


Article

# Application of Artificial Neural Network and Information Entropy Theory to Assess Rainfall Station Distribution: A Case Study from Colombia

Augusto Rafael Garrido-Arévalo <sup>1</sup>, Luis Mauricio Agudelo-Otálora <sup>1</sup>, Nelson Obregón-Neira <sup>2</sup>, Victor Garrido-Arévalo <sup>3</sup>, Edgar Eduardo Quiñones-Bolaños <sup>4</sup>, Parisa Naraei <sup>5</sup>, Mehrab Mehrvar <sup>6</sup> and **Ciro Fernando Bustillo-Lecompte** <sup>7,\*</sup> 

<sup>1</sup> Faculty of Engineering, Universidad de La Sabana, Bogota 140013, Colombia; augustoga@unisabana.edu.co (A.R.G.-A.); mauricio.agudelo@unisabana.edu.co (L.M.A.-O.)

<sup>2</sup> Faculty of Engineering, Pontificia Universidad Javeriana, Bogota 110231, Colombia; nobregon@javeriana.edu.co

<sup>3</sup> Faculty of Engineering, Universidad Tecnológica de Bolívar, Cartagena 131001, Colombia; vgarrido@utb.edu.co

<sup>4</sup> Faculty of Engineering, Universidad de Cartagena, Cartagena 130015, Colombia; equinonesb@unicartagena.edu.co

<sup>5</sup> Department of Computer Science, Ryerson University, Toronto, ON M5B 2K3, Canada; parisa.naraei@ryerson.ca

<sup>6</sup> Department of Chemical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada; mmehrvar@ryerson.ca

<sup>7</sup> Graduate Programs in Environmental Applied Science and Management, Ryerson University, Toronto, ON M5B 2K3, Canada

\* Correspondence: ciro.lecompte@ryerson.ca

Received: 30 May 2020; Accepted: 10 July 2020; Published: 12 July 2020



**Abstract:** An assessment of the rainfall station distribution in the mountainous area of the Regional Autonomous Corporation of Cundinamarca (CAR, for its acronym in Spanish), Colombia, was conducted by applying concepts from information entropy and artificial neural networks (ANNs). This study was divided into two phases: first, a classification of the meteorological stations using two-dimensional self-organizing maps; second, the evaluation of the performance of the ANN by applying concepts of information entropy. Three scenarios were raised for the classification of the meteorological stations by adjusting the number of neurons in the output layer. A high number of neurons in the output layer were obtained, causing the model to over-fit while emphasizing differences amid patterns. When comparing the results of the scenarios, the permanence of certain characteristics and features was found in the system, validating the model classification. Subsequently, the results of the first scenario were used to evaluate the entropy of the historical series. Finally, the results show that the area of study presents a lack of information due to the uncertainty associated with the probabilistic arrangement, which can be corrected with the developed model. Consequently, some recommendations for the redesign of the rainfall are provided.

**Keywords:** hydrology; rainfall; artificial neural networks; information entropy; clustering process

## 1. Introduction

An appropriate rainfall network is fundamental when planning watershed management strategies because it must capture and supply reliable spatial and temporal precipitation data needed for the design, construction, and operation of hydraulic structures such as urban stormwater drainage systems [1]. The design of a rainfall network consists of the determination of the number and location

of stations over a region to obtain a historical record of data that can characterize the phenomenon of precipitation in space and time.

Commonly, estimations of areal rainfall are prepared from point observations and to interpolate precipitation; there is a need to comprehend the entire spatial array and the effect of topography. In mountainous areas with complex rainfall gradients, station density is usually low, and rainfall stations are mostly located in valleys, which causes inaccurate measurements of precipitations compared with the higher landscape [2,3]. That is why the design should incorporate essentially two considerations: first, the knowledge of the physical nature and stochastic meteorological processes and second, the use of the data. Thus, the rainfall network should integrate both the efficiency in data collection and the effectiveness of the measured information.

Different methods are proposed for the design of an appropriate rainfall network, including kriging, Ward's method, self-organizing maps (SOMs), and the K-means method [4]. Kriging is a method that allows for a best linear unbiased estimator (BLUE), with which is the best possible estimate or linear prediction of a variable from the available information and relationships of spatial dependence, minimizing bias and the variance of the estimate [5]. Kriging is also one of the most widely used methods for network optimization. However, the use of a non-linear pattern learning method, such as an artificial neural network (ANN), has been found to yield 15% more reliable results under the same constraints when compared to the conventional kriging method under region and country-wide scenarios [6,7].

The concept of ANNs is inspired by the biological neural networks, which store experiential knowledge. The process of learning is branded as the learning algorithm. The function of the learning algorithm is to vary the synaptic weights of the neural networks to reach a target identified in advance, where the neuron is essential for the functioning of the neural network [8]. ANNs are computer programs that can simulate the behavior of human neural networks with self-learning mechanisms, and besides that, ANNs can generate interactions between memorization and the information itself [9]. ANNs have been used in fields as varied as electricity [10], chemistry [11], chemical engineering [12], and visual pattern recognition [13], and materials engineering [14]. Furthermore, ANNs have also shown positive results in the field of hydraulics and hydrology [15–17].

Kar et al. [18] established different key rainfall networks using SOM, Hall's method, analytical hierarchy process (AHP), and hierarchical clustering (HC) combined with ANN. Wei et al. [19] studied the spatiotemporal scaling effect on rain gauge networks using the entropy concept. Rain gauge prioritization for different spatiotemporal scales was achieved for a reduced number and a low percentage of stations, and the influence of spatial scales was not significant compared to that of temporal scales.

On the other hand, the foundations of the theory of information entropy were established by Shannon [20] and defined as the measure of disorder or peculiarity of certain combinations; it was conceived as a tool for designing communication systems. However, in practice, the ideas of information entropy have been used in different applications to the original target. The concept of information entropy has been used in diverse fields, including economics [21,22], ecology [23], statistical inference [24], agriculture [25], and image processing [26], with positive results that validate its applicability to each sector. Moreover, the entropy of information has been used in the fields of hydraulics and hydrology for different purposes, such as determining a methodology for gauging of rivers [27], regional analysis of precipitation [28], and evaluating precipitation variability in a given area [29]. Furthermore, Liu et al. [16] conducted an entropy-based assessment and rainfall distribution zoning. Although the rain gauge distribution approach has shown to be highly unreliable due to climate change and human activities, its appraisal still has a connotation in understanding the water cycle, and it is critical for the management of water resources. Therefore, the similarities between rainfall gauging stations should be evaluated to categorize precipitation stations into distribution areas. The proper organization of a rainfall network is vital for the adequate allocation and management of water resources to meet increasing water demand [16].

The combined approach of ANN and SOM is recommended for the design of rainfall networks where there are large scale requirements and random criteria for station location, which makes the application of conventional methods not appropriate. This approach is reflected in monitoring stations being located in redundant sites, neglecting other areas. In the case of the studied region of this paper, the lack of a single design criterion is justified in part because stations were installed at different times in the last eight decades. Thus, it is relevant to assess the performance of the rainfall network to determine if there is such redundancy in the location of the monitoring stations. Consequently, the current rainfall stations distribution of 2016, under the jurisdiction of the Regional Autonomous Corporation of Cundinamarca (CAR, for its acronym in Spanish) in Colombia, was evaluated using applied concepts of the information entropy and ANNs to provide recommendations for the redesign of the rainfall network in the studied mountainous region.

## 2. Materials and Methods

### 2.1. Characteristics of the Studied Region

The studied region covers 18,615 km<sup>2</sup>; there are 104 municipalities, of which 98 belong to the Department of Cundinamarca, and six are under the jurisdiction of the Department of Boyacá and the rural area of the City of Bogota, Colombia. Although this territory is mostly for agricultural and livestock use (Soacha, Central Sabana, and West Sabana), there exists a major industrial and mining development plan for Ubaté, Cundinamarca. Figure 1 shows the map of the studied region, which is divided into nine watersheds, including the Sumapaz, Bogota, Magdalena, Black, Minero, Ubaté and Suarez, White, Gacheté, and Macheté rivers. Table 1 lists the corresponding areas for each of these watersheds.

