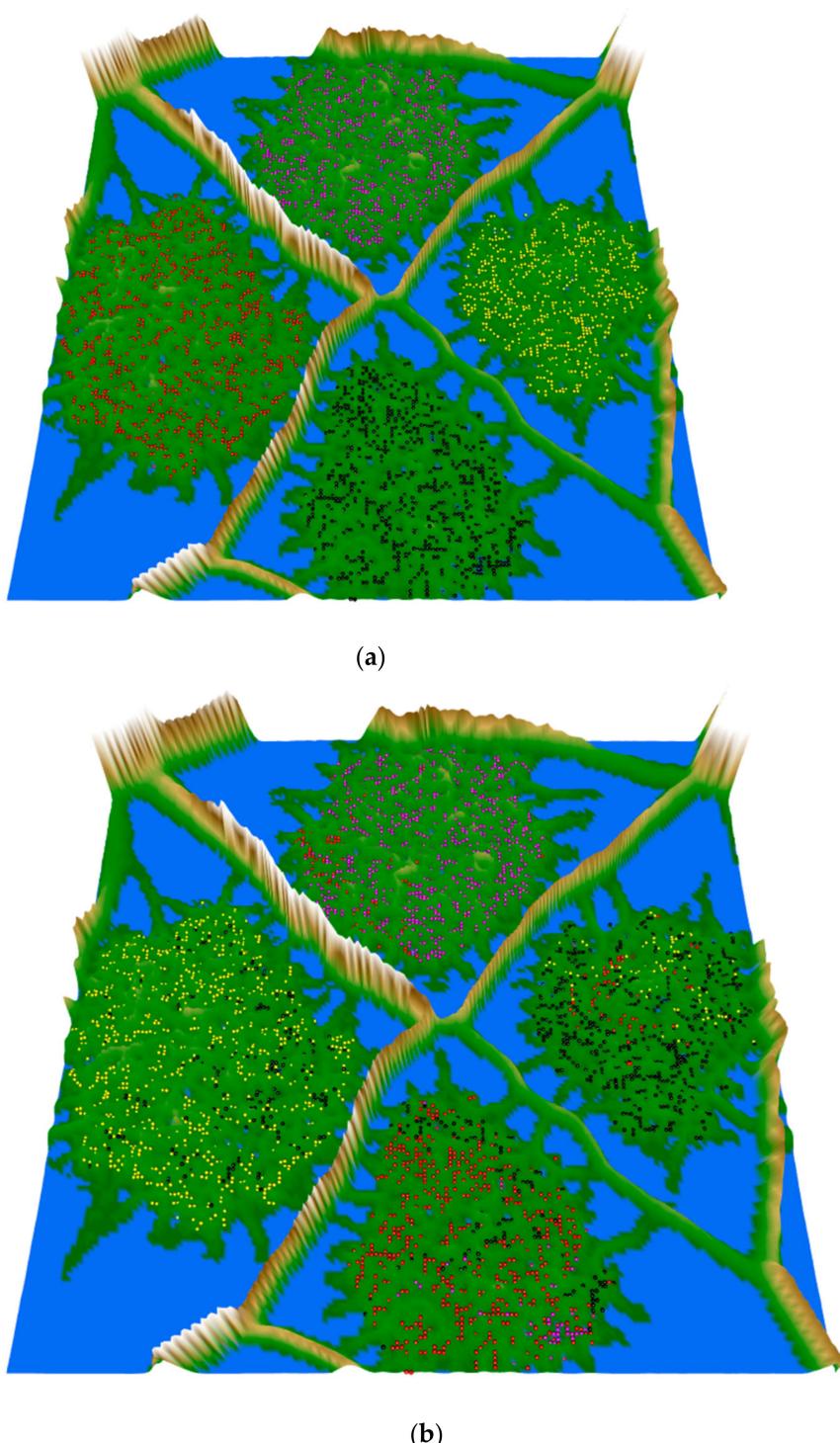


Figure 10. Topographic maps with hypsometric tints using t stochastic neighbor embedding (t-SNE) and the generalized U-matrix. The topology of the map is toroidal (the top and bottom and as well the right and left are joined). Each colored sphere represents a rainstorm with its different classification, as it has already computed. (a) Results using FCM: the rainstorms that belong to the same cluster are separated by mountain ranges. (b) Results using Huff's classification: the rainstorms of different classification are mixed, despite the visual separation of mountain ridges in the map.