**Table 1.** Watersheds of the study region.

Watershed	Area (km <sup>2</sup> )
Sumapaz River	2527
Bogota River	5671
Magdalena River	2191
Negro River	4239
Minero River	990.8
Ubate-Suarez River	1965
Blanco River	471.0
Gacheté River	97.30
Macheté River	508.7

Topographically, 30% of the study area is located at heights between 2500 and 3000 m above sea level (Figure 2), which forms the so called high mountain area in the region of Bogota Sabana and the Ubaté-Chiquinquirá valley with its adjacent hillsides. On the other hand, about 16% of the area corresponds to the Andean peaks, with heights of 3000 m above sea level [30]. In the study area, the bimodal regime of rains predominates, typical of the Andean region, which extends over the western slope and the highlands of the Eastern Cordillera. On the eastern slope of the mountain range, the rain regime is monomodal.

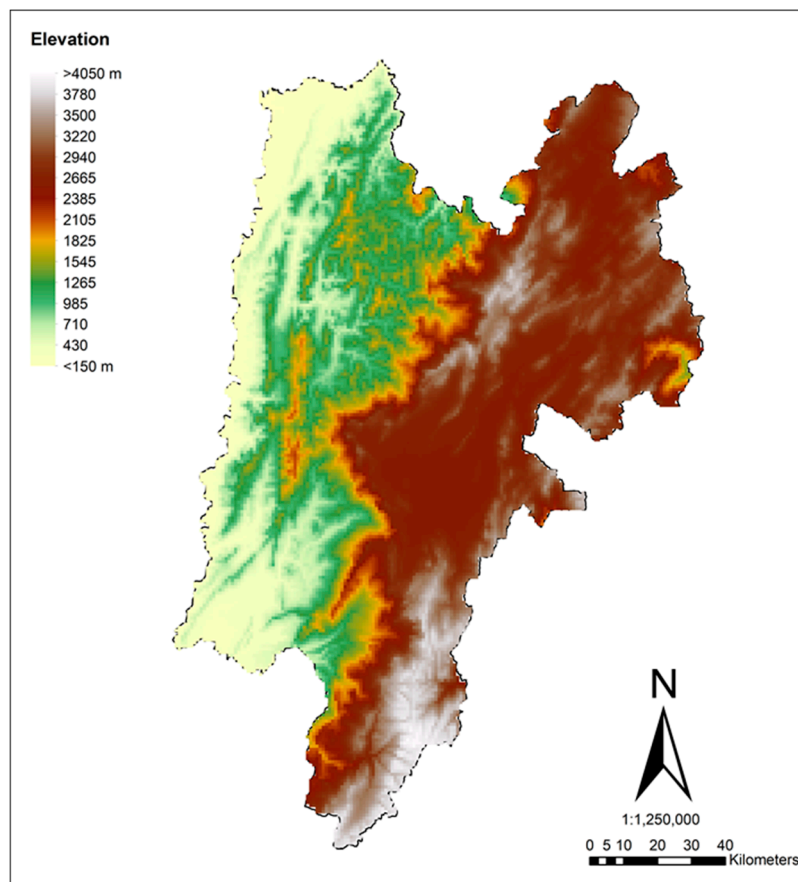
The spatial and temporal variations in precipitation in the study area are governed by three climatological phenomena. The first is the circulation of the atmosphere through the equatorial zone, which, affected by the Intertropical Confluence Zone (ITCZ), is where warm and humid air currents converge from the large high-pressure belts of the southern and northern hemispheres, giving rise to large water-laden clouds. In most of the region, the movement of the ITCZ causes, during the year, a double maximum and a double minimum of rainfall, associated with other meteorological elements during the wet and dry seasons.



Figure 1. Location of the regional autonomous corporation study area in Cundinamarca, Colombia.

The second climatological phenomenon is due to the circulation of air masses arising locally due to thermal differences, which produces cloudiness and precipitation in the upper parts of the valleys and clear skies in the center of them; at night, this phenomenon is reversed. This phenomenon is influenced by the shape and orientation of the terrain, altitude, vegetation, and presence of water [31]. A third phenomenon is that of the southeast trade winds from the Orinoquía (Eastern border of Colombia with Venezuela), which blow with higher intensity from June to September, discharging large amounts

of moisture onto the eastern side of the Cordillera Oriental and cause a maximum rainfall from June to August [32].



**Figure 2.** Terrain heightmap for the study area in Cundinamarca, Colombia.

Regarding the minimum number of rain gauges in the network located in the studied area, the World Meteorological Organization (WMO) proposes a specific number of stations depending on the physiographic unit of the region where the network will be installed. Thus, a minimum density per station of 250 km<sup>2</sup> is recommended for mountainous areas [33].

## 2.2. Methods

The study involved two phases. In the first phase, a classification of the meteorological stations was conducted using two-dimensional SOMs for the studied region. In the second phase, we took advantage of the fact that the entropy of the information cannot only represent the uncertainty of the rainfall distribution, but it can also reflect the correlation and the transmission of information between the rainfall stations [34]; through this, the performance of the rainfall gauge of the Cundinamarca region was evaluated.

### 2.2.1. Meteorological Network Data

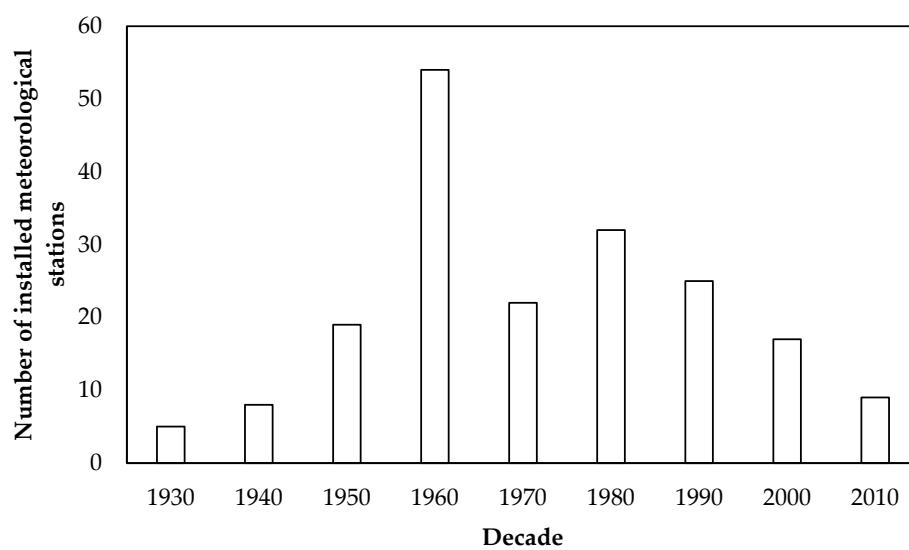
The meteorological network data were gathered from CAR [35] and its area of jurisdiction. The CAR has a historical record of 182 stations for measuring precipitation distribution in the studied area, with historical records since 1931. Some of these stations have been installed recently. Thus, the historical series is not uniform in all cases. Of the total of registered stations, 37% corresponds to rainfall stations, 28% to pluviographic stations, 15% to primary meteorological stations, 14% to secondary meteorological stations, and 5% to the remaining automatic and satellite stations. Gathered



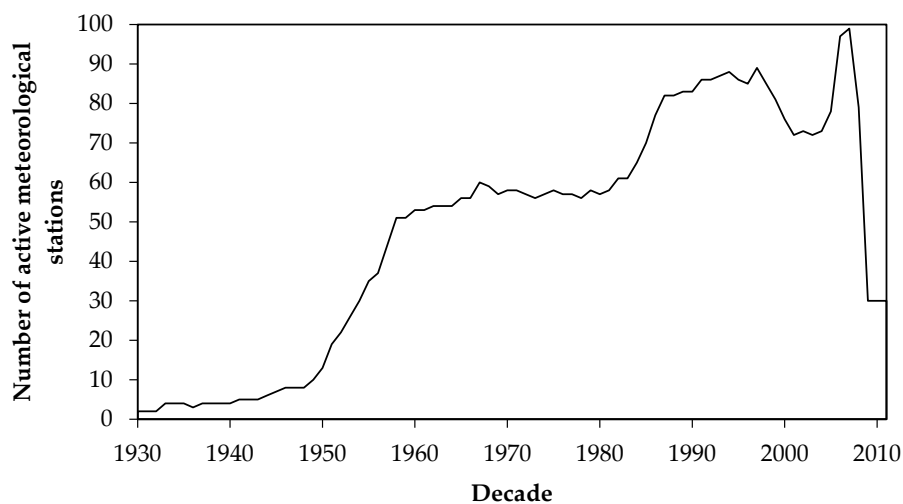
data included the meteorological station locations with coordinates and the historical record of monthly precipitation.

### 2.2.2. Data Processing

Most meteorological stations were installed in the sixties, and the CAR has operated the rainfall network since 1961. There are currently 47 non-active stations and 135 active stations registered under the CAR jurisdiction. Figure 3a depicts the number of stations installed per decade from the 1930s to 2010s. The rainfall records in the studied meteorological stations are not homogeneous in terms of the amount of information. The period with the highest possible number of stations with available data was selected, considering that the meteorological stations in the analyzed network have come into operation at different times. Thus, the stations holding more than 80% of monthly data for not less than twenty years were selected for this study (Figure 3b).



(a)



(b)

**Figure 3.** Meteorological stations installed by the regional autonomous corporation for the study area in Cundinamarca, Colombia. (a) Distributed by number installed per decade; (b) Distributed by active stations.

The monthly records were enumerated, starting with the oldest of the series from January 1927 to the most recent one. They were checked one by one to find out which station had information about the precipitation in every particular month. The obtained information was plotted, as shown in Figure 3b. The period between 1986 and 2008 was found to have the most substantial number of active stations. As a result, ten stations with at least 80% available monthly data were selected for this study.

### 2.2.3. Development of the Artificial Neural Network Model

The classification was performed using the MATLAB Neural Network Toolbox and the SOM approach, which identifies the homogeneous regions with more precision than the K-means and Ward methods, two of the most commonly used classification methods [4]. The networks used are made up of an input layer, in which the input patterns are entered into the model, and an output layer, where the weights of neurons are updated based on the input patterns. Moreover, the output layer is the two-dimensional space that is self-organized based on the structure of the input patterns. Thus, each of the neurons in the input layer connects to all the neurons in the output layer.

Although there are no lateral connections between neurons in the same layer, updating the output layer weights based on the neighborhood of the winning neuron creates a similarity link between nearby neurons that leads to the grouping or self-organization of neurons with similar characteristics.

In this case, a two-dimensional array was made, setting the number of rows and columns of the array of neurons. To accurately differentiate pattern groups, it is recommended to use neuron arrays in the output layer as large as possible. However, it is essential to note that if the number of neurons in the output layer is quite large, the model can over-train and highlight the differences between each of the patterns, yielding the same number of groups as patterns [36].

Consequently, and considering the above, three scenarios were defined with different numbers of neurons in the output layer. The number of iterations is given, according to González-Cuellar [36], by the number of neurons in the output layer multiplied by 500. Finally, with the help of the MATLAB Neural Network Toolbox, a computational application was developed to classify the rainfall stations of the studied region, considering the defined typologies, and the built-in visualization of the results. To prevent the fact that difference of scale between variables affects the classification, these were transformed using the relative change and difference normalization approach so that their ranges were homogeneous and the results comparable [4]. Table 2 shows the applied transformation for each variable to obtain values between zero and one.

**Table 2.** Transformation of variables.

Input Variable	Transformation
Latitude (m)	$y = (x - x_{\min}) / (x_{\max} - x_{\min})$
Longitude (m)	$y = (x - x_{\min}) / (x_{\max} - x_{\min})$
Elevation (m)	$y = x / x_{\max}$
Annual rainfall (mm)	$y = x / x_{\max}$
Standard deviation of annual rainfall (mm)	$y = x / x_{\max}$
Monthly rainfall (mm)	$y = x / x_{\max}$

Large neuron arrays were used in the output layer to differentiate the patterns of groups adequately. However, it is essential to note that if the number of neurons in the output layer is vast, the model may over-fit and highlight the differences between each of the patterns, generating as many groups as patterns [36,37]. Three scenarios with different numbers of neurons in the output layer were defined (100, 400, and 900 neurons for Types 1, 2, and 3, respectively) as per preliminary studies by the authors [17,36]; thus, the number of iterations was estimated by the number of neurons in the output layer multiplied by 500 (50,000, 200,000, and 450,000 iterations for Types 1, 2, and 3, respectively).

As mentioned earlier, a computer application was developed using the MATLAB ANN Toolbox to classify rainfall stations of the CAR based on the classification types defined in each of the three

developed scenarios. Type-1 was based on a Kohonen network model of 100 neurons distributed in an array of ten rows and ten columns. The number of iterations was defined as the number of neurons in the output layer multiplied by 500 (i.e., about 50,000 iterations). Type-2 was based on a Kohonen network model of 400 neurons distributed in an array of 20 rows and 20 columns. Thus, 200,000 iterations were used. Type-3 was based on a Kohonen network model of 900 neurons distributed in an array of 30 rows and 30 columns. Consequently, 450,000 iterations were used. In all three cases, a hexagonal topology, with neurons also in hexagonal shape, was chosen. Thus, neurons that are not on the edges of the Kohonen layer have six neighboring neurons that are connected virtually.

#### 2.2.4. Performance Evaluation of the Rainfall Network in the Cundinamarca Region

The information entropy is a measure of the uncertainty of a specific outcome in a random process [20]. The concept of entropy has been used to investigate the variability associated with monthly, seasonal, and annual series of precipitation, and thus, characterize the precipitation to generate formulations on the efficient management of rainfall water [29]. In a study by Lohani et al. [38], artificial neural networks and the neuro-fuzzy system were used to forecast monthly inflow in a reservoir; results from this study were useful to understand how water supply and flood control measures can be generated from these models.

By using the information entropy concept, this section describes the distribution of information in each of the developed groups. The length of the time series data in each case was established, having no less than 20 year periods and missing data not exceeding 20% [28]. The following equation calculated the marginal entropy for each station:

$$H = - \sum_{k=1}^k p(x_k) \text{Log}[p(x_k)] \quad (1)$$

where  $k$  is the discrete data interval,  $x_k$  is the result corresponding to the interval  $k$ , and  $p(x_k)$  is the probability of  $x_k$ . For each station, an estimated series of predicted values was obtained by multiple linear regression from the data of the meteorological stations in the same group. Thus, the marginal entropy for the estimated series was also calculated by Equation (1). Subsequently, the joint marginal entropy between actual and estimated values for each station was calculated by Equation (2):

$$H(x, y) = - \sum_{k=1}^k \sum_{l=1}^l p(x_k y_l) \text{Log}[p(x_k y_l)] \quad (2)$$

where  $l$  is the discrete data interval for the estimated values,  $y_l$  is the result corresponding to the interval  $l$ , and  $p(x_k y_l)$  is the probability of  $x_k y_l$ . Finally, the mutual information of each station was calculated from the values found previously by applying Equation (2). The mutual information is the amount of information contained in a process to another process [28]. In this case, it corresponds to the rainfall data contained in one station and simultaneously within others in its own group, which is calculated by Equation (3):

$$T(x, y) = H(x) + H(y) - H(x, y) \quad (3)$$

where  $H(x)$ ,  $H(y)$ , and  $H(x,y)$  are the marginal entropy of the actual data, the marginal entropy of the simulated series, and the joint entropy, respectively; stations should have minimum mutual information as possible as the fundamental basis of the monitoring network design using the information entropy approach. Thus, stations must be independent of each other.

Low mutual information values indicate that those stations are more independent and share little information, while high mutual information values represent those with more dependency. Therefore, there may be no need for more stations in that area. It is recommended to install additional rainfall stations where mutual information values are close to zero.



Table 3 shows the classification of mutual information according to the obtained values. This criterion was used to produce the corresponding recommendations for a proper redesign of the rainfall network in the studied area based on precipitation variability and rainfall antecedents, as suggested by Mishra [29] and Chang et al. [39], respectively.

Table 3. Mutual information classification.

Mutual Information Range	Index
0–0.5	High deficit
0.5–1.0	Deficit
1.0–1.5	Acceptable
1.5–2.0	Above average
>2.0	Excess

### 3. Results and Discussion

As mentioned earlier, the studied region encompasses 18,615 km<sup>2</sup>, where 30% is located at altitudes between 2500 and 3000 m above sea level, and 16% of the area is mountainous. The region is divided into nine second-order basins, where the Bogota river basin is the largest. The precipitation has a bimodal behavior, a period of rain and drought. During the period from December to March, rainfall is equivalent to only 19% of the annual average. The rainiest months are October and April, with 15 and 16% of the total annual precipitation, respectively. The driest months are January and February, and rains are distributed throughout the year (142 days on average).

#### 3.1. Classification of Rainfall Stations

As mentioned in Section 2.2.2 Data Processing, ten stations with at least 80% of available monthly data were selected for this study. Table 4 shows the normalized input variables for the selected ten stations and their corresponding normalization.