4. Conclusions

In this paper, the timeseries of rainfall data were processed in order to detect patterns. The data, coming from the Greek National Bank of Hydrological and Meteorological

Information, cover the totality of the country's regions. In the various stages of the project, novel methods were presented and implemented. First, extraction of rainstorms was executed via a stochastic method. Then, the timeseries were properly transformed to unitless form. The data were tested for clustering tendency prior to the application of clustering algorithms. An iterative form of fuzzy clustering was presented that does not assume a priori knowledge of the number of clusters. It consists of a repeated execution of fuzzy c-means clustering in combination with statistical testing for the choice of the relevant number of clusters. Finally, a verification of the clustering results and a comparison to the widely used Huff's method was attained through topographic maps. The overall approach extends previous rainfall pattern recognition efforts of the authors and of the literature in the sense of the unsupervised learning and in the direction of fuzzy analysis in rainfall and in any follow-up study of floods or water management problems.

Author Contributions: Conceptualization, methodology and software K.V.; writing—original draft preparation, K.V. and E.S.; writing—review and editing, E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Precipitation data are available from the Greek National Bank of Hydrological and Meteorological Information.

Acknowledgments: The data importing, analysis and presentation were done using the open source R language for statistical computing and graphics [68] using the packages: hydroscoper [43], hyetor [69], e1071 [70], FCPS [71], GeneralizedUmatrix [63], factoextra [72] and ggplot2 [73].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Haan, C.T.; Allen, D.M.; Street, J.O. A Markov Chain Model of Daily Rainfall. *Water Resour. Res.* **1976**, *12*, 443–449. [[CrossRef](#)]
2. Gao, C.; Xu, Y.-P.; Zhu, Q.; Bai, Z.; Liu, L. Stochastic Generation of Daily Rainfall Events: A Single-Site Rainfall Model with Copula-Based Joint Simulation of Rainfall Characteristics and Classification and Simulation of Rainfall Patterns. *J. Hydrol.* **2018**, *564*, 41–58. [[CrossRef](#)]
3. Urdiales, D.; Meza, F.; Gironás, J.; Gilabert, H. Improving Stochastic Modelling of Daily Rainfall Using the ENSO Index: Model Development and Application in Chile. *Water* **2018**, *10*, 145. [[CrossRef](#)]
4. Onof, C.; Wang, L.-P. Modelling Rainfall with a Bartlett–Lewis Process: New Developments. *Hydrol. Earth Syst. Sci.* **2020**, *24*, 2791–2815. [[CrossRef](#)]
5. Rodriguez-Iturbe, I.; De Power, B.F.; Valdés, J.B. Rectangular Pulses Point Process Models for Rainfall: Analysis of Empirical Data. *J. Geophys. Res. Atmos.* **1987**, *92*, 9645–9656. [[CrossRef](#)]
6. Burton, A.; Fowler, H.J.; Blenkinsop, S.; Kilsby, C.G. Downscaling Transient Climate Change Using a Neyman–Scott Rectangular Pulses Stochastic Rainfall Model. *J. Hydrol.* **2010**, *381*, 18–32. [[CrossRef](#)]
7. Onof, C.; Chandler, R.E.; Kakou, A.; Northrop, P.; Wheater, H.S.; Isham, V. Rainfall Modelling Using Poisson-Cluster Processes: A Review of Developments. *Stoch. Environ. Res. Risk Assess.* **2000**, *14*, 384–411. [[CrossRef](#)]
8. Rodriguez-Iturbe, I.; Cox, D.R.; Isham, V. Some Models for Rainfall Based on Stochastic Point Processes. *Proc. R. Soc. Lond. Math. Phys. Sci.* **1987**, *410*, 269–288.
9. Rodriguez-Iturbe, I.; Cox, D.R.; Isham, V. A Point Process Model for Rainfall: Further Developments. *Proc. R. Soc. Lond. Math. Phys. Sci.* **1988**, *417*, 283–298.