Table 4. Normalized input variables for selected first ten rainfall stations.

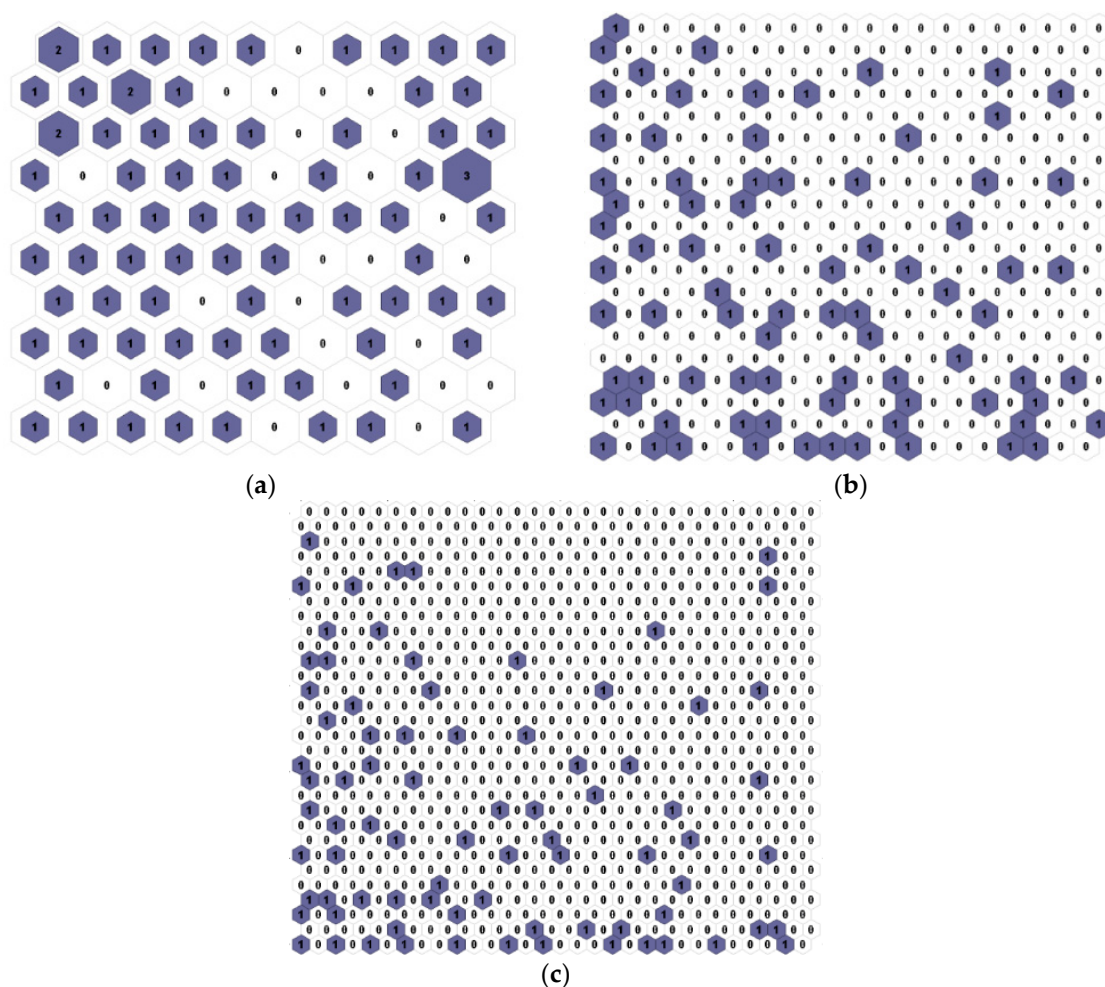
Parameter	Rainfall Station Number										
	S1	S4	S8	S9	S10	S13	S15	S16	S18	S22	
Longitude	0.6582	0.1046	0.1765	0.6240	0.3406	0.4603	0.9768	0.7852	0.6414	0.4045	
Latitude	0.4806	0.0256	0.1008	0.5132	0.3561	0.3755	0.6668	0.7215	0.3214	0.3260	
Elevation	0.7636	0.1150	0.1093	0.8806	0.7688	0.7837	0.8053	0.9002	0.7417	0.7506	
Average monthly rainfall	January	0.3197	0.4796	0.5404	0.2984	0.4240	0.3530	0.3445	0.2358	0.5877	0.3984
	February	0.3681	0.5415	0.5799	0.3155	0.4837	0.4429	0.3097	0.2851	0.5078	0.5936
	March	0.3438	0.5556	0.5726	0.4048	0.4610	0.4513	0.3644	0.3543	0.4520	0.5175
	April	0.2817	0.4985	0.4060	0.2883	0.2691	0.3447	0.2279	0.2684	0.3012	0.3315
	May	0.3257	0.5299	0.4270	0.3777	0.2608	0.3329	0.4092	0.2680	0.2975	0.3670
	June	0.3446	0.2578	0.2759	0.4483	0.2356	0.3174	0.5403	0.2651	0.3430	0.4590
	July	0.4423	0.1983	0.2441	0.5927	0.2765	0.3589	0.6043	0.3653	0.4742	0.5490
	August	0.2833	0.1837	0.2197	0.5957	0.2626	0.3128	0.4148	0.2476	0.3009	0.4912
	September	0.3041	0.4688	0.4408	0.3927	0.3009	0.3613	0.3561	0.2203	0.3655	0.4860
	October	0.4040	0.3501	0.3811	0.3445	0.3192	0.3827	0.3469	0.3007	0.3461	0.4283
	November	0.2871	0.3532	0.4962	0.3086	0.3615	0.3209	0.2984	0.2276	0.3648	0.3564
	December	0.2593	0.3984	0.4038	0.2600	0.3633	0.4133	0.3635	0.3224	0.3810	0.4007
Annual rainfall	0.3286	0.4104	0.4169	0.3749	0.3246	0.3627	0.3683	0.2782	0.3720	0.4310	
Standard deviation of annual rainfall	0.3377	0.5635	0.4935	0.3289	0.3004	0.3365	0.3209	0.2699	0.2519	0.3057	

The input variables used for the classification were the annual rainfall (mm), elevation (m), latitude (m), longitude (m), monthly precipitation (mm), and standard deviation of the annual rainfall in each meteorological station. The values were transformed using the relative change and

difference normalization approach, previously shown in Table 2, for obtaining homogeneous and comparable ranges.

### 3.2. Scenario Configurations

The classification method by the SOM approach offers the advantage of indicating the results on two-dimensional maps regardless of the number of variables included using the map of Hits and the distance between neurons by the SOM neighbor weight distances (U-matrix map). Three scenarios were analyzed, as described in previous sections. In the Type-1 scenario, the winning neurons were identified in the map of Hits shown in Figure 4a, where the number within neurons indicates the number of stations represented (i.e., the number of wins for each neuron). Neurons with a value of zero are those that pose no pattern, in this case, no station. On the other hand, the U-matrix map shows how different a neuron is from another. Consequently, it is possible to identify the groups in which the information is divided.

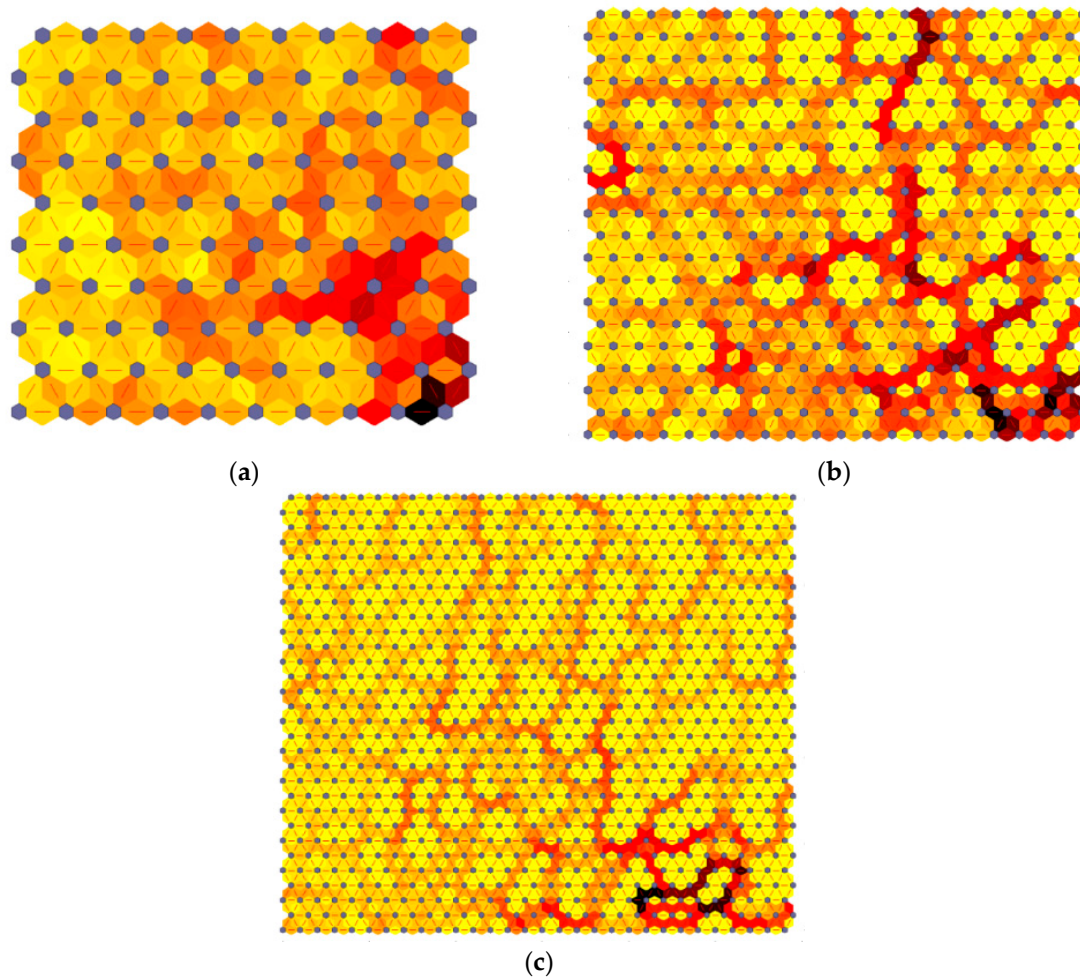


**Figure 4.** Map of Hits for the three evaluated scenarios: (a) Type-1 with 100 neurons and 13 groups, (b) Type-2 with 400 neurons and 47 groups, and (c) Type-3 with 900 neurons and 56 groups.

Figure 5a depicts the distance between neurons map for Type-1, where the presence of darker areas marks the division between the sets shown. Thirteen groups were observed with different numbers of neurons. For this classification, the lines of higher intensity were considered.

Likewise, Figure 4b depicts the map of Hits for the visualization of the results in the Type-2 scenario, where the winning neurons are identified. It is noted that winning neurons are more dispersed in Type-2 than in Type-1, due to the increased number of neurons in the network. Furthermore, the distance

between neurons U-matrix map for Type-2 is provided in Figure 5b. In this case, the demarcation is more noticeable than in Type-1. Fifty groups were distinguished; some of these were formed by not winning neurons, becoming irrelevant to the study. Thus, only 47 groups remained.



**Figure 5.** Self-organizing map of neighbor weight distances for the three evaluated scenarios: (a) Type-1 with 100 neurons and 13 groups, (b) Type-2 with 400 neurons and 47 groups, and (c) Type-3 with 900 neurons and 56 groups.

Finally, the map of Hits for the Type-3 scenario is shown in Figure 4c. It is noted that winning neurons are scattered similarly to the Type-2 case, due to an over-increased number of neurons in the network. This outcome indicates that the groups will consist of fewer stations. Besides, the distance between neurons in the map for the Type-3 scenario shows the presence of darker areas, marking the division between the sets, as shown in Figure 5c. In this case, the demarcation is more noticeable than in Type-1. Sixty-five groups were distinguished. Some of these remained comprised of non-winning neurons and were found irrelevant to the study; thus, only 56 groups were used.

### 3.3. Mutual Information Classification Analysis

The results obtained for the different scenario types, in which at least one neuron is the winner, are 13, 47, and 56 groups formed in Type-1 (100 neurons in the output layer), Type-2 (400 neurons), and Type-3 (900 neurons), respectively. By increasing the number of neurons, the number of groups formed also increases. Therefore, considering a clustering process, a high number of neurons in the output layer can result in the over-training of the model, emphasizing the differences amid patterns; thus, in the Type-3 scenario, with 900 neurons, about 55% of the stations were classified individually.

It is noteworthy that when making a comparative analysis of the different scenarios, specific patterns remain in the classification. A particular case is station S140, which was classified individually in all scenarios, and station S80, which was classified individually in Types 1 and 3.

For the mutual information classification, the input variable to measure the entropy was monthly precipitation, as explained in previous sections. Although the CAR rainfall network has historical records of monthly rainfall since 1931, the length of the series for each of the stations varies according to the installation date. The amount of input data in each group varies based on the available information. Consequently, a minimum of 120 datasets was established to execute the classification procedure; thus, the mutual information classification of each combination of stations for each of the groups formed was performed. The results show that stations present deficit information, indicating that each of them is independent. This outcome makes it unfeasible to rebuild the historical series of precipitation of each station from other stations with which it shares the same group. As a result, groups formed in Type-1 were used, as shown in Table 5.

**Table 5.** Distribution stations by groups.

Group Number	Station List	Number of Grouped Stations
1	S16, S25, S27, S36, S37, S52, S60, S61, S65, S71, S86, S89, S101, S102, S126, S138, S144, S149, S150, S156, S166	21
2	S29, S31, S40, S76, S96, S122, S158	7
3	S140	1
4	S15, S91, S105, S118, S131	5
5	S85, S97, S109	3
6	S80	1
7	S63, S88, S123	3
8	S9, S18, S28, S41, S45, S108, S142, S151, S174	9
9	S44, S64, S77, S78, S124, S130, S159, S163	8
10	S32, S42, S62, S67, S116, S129, S147, S154, S162	9
11	S10, S13, S22, S81	4
12	S4, S8, S55	3
13	S1, S33, S49, S53, S54, S161	6

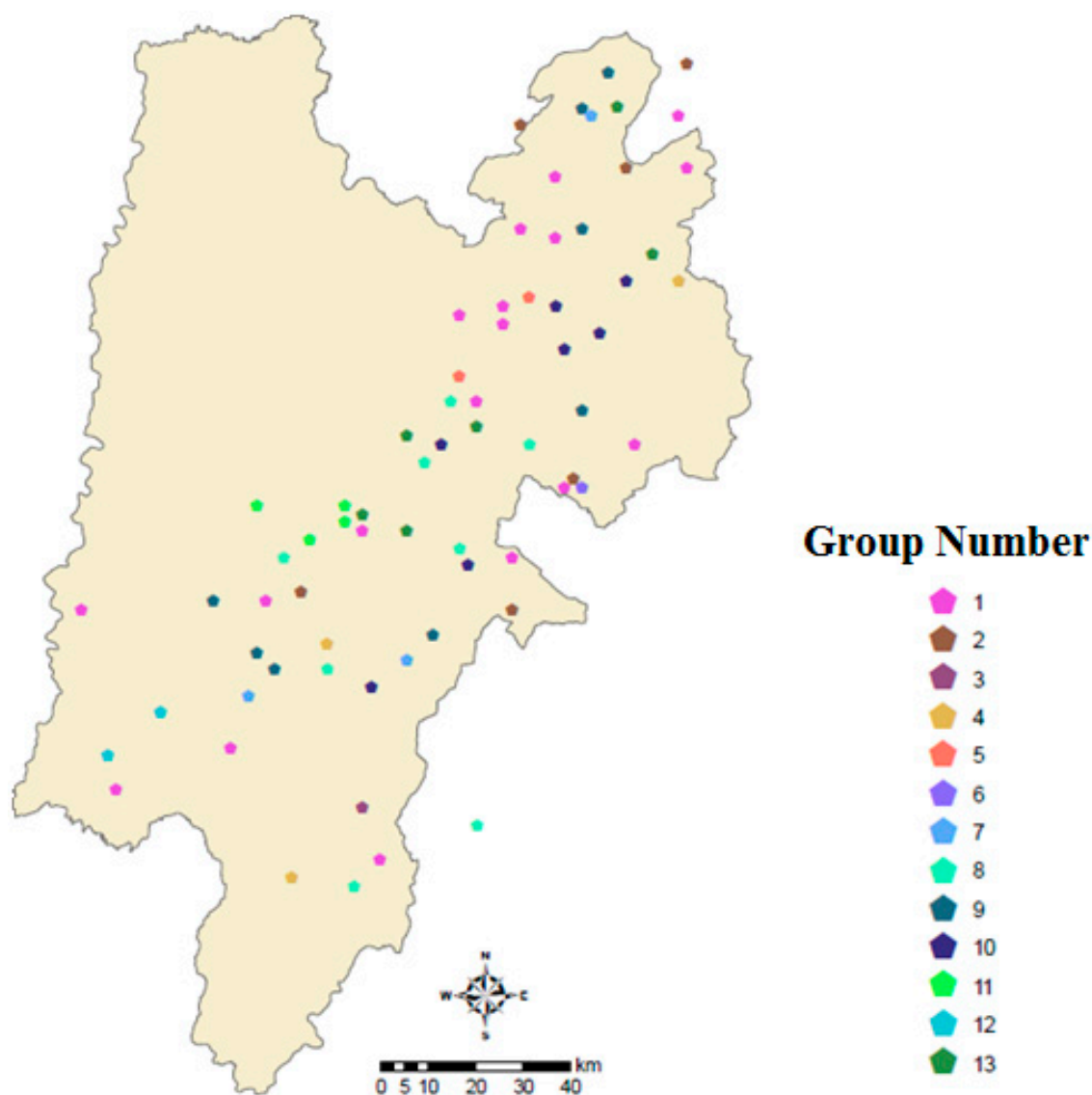
Figure 6 shows a schematic representation of the location of each of the selected groups within the area of CAR jurisdiction. Despite the change in the number of neurons in the output layer, specific patterns were found in the classification, which ensures the existence of similar configurations in the groups obtained. This is the case of the stations S142 (La Casita) and S28 (Santa Teresa), located near the municipality of La Calera, which were classified in the same group in each of the three scenarios.