10. Brigandì, G.; Aronica, G.T. Generation of Sub-Hourly Rainfall Events through a Point Stochastic Rainfall Model. *Geosciences* **2019**, *9*, 226. [[CrossRef](#)]
11. Vandenberghe, S.; Verhoest, N.E.C.; Buyse, E.; De Baets, B. A Stochastic Design Rainfall Generator Based on Copulas and Mass Curves. *Hydrol. Earth Syst. Sci.* **2010**, *14*, 2429–2442. [[CrossRef](#)]
12. Huff, F.A. Time Distribution of Rainfall in Heavy Storms. *Water Resour. Res.* **1967**, *3*, 1007–1019. [[CrossRef](#)]
13. Bonta, J. Development and Utility of Huff Curves for Disaggregating Precipitation Amounts. *Appl. Eng. Agric.* **2004**, *20*, 641. [[CrossRef](#)]
14. Yin, S.; Xie, Y.; Nearing, M.A.; Guo, W.; Zhu, Z. Intra-Storm Temporal Patterns of Rainfall in China Using Huff Curves. *Trans. ASABE* **2016**, *59*, 1619–1632.

15. Loukas, A.; Quick, M.C. Spatial and Temporal Distribution of Storm Precipitation in Southwestern British Columbia. *J. Hydrol.* **1996**, *174*, 37–56. [[CrossRef](#)]
16. Dunkerley, D. Identifying Individual Rain Events from Pluviograph Records: A Review with Analysis of Data from an Australian Dryland Site. *Hydrol. Process. Int. J.* **2008**, *22*, 5024–5036. [[CrossRef](#)]
17. Yu, R.; Xu, Y.; Zhou, T.; Li, J. Relation between Rainfall Duration and Diurnal Variation in the Warm Season Precipitation over Central Eastern China. *Geophys. Res. Lett.* **2007**, *34*. [[CrossRef](#)]
18. USDA-ARS. *Science Documentation: Revised Universal Soil Loss Equation, Version 2 (RUSLE 2)*; USDA-Agricultural Research Service: Washington, DC, USA, 2013.
19. Restrepo-Posada, P.J.; Eagleson, P.S. Identification of Independent Rainstorms. *J. Hydrol.* **1982**, *55*, 303–319. [[CrossRef](#)]
20. Wang, W.; Yin, S.; Xie, Y.; Nearing, M.A.; Wang, W.; Yin, S.; Xie, Y.; Nearing, M.A. Minimum Inter-Event Times for Rainfall in the Eastern Monsoon Region of China. *Trans. ASABE* **2019**, *62*, 9–18. [[CrossRef](#)]
21. Vantas, K.; Sidiropoulos, E.; Loukas, A. Robustness Spatiotemporal Clustering and Trend Detection of Rainfall Erosivity Density in Greece. *Water* **2019**, *11*, 1050. [[CrossRef](#)]
22. Hartigan, J.A. *Clustering Algorithms*; John Wiley & Sons: New York, NY, USA, 1975.
23. Sheikholeslami, G.; Chatterjee, S.; Zhang, A. Wavecluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In Proceedings of the 24th International Conference on Very Large Databases (VLDB), New York, NY, USA, 24–27 August 1998; Volume 98, pp. 428–439.
24. Nayak, J.; Naik, B.; Behera, H. Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014. *Comput. Intell. Data Min. Vol.* **2015**, *2*, 133–149.
25. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Academic Press: Burlington, MA, USA, 2009; ISBN 978-1-59749-272-0.