It is recommended to include this correlation in future rainfall network studies considering a possible reconstruction of the historical series using information from one station to the other. Thus, this allows the relocation of one of the stations to places with fewer stations. The same situation occurs between stations S42 (San Jorge) and S129 (Doña Juana), S76 (Monserrate) and S122 (Esclusa), as well as S16 (Tres Esquinas) and S101 (El Hato No. 2), which were classified into similar groups in different types. Conversely, stations S80 (Las Margaritas) and S140 (Central No. 2) were classified individually. Thus, their information could not be reconstructed from other nearby stations. Consequently, it is recommended that the Regional Autonomous Corporation of Cundinamarca, in the case of a future re-engineering of the network, should not relocate these stations since their information is unique, with valuable historical records since 1959.

It was found that the northeastern part (corresponding to the municipalities of Yacopí and Puerto Salgar) and the south of the studied area (corresponding to the municipality of Cabrera) have low coverage of meteorological stations, so it is recommended to increase the number of stations in this area. In this case, the CAR should evaluate transferring some stations classified under the same group in all three scenarios, such as S16, S28, S42, S76, S101, S122, S129, and S142. Moreover, despite the high



number of stations in the central area of the studied area, the transfer of stations different to those mentioned above is not recommended since, with this method, it is not possible to reconstruct the information that each of these stations provides to neighboring stations.



**Figure 6.** Geographical distribution of groups for the study area in Cundinamarca, Colombia.

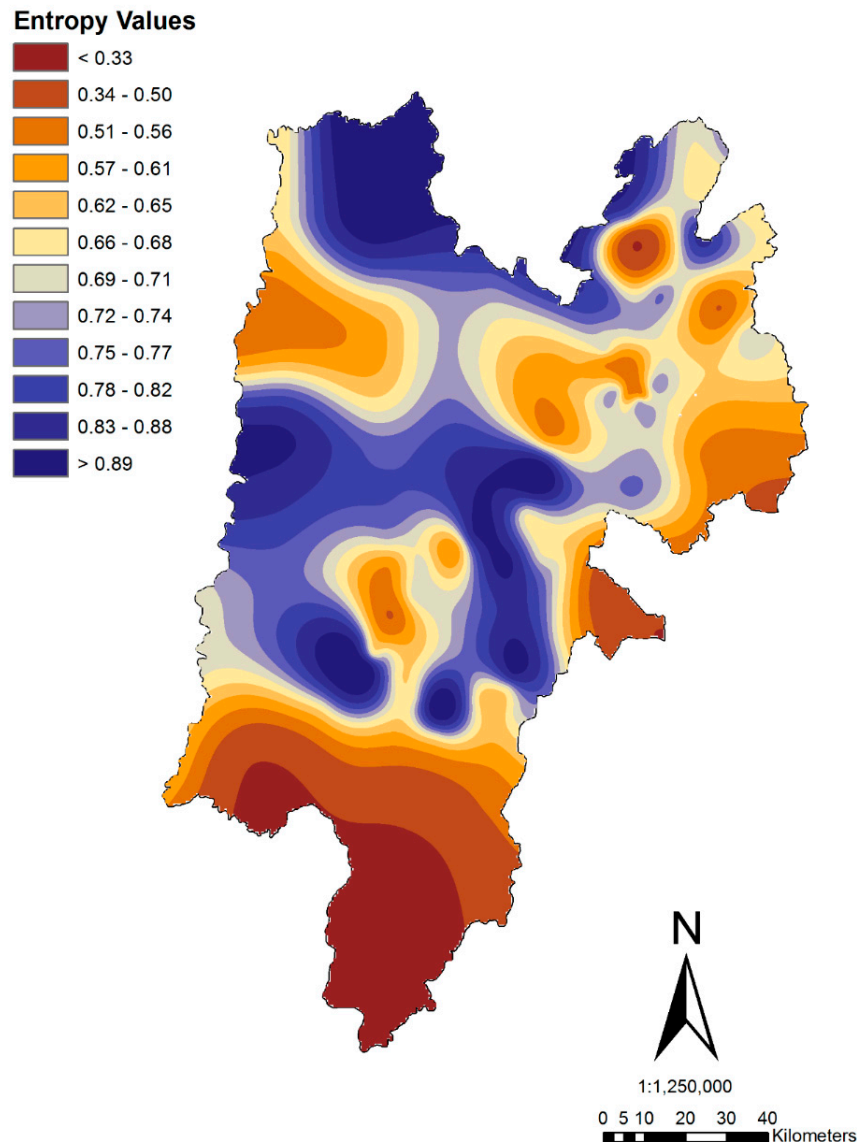
Applied spatial analyst tools were used to construct Figures 7 and 8, where the images show interpolated surfaces calculated using the kriging method within ArcMap software. As discussed earlier, kriging is a method of applied spatial analysis that allows the estimation of values in unsampled locations using the information provided by the sample [4–7]. Figure 7 shows the distribution of mutual information in terms of entropy from values of precipitation of 80 stations within the studied region from 1986 and onwards.

It is noted that the variation of mutual information shown in Figure 7 is not homogeneous, presenting higher values in the center and north of the study area ( $>0.89$ ). These results may be justified for the regions with a large number of stations; however, the values found for mutual information are deficient.

As mentioned in Section 2.1 Characteristics of the studied region, the complex topography typical of the Andean region can cause a considerable variation in precipitation; thus, when comparing nearby



stations, low values for mutual information are obtained. Consequently, it can be stated that the criterion of proximity does not determine the homogeneity of the records.



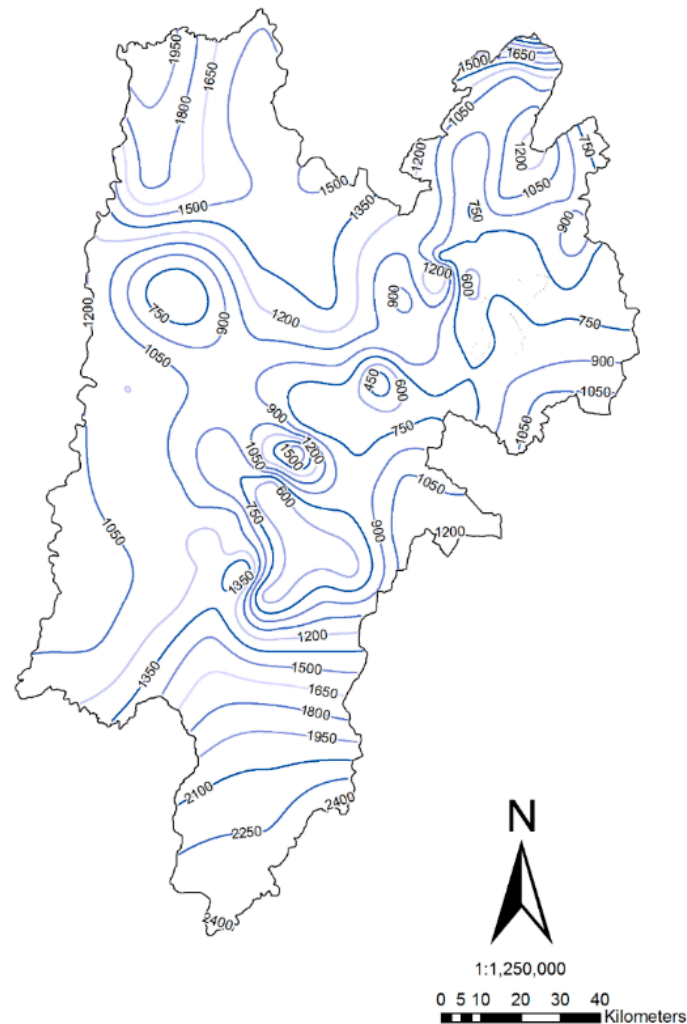
**Figure 7.** Mutual information based on entropy values for the study area in Cundinamarca, Colombia.

It should be noted here that the topographic factor is of great importance to explain the spatial variability of rain in the study area [31], considering that 30% of the study area is located in heights between 2500 and 3000 m above sea level, which makes up the so-called High Mountain area. In comparison, 16% of the area corresponds to the so called Andean peaks (3000 m above sea level).

Figure 7 also shows that the rainfall gauge in the studied area presents information deficiency. Only two stations, E32 (Santa Isabel) and E154 (Campobello), yielded acceptable mutual information values (between 1 and 1.5), which are in the remote municipalities of Tabio and Madrid, respectively. However, it is not recommended to relocate one of these stations based on this analysis but to expand the network coverage, especially in the northeastern and southern parts of the studied area.

Figure 8 shows the map of isohyets corresponding to the distribution of the annual precipitation values in the studied area, which is characterized by the influence of geography of the Andean region. By analyzing the map of isohyets, a heterogeneous behavior and development of a gradient in the South–North direction are identified, with increasing cumulative annual rainfall values ranging from 1200 to 2400 mm. This map of isohyets will support future estimation of design rainfalls, especially

in ungauged areas in Colombia, where the lack of readily available and processed information often becomes an obstacle in the development of hydrological studies [40]; thus, the predictive capabilities of data-driven modeling applied to hydrology are demonstrated [41].



**Figure 8.** Map of isohyets for the study area in Cundinamarca, Colombia.

The heterogeneous behavior of the precipitation and the presence of gradients observed in Figure 8 correspond to the mutual information values shown in Figure 7. This relationship can be established considering that heterogeneity in the distribution of precipitation allows for different behavior from their neighboring stations in the historical record.

As a result, deficient values of mutual information are found when comparing one station to another. It is important to note, as stated above, that the complex Andean topography turns out to be decisive on the distribution of the rainfall, and hence, of the values of the mutual information.

The SOM approach was used to identify homogeneous areas from the time series of precipitation. This decision is justified in that this method has been previously compared with two of the most commonly used methods for classification—the methods of Ward and K-means—through experimental design.

All three methods were tested by experimental datasets, where both the number of groups and their members were previously known. Accuracy values of the groups obtained by the SOM method were 100%, whereas the accuracy values for the K-means' and Ward methods were 97% and 95%, respectively. This comparison ensures that the SOM approach determines the exact number of groups and elements belonging to them; thus, the results are validated [4]. Such comparison has

also been the objective of other studies, in which the SOM approach allows the identification of the homogeneous regions with more precision than the most used classification methods, the K-means and Ward methods [42].

#### 4. Conclusions

By combining self-organizing maps and the concept of entropy of information, it was possible to evaluate the distribution of the rainfall stations network of the Cundinamarca region. The application of ANNs to classify rainfall stations in the studied region showed that such stations could be grouped into 13, 47, and 56 groups, depending on the number of neurons in the output layer.

Three scenarios were raised for the classification of the meteorological stations by varying the number of neurons in the output layer. Results show that increasing the number of these neurons increased the number of groups formed. In Type-1 with 100 neurons in the output layer, 13 groups were obtained. In Type-2 with 400 neurons in the output layer, 47 groups were obtained, and finally, in Type-3 with 900 neurons in the output layer, 56 groups were obtained. Consequently, in a clustering process, a high number of neurons in the output layer can over-train the model.

It was expected that the best values of mutual information were present at stations in the central zone of the studied area for its high density. However, the results show that the same trend was present in the rest of the studied area with deficit values. Thus, it can be concluded that the criterion of proximity between stations does not guarantee the homogeneity of the information provided; it would be unwise to transfer any of these stations under this criterion.

Because of the low mutual information values, it is recommended to integrate the available information from the CAR with other entities, such as the Institute of Hydrology, Meteorology and Environmental Studies of Colombia (IDEAM, for its acronym in Spanish) to complement the data from the stations in the same area to expand network coverage. Detailed maps of isohyets such as the one shown in this study can be used for both regional and global hydrological studies to develop water management strategies, considering the lack of readily available processed data in Colombia.

Finally, it is recommended to extend the coverage of the rainfall network in the northeastern and southern parts of the studied area using a new group of stations with similar characteristics based on the results of the classification by ANNs. However, this relocation must be assessed in detail because the values of mutual information obtained for the same stations are deficient. Consequently, it is recommended to assess, in a future study, the mutual information of these stations with their nearest neighbors. If in such a case, there were acceptable values of mutual information, the transfer would be feasible.

**Author Contributions:** Conceptualization, A.R.G.-A., L.M.A.-O., and N.O.-N.; methodology, A.R.G.-A.; software, A.R.G.-A. and C.F.B.-L.; validation, A.R.G.-A., L.M.A.-O., N.O.-N., and C.F.B.-L.; formal analysis, A.R.G.-A., E.E.Q.-B., and C.F.B.-L.; investigation, A.R.G.-A., E.E.Q.-B., P.N., C.F.B.-L.; resources, L.M.A.-O., N.O.-N., V.G.-A., and M.M.; data curation, A.R.G.-A., P.N., and C.F.B.-L.; writing—original draft preparation, A.R.G.-A., E.E.Q.-B., and C.F.B.-L.; writing—review and editing, A.R.G.-A., L.M.A.-O., N.O.-N., V.G.-A., E.E.Q.-B., P.N., M.M., and C.F.B.-L.; visualization, A.R.G.-A. and C.F.B.-L.; supervision, L.M.A.-O., N.O.-N., E.E.Q.-B., and M.M.; project administration, A.R.G.-A., L.M.A.-O., N.O.-N., and C.F.B.-L.; funding acquisition, L.M.A.-O., N.O.-N., V.G.-A., M.M., and C.F.B.-L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** The financial support of Universidad de La Sabana, Pontificia Universidad Javeriana, Universidad Tecnológica de Bolívar, Universidad de Cartagena, and Ryerson University is appreciated.

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

#### References

1. Zoppou, C. Review of urban storm water models. *Environ. Model. Softw.* **2001**, *16*, 195–231. [[CrossRef](#)]
2. Daly, C.; Neilson, R.P.; Phillips, D.L. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **1994**, *33*, 140–158. [[CrossRef](#)]