26. Milligan, G.W.; Cooper, M.C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **1985**, *50*, 159–179. [[CrossRef](#)]
27. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **2014**, *61*, 1–36. [[CrossRef](#)]
28. Dikbas, F.; Firat, M.; Koc, A.C.; Gungor, M. Classification of Precipitation Series Using Fuzzy Cluster Method. *Int. J. Climatol.* **2012**, *32*, 1596–1603. [[CrossRef](#)]
29. Zeybekoglu, U.; Keskin, A.Ü. Defining Rainfall Intensity Clusters in Turkey by Using the Fuzzy C-Means Algorithm. *Geofizika* **2020**, *37*, 181–195. [[CrossRef](#)]
30. Hsu, K.-C.; Li, S.-T. Clustering Spatial-Temporal Precipitation Data Using Wavelet Transform and Self-Organizing Map Neural Network. *Adv. Water Resour.* **2010**, *33*, 190–200. [[CrossRef](#)]
31. Lana, X.; Rodríguez-Solà, R.; Martínez, M.; Casas-Castillo, M.; Serra, C.; Burgueño, A. Characterization of Standardized Heavy Rainfall Profiles for Barcelona City: Clustering, Rain Amounts and Intensity Peaks. *Theor. Appl. Climatol.* **2020**, *142*, 255–268. [[CrossRef](#)]
32. Nojumuddin, N.S.; Yusof, F.; Yusop, Z. Identification of Rainfall Patterns in Johor. *Appl. Math. Sci.* **2015**, *9*, 1869–1888. [[CrossRef](#)]
33. Bonta, J.V. Stochastic Simulation of Storm Occurrence, Depth, Duration and within-Storm Intensities. *Trans. ASAE* **2004**, *47*, 1573–1584. [[CrossRef](#)]
34. Wu, S.-J.; Yang, J.-C.; Tung, Y.-K. Identification and Stochastic Generation of Representative Rainfall Temporal Patterns in Hong Kong Territory. *Stoch. Environ. Res. Risk Assess.* **2006**, *20*, 171–183. [[CrossRef](#)]
35. Williams-Sether, T. *Empirical, Dimensionless, Cumulative-Rainfall Hyetographs Developed from 1959–1986 Storm Data for Selected Small Watersheds in Texas*; US Geological Survey: Reston, VA, USA, 2004.
36. Azli, M.; Rao, A.R. Development of Huff Curves for Peninsular Malaysia. *J. Hydrol.* **2010**, *388*, 77–84. [[CrossRef](#)]
37. Amponsah, A.O.; Daraio, J.A.; Khan, A.A. Implications of Climatic Variations in Temporal Precipitation Patterns for the Development of Design Storms in Newfoundland and Labrador. *Can. J. Civ. Eng.* **2019**, *46*, 1128–1141. [[CrossRef](#)]
38. Zeimetz, F.; Schaefler, B.; Artigue, G.; Hernández, J.G.; Schleiss, A.J. Swiss Rainfall Mass Curves and Their Influence on Extreme Flood Simulation. *Water Resour. Manag.* **2018**, *32*, 2625–2638. [[CrossRef](#)]
39. Bezak, N.; Šraj, M.; Rusjan, S.; Mikoš, M. Impact of the Rainfall Duration and Temporal Rainfall Distribution Defined Using the Huff Curves on the Hydraulic Flood Modelling Results. *Geosciences* **2018**, *8*, 69. [[CrossRef](#)]
40. Jiang, P.; Yu, Z.; Gautam, M.R.; Yuan, F.; Acharya, K. Changes of Storm Properties in the United States: Observations and Multimodel Ensemble Projections. *Glob. Planet. Chang.* **2016**, *142*, 41–52. [[CrossRef](#)]
41. Vantas, K.; Sidiropoulos, E.; Vafeiadis, M. Optimal Temporal Distribution Curves for the Classification of Heavy Precipitation Using Hierarchical Clustering on Principal Components. *Glob. NEST J.* **2019**, *21*, 530–538.
42. Vantas, K.; Sidiropoulos, E.; Vafeiadis, M. A Data Driven Approach for the Temporal Classification of Heavy Rainfall Using Self-Organizing Maps. In Proceedings of the EGU General Assembly 2019, Vienna, Austria, 7–12 April 2019; Volume 21, p. 1.
43. Vantas, K. Hydroscoper: R Interface to the Greek National Data Bank for Hydrological and Meteorological Information. *J. Open Source Softw.* **2018**, *3*, 625. [[CrossRef](#)]
44. Babu, G.J.; Rao, C.R. Goodness-of-Fit Tests When Parameters Are Estimated. *Sankhyā Indian J. Stat.* **2004**, *66*, 63–74.
45. Vantas, K.; Sidiropoulos, E.; Vafeiadis, M. Rainfall Temporal Distribution in Thrace by Means of an Unsupervised Machine Learning Method. In Proceedings of the Protection and Restoration of the Environment XIV, Thessaloniki, Greece, 3–6 July 2018; Volume 1, pp. 555–564.

46. Bezdek, J.C.; Hathaway, R.J. VAT: A Tool for Visual Assessment of (Cluster) Tendency. In Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02 (Cat. No.02CH37290). Honolulu, HI, USA, 12–17 May 2002; pp. 2225–2230.
47. Hopkins, B.; Skellam, J.G. A New Method for Determining the Type of Distribution of Plant Individuals. *Ann. Bot.* **1954**, *18*, 213–227. [[CrossRef](#)]
48. Husson, F.; Lê, S.; Pagès, J. *Exploratory Multivariate Analysis by Example Using R*, 2nd ed.; Chapman and Hall/CRC: New York, NY, USA, 2017; ISBN 978-0-429-22543-7.
49. Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57. [[CrossRef](#)]
50. Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Springer: Boston, MA, USA, 1981; ISBN 978-1-4757-0452-5.
51. Huang, M.; Xia, Z.; Wang, H.; Zeng, Q.; Wang, Q. The Range of the Value for the Fuzzifier of the Fuzzy C-Means Algorithm. *Pattern Recognit. Lett.* **2012**, *33*, 2280–2284. [[CrossRef](#)]
52. Oppel, H.; Fischer, S. A New Unsupervised Learning Method to Assess Clusters of Temporal Distribution of Rainfall and Their Coherence with Flood Types. *Water Resour. Res.* **2020**, *56*. [[CrossRef](#)]
53. Lin, G.-F.; Wu, M.-C. A SOM-Based Approach to Estimating Design Hyetographs of Ungauged Sites. *J. Hydrol.* **2007**, *339*, 216–226. [[CrossRef](#)]
54. Ester, M.; Kriegel, H.-P.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96*; AAAI Press: Palo Alto, CA, USA, 1996.
55. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons: New York, NY, USA, 1980.
56. Bonta, J.V.; Rao, A.R. Regionalization of Storm Hyetographs. *JAWRA J. Am. Water Resour. Assoc.* **1989**, *25*, 211–217. [[CrossRef](#)]
57. Bonta, J.V.; Shahalam, A. Cumulative Storm Rainfall Distributions: Comparison of Huff Curves. *J. Hydrol.* **2003**, *42*, 65–74.
58. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [[CrossRef](#)]
59. Ultsch, A. *U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data*; Technical Report, Nr. 36; University of Marburg, Department of Computer Science: Marburg, Germany, 2003.
60. Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
61. Ultsch, A.; Mörschen, F. *ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM*; Technical Report, Nr. 46; University of Marburg, Department of Computer Science: Marburg, Germany, 2003.
62. Dasgupta, S.; Gupta, A. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms* **2003**, *22*, 60–65. [[CrossRef](#)]
63. Thrun, M.C.; Ultsch, A. Uncovering High-Dimensional Structures of Projections from Dimensionality Reduction Methods. *MethodsX* **2020**, *7*, 101093. [[CrossRef](#)]
64. Thrun, M.C.; Lerch, F. Visualization and 3D Printing of Multivariate Data of Biomarkers. In Proceedings of the 24th Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 6 March 2016.
65. Vantas, K.; Sidiropoulos, E.; Evangelides, C. Rainfall erosivity and its estimation: Conventional and machine learning methods. In *Soil Erosion-Rainfall Erosivity and Risk Assessment*; IntechOpen: London, UK, 2019.
66. Koutsogiannis, D. A Stochastic Disaggregation Method for Design Storm and Flood Synthesis. *J. Hydrol.* **1994**, *156*, 193–225. [[CrossRef](#)]
67. Hensel, D.R.; Hirsch, R.M. *Statistical Methods in Water Resources*; Book 4, Hydrologic Analysis and Interpretation; U.S. Geological Survey: Reston, VA, USA, 2002.
68. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2021.
69. Vantas, K. Hyetor: R Package to Analyze Fixed Interval Precipitation Time Series. Available online: <https://github.com/kvantas/hyetor> (accessed on 20 March 2021).
70. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). TU Wien. Available online: <https://cran.r-project.org/web/packages/e1071> (accessed on 20 March 2021).
71. Thrun, M.C.; Stier, Q. Fundamental Clustering Algorithms Suite. *SoftwareX* **2021**, *13*, 100642. [[CrossRef](#)]
72. Kassambara, A.; Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Available online: <https://cran.r-project.org/web/packages/factoextra/> (accessed on 20 March 2021).
73. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.