3. Johansson, B.; Chen, D. The influence of wind and topography on precipitation distribution in Sweden: Statistical analysis and modelling. *Int. J. Climatol.* **2003**, *23*, 1523–1535. [[CrossRef](#)]
4. Lin, G.F.; Chen, L.H. Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *J. Hydrol.* **2006**, *324*, 1–9. [[CrossRef](#)]
5. Rojas-Polanco, M.I.; Mora-Mora, L.E. Optimum design of rainfall network. *Rev. For. Venez.* **2009**, *53*, 9–22.
6. Chowdhury, M.; Alouani, A.; Hossain, F. Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater. *Stoch. Environ. Res. Risk Assess.* **2010**, *24*, 1–7. [[CrossRef](#)]
7. Chen, Y.C.; Wei, C.; Yeh, H.C. Rainfall network design using kriging and entropy. *Hydrol. Process.* **2008**, *22*, 340–346. [[CrossRef](#)]
8. Karabacak, K.; Cetin, N. Artificial neural networks for controlling wind-PV power systems: A review. *Renew. Sustain. Energy Rev.* **2014**, *29*, 804–827. [[CrossRef](#)]
9. Dursun, M.; Özden, S. An efficient improved photovoltaic irrigation system with artificial neural network based modeling of soil moisture distribution—A case study in Turkey. *Comput. Electron. Agric.* **2014**, *102*, 120–126. [[CrossRef](#)]
10. Asimakopoulou, F.E.; Tsekouras, G.J.; Gonos, I.F.; Stathopoulos, I.A. Estimation of seasonal variation of ground resistance using Artificial Neural Networks. *Electr. Power Syst. Res.* **2013**, *94*, 113–121. [[CrossRef](#)]
11. Adib, H.; Haghbakhsh, R.; Saidi, M.; Takassi, M.A.; Sharifi, F.; Koolivand, M.; Rahimpour, M.R.; Keshtkari, S. Modeling and optimization of Fischer-Tropsch synthesis in the presence of Co (III)/Al<sub>2</sub>O<sub>3</sub> catalyst using artificial neural networks and genetic algorithm. *J. Nat. Gas Sci. Eng.* **2013**, *10*, 14–24. [[CrossRef](#)]
12. Mohajerani, M.; Mehrvar, M.; Ein-Mozaffari, F. Using an external-loop airlift sonophotoreactor to enhance the biodegradability of aqueous sulfadiazine solution. *Sep. Purif. Technol.* **2012**, *90*, 173–181. [[CrossRef](#)]
13. Fukushima, K. Artificial vision by multi-layered neural networks: Neocognitron and its advances. *Neural Netw.* **2013**, *37*, 103–119. [[CrossRef](#)]
14. Mallela, U.K.; Upadhyay, A. Buckling load prediction of laminated composite stiffened panels subjected to in-plane shear using artificial neural networks. *Thin-Walled Struct.* **2016**, *102*, 158–164. [[CrossRef](#)]
15. Turlapaty, A.C.; Anantharaj, V.G.; Younan, N.H.; Joseph Turk, F. Precipitation data fusion using vector space transformation and artificial neural networks. *Pattern Recognit. Lett.* **2010**, *31*, 1184–1200. [[CrossRef](#)]
16. Liu, Q.J.; Shi, Z.H.; Fang, N.F.; Zhu, H.D.; Ai, L. Modeling the daily suspended sediment concentration in a hyperconcentrated river on the Loess Plateau, China, using the Wavelet-ANN approach. *Geomorphology* **2013**, *186*, 181–190. [[CrossRef](#)]
17. Garrido-Arévalo, A.R.; Agudelo, L.M.; Obregon, N.; Garrido, V.M. Classification of pluviometric networks located in the region of Bogotá, Colombia using artificial neural networks. *J. Phys. Conf. Ser.* **2020**, *1448*. [[CrossRef](#)]
18. Kar, A.K.; Lohani, A.K.; Goel, N.K.; Roy, G.P. Rain gauge network design for flood forecasting using multi-criteria decision analysis and clustering techniques in lower Mahanadi river basin, India. *J. Hydrol. Reg. Stud.* **2015**, *4*, 313–332. [[CrossRef](#)]
19. Wei, C.; Yeh, H.C.; Chen, Y.C. Spatiotemporal Scaling Effect on Rainfall Network Design Using Entropy. *Entropy* **2014**, *16*, 4626–4647. [[CrossRef](#)]
20. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
21. Vinod, H.D. Maximum entropy ensembles for time series inference in economics. *J. Asian Econ.* **2006**, *17*, 955–978. [[CrossRef](#)]
22. Noonan, J.P.; Basu, P. On estimation error using maximum entropy density estimates. *Kybernetes* **2007**, *36*, 52–64. [[CrossRef](#)]
23. Weber, T.C. Maximum entropy modeling of mature hardwood forest distribution in four U.S. states. *For. Ecol. Manag.* **2011**, *261*, 779–788. [[CrossRef](#)]
24. Payandeh Najafabadi, A.T.; Hatami, H.; Omid Najafabadi, M. A maximum-entropy approach to the linear credibility formula. *Insur. Math. Econ.* **2012**, *51*, 216–221. [[CrossRef](#)]
25. Aurbacher, J.; Dabbert, S. Generating crop sequences in land-use models using maximum entropy and Markov chains. *Agric. Syst.* **2011**, *104*, 470–479. [[CrossRef](#)]
26. Xie, L.; Li, G.; Xiao, M.; Peng, L. Novel classification method for remote sensing images based on information entropy discretization algorithm and vector space model. *Comput. Geosci.* **2016**, *89*, 252–259. [[CrossRef](#)]

27. Calisto Acosta, O.E. River gauging with one velocity point based on the principle of maximum entropy. *Ing. Hidráulica Mex.* **2002**, *17*, 5–19.
28. Dalezios, N.R.; Tyraskis, P.A. Maximum entropy spectra for regional precipitation analysis and forecasting. *J. Hydrol.* **1989**, *109*, 25–42. [[CrossRef](#)]
29. Mishra, A.K.; Özger, M.; Singh, V.P. An entropy-based investigation into the variability of precipitation. *J. Hydrol.* **2009**, *370*, 139–154. [[CrossRef](#)]
30. CAR. *2012–2015 Master Plan*; Corporacion Autonoma Regional de Cundinamarca (CAR): Bogota, Colombia, 2012.
31. Hurtado-Montoya, A.F.; Mesa-Sánchez, Ó.J. Reanalysis of monthly precipitation fields in Colombian territory. *DYNA* **2014**, *81*, 251–258. [[CrossRef](#)]
32. CAR. *2016–2019 Master Plan*; Corporacion Autonoma Regional de Cundinamarca (CAR): Bogota, Colombia, 2016.
33. OAS. *Manual for Design, Installation, Operation and Maintenance of Systems of Flood Early Warning and Online Database*; The Organization of American States (OAS), Department of Sustainable Development: Washington, DC, USA, 2010.
34. Wang, W.; Wang, D.; Singh, V.P.; Wang, Y.; Wu, J.; Wang, L.; Zou, X.; Liu, J.; Zou, Y.; He, R. Optimization of rainfall networks using information entropy and temporal variability analysis. *J. Hydrol.* **2018**, *559*, 136–155. [[CrossRef](#)]
35. CAR. *SICLICA—Sistema de Información Climatológica e Hidrológica: Valores Totales Mensuales de Precipitación, Máxima en 24 Horas (mm)*; Corporacion Autonoma Regional de Cundinamarca (CAR): Bogota, Colombia, 2010.
36. González-Cuéllar, F.; Obregón-Neira, N. Self-organizing maps of Kohonen as a river clustering tool within the methodology for determining regional ecological flows ELOHA. *Ing. Univ.* **2013**, *17*, 311–323.
37. Hamzehie, M.E.; Fattahi, M.; Najibi, H.; Van der Bruggen, B.; Mazinani, S. Application of artificial neural networks for estimation of solubility of acid gases (H<sub>2</sub>S and CO<sub>2</sub>) in 32 commonly ionic liquid and amine solutions. *J. Nat. Gas Sci. Eng.* **2015**, *24*, 106–114. [[CrossRef](#)]
38. Lohani, A.K.; Kumar, R.; Singh, R.D. Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. *J. Hydrol.* **2012**, *442*, 23–35. [[CrossRef](#)]
39. Chang, T.K.; Talei, A.; Alaghmand, S.; Ooi, M.P.L. Choice of rainfall inputs for event-based rainfall-runoff modeling in a catchment with multiple rainfall stations using data-driven techniques. *J. Hydrol.* **2017**, *545*, 100–108. [[CrossRef](#)]
40. González-álvarez, A.; Vilorio-Marimón, O.M.; Coronado-Hernández, O.E.; Vélez-Pereira, A.M.; Tesfagiorgis, K.; Coronado-Hernández, J.R. Isohyetal maps of daily maximum rainfall for different return periods for the Colombian Caribbean Region. *Water* **2019**, *11*, 358. [[CrossRef](#)]
41. Elshorbagy, A.; Corzo, G.; Srinivasulu, S.; Solomatine, D.P. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 1: Concepts and methodology. *Hydrol. Earth Syst. Sci.* **2010**, *14*, 1931–1941. [[CrossRef](#)]
42. Farsadnia, F.; Rostami Kamrood, M.; Moghaddam Nia, A.; Modarres, R.; Bray, M.T.; Han, D.; Sadatinejad, J. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *J. Hydrol.* **2014**, *509*, 387–397. [[CrossRef](#)]

